Multiple Imputation for Income

Seppo Laaksonen[1]

[1]University of Helsinki, e-mail: Seppo.Laaksonen@Helsinki.Fi

Abstract

Income is a demanding continuous variable in the case of missing data. One reason is that income of the non-respondents is often low although some high income people are not either good respondents. Imputation is in general a better method than weighting to solve the problem. The second point concerning income is that the average is not most interesting but income differences. This paper is focused on both these estimates using international data when we know true values. The missingness is created by the unknown person and hence we do not know its mechanism. The auxiliary variables available are not ideal leading to difficulties when implementing the imputation model. On the other hand, some imputation tasks are not working since they give negative incomes that are not correct at all.

Keywords: model-donor imputation, real-donor imputation, single vs multiple imputation

1 Introduction

Imputation is for replacing missing values with plausible ones. If this procedure has been done 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, 118-119) says that each imputation should be ‘proper’. He 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?

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

Several proper MI implementations are given in Rubin’s books and in software packages (e.g. SAS and SPSS) using his book. He thus recommends that imputations should be created through a Bayesian process as follows: (i) specify a parametric model for the complete data, (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 but his Bayesian rules might implicitly include the same; this is however difficult to recognize.

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 used in ordinary imputation literature; it cannot be found 9 years alter from a book by Carpenter and Kenward (2013) that much follows Rubin's framework but they use the term 'frequentist'. We still use the term 'non-Bayesian,' since we cannot say whether it is equal to 'frequentist.'

Björnstad motivates his approach also from the practical points of view saying that in national statistical institutes the methods used for imputing for nonresponse very seldom if ever satisfy the requirement of being "proper." Moreover, Muñoz and Rueda (2009) say that several statistical agencies seem to prefer single imputation, mainly due to operational difficulties in maintaining multiple complete data sets, especially in large-scale surveys. We agree with these views. Since a non-Bayesian approach also leads to single imputation, that is commonly used 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 applied non-Bayesian tools both for single and multiple imputation, although most often for single imputation. This paper first summaries our approach to imputation.

This approach first makes attempts to impute the missing values once. That is, the focus is first on 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 continuous variable, that is, income, two types of estimates are of a special importance. One is income average and the other is income distribution, respectively. Income distribution can be measured by various indicators such as quantiles or Gini coefficient, but the coefficient of variation is here considered to be simple enough to indicate well income differences between people.

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. We 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 application of SAS and SPSS. Respectively, our own imputation framework is called ‘Non-Bayesian MI.’

2 Imputation framework

In order to succeed in imputation, good auxiliary data or covariates 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 non-respondents, ‘proper’ imputation methods can be used. In this case, the imputation framework (cf. Laaksonen 2016) includes the two core stages:

(i) Construction and implementing of *the imputation model*

(ii) Imputation itself or *imputation task.*

These two terms are also used by Rubin (2004) but these are integrated well together in our framework. 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 model is estimated from the same data set, it is expected that this replacer behaves more surely well in imputations. Hence 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 response indicator of the variable being imputed.*’ The same auxiliary variables can be used in both models. Naturally, the estimations that are needed in the next step are derived from the different data sets, from the respondents for the model (a) and from both the respondents and the non-respondents for the model (b). The covariates need to be completely observed to compute the predicted values for the stage (ii).

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 option is called ‘*model-donor*’ imputation, and the second is ‘*real-donor’* imputation, respectively. The latter one is 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 regression 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 is needed. The natural criterion is to select an as a similar real-donor (observed value) as possible. This may be based on a kind of nearness metrics. If a clear criterion exists, it is good to select the nearest or another from the 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 can be called ‘an imputation cell,’ In our approach, the predicted values of either the model (a) or the model (b) are used as the nearness metrics, leading to real-donor methods. We focus on multiple imputation and hence we impute everything 10 times and calculate their average as the point estimate. The variance estimate is the sum of the between variance and the within variance. Rubin's formula does not include the response rate meaning the variance is smaller than in the case of Björnstad's formula.

Our framework thus is non-Bayesian and so we simply add the noise term to the predicted values. We test two types of the noise term using random numbers: (i) normally distributed residuals, (ii) normally distributed standard errors. We test several imputation models: (i) linear regression, (ii) log-linear regression, (iii) logistic regression, (iv) probit regression, (v) log-log regression (LL), (vi) complementary log-log regression (CLL).

SPSS and SAS use their methods and we simply apply them but we test two imputation models: (i) linear regression, and (ii) log-linear egression. They thus are Bayesian.

3 Empirical examples

The number of missing values or the imputation size is 3133 (out of 10000)) that is fairly realistic. The data set consists of a quite good number of covariates which all except age are categorical. The age was however categorized. The full list with the number of categories that is used in all imputation models is as follows: gender (2), five-year age group (11), marriage (2), civil status (2), education level (4), region (12), Internet at home or not (2), socio-economic status (4), unemployed or not (2), children or not (2). As seen any of these covariates is not well predicting yearly income (R-square of the linear regression model is about 40%).

Model-donor methods

The linear regression model is easy to apply for model-donor imputation but it does not give excellent results due to many negative values. Table 1 gives the results.

Table 1. Negative values of model-donor methods (NB = Non-Bayesian, B = Bayesian)

Method	Negative values, %
Using residuals NB	8.5
Using standard errors NB	0.3
SPSS B	16.8
SAS B	16.6

We find that all methods give negative values but Bayesian methods much more. Hence we do not use more model-donor methods but go to real-donor methods. We have explained already the basics of non-Bayesian methods but do not go details as far as Bayesian methods are concerned. Both SPSS and SAS have the method called 'Predictive mean matching methods' that always give observed values, thus not negative.

Real-donor methods

Table 2 presents the results. They are ordered by the imputation model applied. The last four methods are for binary regressions where are symmetric (probit, logit) and asymmetric link functions (CLL and LL). We find that log-linear regression is worst but it is not easy to know the reason. All imputed averages seem to be too big but the CV's almost always too small. Some imputation methods are however fairly good as far as income differences are concerned. One general conclusion could be that the imputations are leading to reduce the bias but not enough, concerning averages especially.

Table 2. Averages and coefficients of variation of yearly income and standard errors by Rubin and Björnstad

			Ranking		Mean ranking	Standard error of the mean	
Method	**Average**	**CV**	**Average**	**CV**		**Rubin**	**Björnstad**
Linear regression NB	46178	66.3	8	7	7.5	692	729
Linear regression SAS	45121	68.0	3	3	3	896	1017
Linear regression SPSS	45471	66.2	5	8	6.5	710	757
Log regression NB	46722	65.2	10	10	10	772	846
Log regression SAS	46034	66.7	7	5	6	864	973
Log regression SPSS	46179	66.1	9	9	9	692	728
Logit regression NB	45468	67.7	4	1	2.5	845	950
Probit regression NB	44785	67.9	1	2	1.5	754	822
CLL regression NB	44898	67.3	2	4	3	915	1047
LL regression NB	45493	66.4	6	6	6	864	976
True value	43531	67.7					

The major part of the standard error is derived from the within variance (from 59% to 79%). This is one reason that the differences between Rubin's and Björnstad's standard errors respectively are not big. They vary fairly much by methods. If the standard error is big, it is easier to get the result that covers the true value. On the other hand, a small standard error is often good. The reader can make his/her interpretation what method is best and which standard error formula. I prefer the probit regression NB.

References

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

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

Laaksonen, S. (2016). A new framework for multiple imputation and applications to a binary variable. *Model Assisted Statistics and Applications*, 11.3, IOS Press. http://content.iospress.com/journals/model-assisted-statistics-and-applications/11/2

Muñoz, J.F. and Rueda, M.M. (2009). New imputation methods for missing dada using quantiles. *Journal of Computational and Applied Mathematics* 232, 305-317.

Rubin, D. (1987/2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library Edition.

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