# Bayesian and non-Bayesian multiple imputation with examples for a continuous variable

Seppo Laaksonen[1]
[1]University of Helsinki, Finland

Imputation is for replacing missing and deficient values with plausible ones. If this procedure has been done once, it is single imputation. This is a usual way in statistical offices, for instance. The single imputation can be performed several times as well. If this procedure is 'coordinated' well, the outcome is 'multiple imputation' (MI). What such a good coordination means, it is a special question? Rubin in his book (1987) says that each imputation should be proper. He also gives some rules for that but they are not necessarily easy to follow, or their implementation is not automatic.

Several proper MI implementations are given in Rubin's book and in software packages after that. Rubin recommends that imputations be created through a Bayesian process: specify a parametric model for the complete, apply a prior distribution to the unknown model parameters, and simulate *L* independent draws from the conditional distribution of the missing data given the observed data by Bayes' Theorem.

I test two general software packages, SAS and SPSS, respectively, and assume that their MI procedures follow a Bayesian process. Correspondingly, I use the term Bayesian MI. Since a Bayesian process is not according to Björnstad (2007) practicable to follow in statistical offices, in particular, other types of MI procedures can be created. For these I use Björnstad's term 'Non-Bayesian MI.'

Interestingly, Björnstad also presents a partially different formula for the MI variance. Fortunately, the estimate itself is equal, that is, the average of the single estimates (*Q* = the estimate and *L* = the number of imputations):

$$Q_{MI} = \frac{\sum_u Q_u}{L}$$

Björnstad's variance is below. The first term here is the average of the variances of the *L* completed data sets (within-variance) whereas the second term is the between-imputation variance). This formula is equal to Rubin's initial formula if the *k* =1. But this does not hold since it increases according to Björnstad while the non-response rate *f* increases.

$$B_{MI} = \frac{\sum_u B_u}{L} + (k + \frac{1}{L}) \frac{1}{L-1} \sum_u (Q_u - Q_{MI})^2 \qquad \text{in which} \qquad k = \frac{1}{1-f}$$

:

To interpret these differences. Rubin's formula does not explicitly take care of missing values to be imputed but Björnstad's does. It is possible that Rubin's Bayesian procedure takes this into account implicitly, but I am not convinced about this.

The empirical part of my paper presents a number of results of the imputation of a challenging continuous variable, that is, income. The data set is rather ordinary in practice. The gross size is about 20000 and the net size nearly 15000. Thus, the nonresponse rate is 27 per cent and these missing values need to be imputed, respectively. The data set includes 9 auxiliary variables, such as age, gender, region, socio-economic status, education and civil status but any of them is not correlated well with income. This is seen from the multivariate regression model in which the R-square is 39 per cent. I bit larger R-square (41 %) is obtained if log-income is used that is usual in econometrics.

My strategy for imputations is the following: (i) construct the imputation model and estimate its important parameters, (ii) using those estimated parameters predicti plausible values for replacing the missing ones. There are also the two alternatives for the imputations themselves, (i) model-donor methods in which case the imputed values are computed in a certain way, (ii) real-donor methods in which case the imputed values are borrowed from the observed values.

In the case of the continuous income variable, the two types of the imputation models are tested so that either (i) the variable being imputed or its logarithm is a dependent variable, or (i) the binary missingness

indicator is the dependent variable, respectively. In the latter case, three link functions (logit, probit and complementary log-log) are tested. In these non-Bayesian tests the same auxiliary variables are used in all the imputation models. This is concerned also Bayesian tests that are made using both newest SAS and SPSS procedures. These procedures give opportunity to test also the predictive mean matching imputation method that does not give negative values for income, for example, unlike the regression based and MCMC methods do. The same is concerned some of my non-Bayesian methods but not so extensively than for 'black box' methods of SAS and SPSS.

The room of the abstract does not give opportunity to present results in details, but I have to say that I have not been very happy with SAS and SPSS ready-made procedures. Naturally, all my methods are not satisfactory either but some will be preferred from the bias point of view that is possible in my simulations ($L = 10$) since we know the true values. An obvious reason why some methods are even bad is that the imputation model is difficult to specify since any strong variable does not exist. But this is an ordinary case in real-life.

Keywords: Single imputation, two formulas for the variance, empirical tests with different MI methods,

## References:

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

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