Journal of Official Statistics, Vol. 23, No. 4, 2007, pp. 467–475

Discussion

Seppo Laaksonen[1]

1. Introduction

Bjørnstad's article is a welcome contribution to the discussion on multiple imputation (MI) in survey estimation that is rarely used by national statistical institutes (NSI), as observed from for instance my own small-scale survey (Laaksonen, 2005). This survey, however, shows that MI methodology has been under discussion and investigation at some NSIs and, consequently, may actually be exploited some day. However, the big question is how to use it. I think that technical aspects are no longer its main obstacle in the sense that standard sample survey data files can be handled without problems with new computers, and even be reproduced several times by imputing. This obstacle was often mentioned in a discussion in the *Journal of the American Statistical Association* in 1996 (see Rubin, 1996 and its discussion). This will continue to be a problem with very large files.

However, MI is not a general tool for handling missing data as Rubin would hope. I have understood especially from discussions on the web that MI is most popular among biometricists, but they focus on multivariate data analysis, not on estimating means, totals and other statistics. Are NSIs more conservative, or are there some other reasons behind this nonuse of MI? I argue that the main reasons lie in the difficulties of using MI well. This is not only due to MI itself but also, and more importantly, to the fact that NSI data files are complex and not well suited as current tools for MI. I am not here saying that MI is to blame, but the main reason is that, on the whole, it is difficult to find any good imputation method for many practical situations.

It should be noted that MI is not (as it is often understood to be) a specific imputation method, but just an option for variance estimation, as Bjørnstad also points out. This, thus, means that before repeating imputations and creating several datasets, one has to concentrate on single imputation (SI) and hope to succeed in the best way at this stage. This fact, unfortunately, has often been forgotten by MI contributors. At the next stage, a good strategy should be found for repeating SIs so that variance estimates could be estimated as correctly as possible. At this stage, SI needs to be extended with a stochastic solution. This MI strategy does not work with many NSI data at all. In many practical situations it is best to only use a deterministic method and, hence, exclude MI (but still try to estimate the variances correctly). It is also difficult for me to encourage the application

[1]Statistics Finland, Department of Statistics, P O Box 68, FIN – 000214 Statistics Finland. Email: seppo.laaksonen@helsinki.fi

of MI in cases where sampling weights are equal to one, or small, as often occurs in business surveys, in particular.

As Bjørnstad points out, the core question in imputations is, to use Rubin's term, requiring the method to be "proper." This term is nice and illustrative, and in his book from 1987 and later on in several articles Rubin (1987) explains what it means. However, the "properness" requirements have been presented at quite a general level and it is not easy to see how they should be interpreted in practice. To me, imputation is proper if the missingness due to nonresponse and other things is traced back as well as possible from the point of view of each estimation task. This requirement also naturally concerns variation. Bjørnstad gives a more general interpretation, saying that imputations should be random draws from a posterior distribution when exploiting a Bayesian framework. Next, he concludes that the methods NSIs use for imputing for nonresponse very seldom, if ever, satisfy the requirement of being "proper." This is not explicitly proven in his article, but in my opinion this argument is correct also outside NSI data files (see e.g., the discussion relating to Rubin 1996, Fay 1996, and Rao 1996).

Thus, imputation results will obviously be improving as imputation strategies approach "properness," whatever this term means. At the same time, the quality of point estimates will improve, too. As already mentioned, the first key requirement for good point estimates is success in SI. If several SIs have been performed, the main advantage of MI will be in the robustness of these point estimates, since some uncertainty will be reduced. However, this advantage is not expected to be substantial for simple point estimates such as averages and totals. Instead, certain quantile estimates for continuous variables could be more robust and reliable when using MI. (This comment is based on my experiment for estimating the poverty rate of Finland in the early 1990s (Laaksonen 1992). Poor households responded worse in this particular survey but I was able to use register data to impute disposable incomes for unit nonrespondents. My MI corresponds to Fay's (1996) "fractionally weighted imputation." This technique was considered the best for estimating the number of poor households, but this was not the case in general.

Properness is even more important for interval estimation. Rubin's well-known formula includes the two terms, which I here call the within-imputation-variance and the between-imputation-variance. Both of these should naturally be based on best possible estimates using m completed datasets. Bjørnstad's basic point is to present an alternative approach to the calculation of the second term. He says that this result does not require any Bayesian framework and his formula is also in his opinion more applicable for NSIs. Bjørnstad develops formulas for different imputation strategies but the basic feature of each formula is essentially the same as that of the next. His coefficient for the between-imputation-variance is of type $(k + 1/m)$ in which k is the inverse of the response rate, whereas in Rubin's formula, k is fixed to be equal to one.

Unless the between-imputation-variance is equal to zero, the new formula increases the variance estimate while the missingness rate increases. Respectively, the standard error and the confidence interval will be larger. Moreover, the imputations do not need to follow any Bayesian rule as would be the case if $k = 1$. It is not easy to act as an adjudicator between these two parties, since both results are conditional on many things. As already discussed, the big question is how to make imputations properly concerning both SI and MI. Further I suppose that Bjørnstad's approach also requires strict rules for performing

his imputations so that the requirements of his approach will be met ideally. He does not discuss this question much, although his examples shed some light on it. I, thus, suggest that he should determine how to make imputations "correctly" (if not "properly") when using his non-Bayesian framework. I will come back to these substantial questions in Section 3 after presenting some empirical tests in the next section.

2. Proper or Improper Imputations for Two NSI Datasets

The title of this section stems from my problem in ascertaining whether a certain imputation strategy is "proper" or not. I hope that having seen the results from the following two patterns of examples the reader will be able to answer this question. The variable being imputed is simple and *binary* in the first pattern, whereas it is quite complex and *continuous* in the second. However, both examples are complex from the point of view of their missingness mechanisms. Awkwardly, the auxiliary variables are by no means ideal, as unfortunately is usually the case with the data of NSIs and other survey institutes. This means that it is difficult to develop "proper" adjustment strategies. In the first example this succeeds slightly better. I show results from several imputation approaches and from some weighting adjustments in order to illustrate the results in a broader perspective. The methods and results of these experiments are presented in the following two subsections.

2.1. Binary Variable

Laaksonen and Chambers (2006) performed survey estimates based on several methods in a fairly simple follow-up situation. In the first phase of the survey, the response rate was about 67 per cent. Hence the follow-up of the nonrespondents was carried out based on a 40 per cent simple random sample. Due to the great efforts made in this survey, full responses were received to the simple question irrespective of whether the particular units (businesses) were innovative or not. The same question was, of course, included in the actual survey.

The dataset gave several options for both point and interval estimation. In all cases, the responses of the second phase of the survey were exploited as auxiliary information. In the original article, two main strategies were used, weighting adjustments and the prediction approach. The weighting adjustments were performed in two ways, both with respondents of the first phase, but with different auxiliary data. This covered the first and second phase respondents in the first approach, but the whole initial sample in the second. The latter approach required imputation of missing values for 60 per cent of the first phase nonrespondents. These same imputed values were also exploited in the prediction approach. For this article, I performed these imputations eight times, and on each occasion the data for MI-based estimations were available.

Our imputation strategy was simply to borrow each imputed value from a respondent of the second phase sample. This was done using a random draw with replacement (often called random hot-decking, but I do not like this term). The method corresponds to the strategy of the first example of Bjørnstad. Since all the draws in MIs are conducted independently of each other, I assume that the strategy is "proper" from both the Bayesian and the non-Bayesian point of view.

Our weighting strategy was based on two steps: response probabilities were calculated first and rescaling was performed next so that the margins were correct. The variance estimation formula of this approach includes two additive terms. The first term is analogous to the relevant sampling variance but is based on adjusted weights. The second term shows specifically the impact of missingness and this thus increases with the missingness rate. It is analogous to the second term of the Bjørnstad formula and also resembles the variance formula of mean imputation. For details of all these formulas and the data see Laaksonen and Chambers (2006).

This experiment was not real although efforts were made to make it as close to real life as possible. Consequently, we used simulations in our analysis. Hence, we were able to compare the results from different aspects but I will here only take a close look at the variance estimates. It should be noted, however, that the point estimates were fairly correct with all these strategies although not complete. The variance estimates varied more, although not dramatically. The largest variances (around 6,000) were obtained using the prediction approach and the imputation-based weighting approach. The first weighting approach gave a slightly lower value (5,700) than might have been considered logical since uncertainty due to imputations was absent. Both Bjørnstad's formula (5,800) and Rubin's formula (5,400) generate logically lower values than the corresponding SI methods. How much lower should these variances be? The value from the Rubin formula could be considered too low but I am not sure about this conclusion. The 95 per cent confidence interval coverage of this method is just below 95, whereas for Bjørnstad it is just above 95. All other methods give somewhat conservative estimates, the coverage varying from 95.2 to 96.3.

2.2. *Continuous Variable*

Wages and incomes and their many components are typical NSI variables. Missing values often very awkwardly violate estimates based on these variables. I was here able to use the income data of the Euredit project (Charlton 2003). Some missingness was found within the data set, but afterwards we had the opportunity to compare these estimates against those computed from the cleaned, "true" data.

The dataset of this experiment is very large; almost 200,000 persons after omitting persons with zero incomes. The relevant missingness rate was about 27 per cent. It was known that the missingness mechanism could not be ignored but the file contained a number of potential auxiliary variables for adjustment attempts. The problem was that, except for age, all these variables were categorical (such as gender, level of education, marital status, region, and employment status) and not well-related to incomes at the individual level (but much better at aggregate levels). This did not allow the construction of a well-fitting adjustment model. In terms of *R*-square, the maximum values were around 40 per cent. The fitting difficulty is a good thing here since it will encourage the development of a strategy for "proper" imputation or for another good adjustment (see also the results from the Euredit project in which a number of SI methods were tested without any magnificent success).

I performed three different imputations using this dataset. One was based on the SAS procedure for MI (lately referred to as "*Proc MI Method"* and also as "*SAS"*). In this

approach I used the same imputation model as with my second method. So, the variable being imputed was the dependent variable of the general linear model. The auxiliary variables were used maximally (as e.g., Rubin 1996 proposes) and exactly identically in all of the three imputation models, thus including the third model as well. However, the dependent variable of the third model was the binary missingness indicator. Consequently, this model was estimated using logistic regression.

The "Proc MI Method" is fairly automatic and finds the imputed values when certain rules have been inserted into the procedure. I did not concern myself with how the SAS really produces the results. I was more like an ordinary user. However, I calculated the variance estimations myself from the multiply imputed datasets so that both Rubin's formula and Bjørnstad's formula were used and results calculated for both.

In the second method, having first estimated the multivariate general linear model (*GLM*) I then calculated the predicted values plus the noise terms first and then used these metrics when searching for the nearest neighbour for each person with missing income data. The noise term was calculated using truncated, normally distributed random numbers with the zero as mean and with residual variance (see details of this approach in Laaksonen 2003, Euredit 2004 and Laaksonen 2005, also known as *IMAI*, which is short for Integrated Modelling Approach to Imputation). This *IMAI* strategy gives an easy opportunity to perform MIs, hoping that each SI also gives quite good point estimates. I suppose the method is fairly "proper" in all senses but further research is still needed.

The third example is analogous to the second one (belongs to the *IMAI* framework as well, and is referred to as "*Logit*") so that each imputed value has been taken from a nearest donor without missingness and the metrics are based on the predicted values plus the noise term of the imputation model. Since these values are now probabilities, it is not automatically clear how the random noise should be performed. I deduced that it would be good to find each donor at random from the neighbourhood of a unit with missingness. This was made possible by adding to each deterministic probability a random number that was drawn from a uniform distribution of the interval $(-5\%, +5\%)$. I know that this is a slightly heuristic method like some other methods in the MI literature (see e.g., the strategy of the Solas software where estimated response probabilities are used to form a certain number of imputation cells and the donor is drawn randomly within each such cell).

Apart from the above selections for SIs and MIs it is not automatically clear how each model should be specified. An advantage of the third approach is that the response variable is simple and needs no transformation. By contrast, the income variable is not as simple as this due to its skew distribution. The usual way in subject-matter research is to use its logarithm. This was attempted here, too, and it worked quite well with the second method. For a reason I do not know, this did not work with the first method when the average was estimated. Instead, it worked better for estimating the coefficient of variation (CV) that measures income differences. In the case of the second method, both models, linear and logarithmic, worked quite well for the average and the CV. The third method also gave reasonable point estimates (although, as usual in my experience, the general tendency was that the bias would not be reduced completely, and only a value approaching the true one would be obtained).

Nevertheless, I did not use logarithms in any of the imputation models, since this strategy was not working well in the case of the CVs of the first method. Thus, although

I am not happy with all point estimates derived with the "Proc MI method," I considered it fair to concentrate on the results in which all point estimates (concerning e.g., average income) are reasonable and compare the variance estimates with each other.

For empirical tests I wanted to have different sizes of datasets. This could be easily achieved by trimming an initial dataset randomly. Likewise, I wanted to see what effects larger missingness would have. For this test, the trimming was performed to the 50 per cent response rate.

As far as repeated SIs are concerned, the MI literature often states that a reasonable number could be in the range from five to seven. I first chose eight imputations for each experiment, but since some of the results were not at all logical, I extended this number to 16. The results in Table 1 are based on this number although I am still not completely satisfied with them. Altogether the table includes $3 \times 2 \times 2 \times 7 = 84$ experiments presented as standard errors, since these are smaller and more illustrative than variances.

Table 1 allows a number of interpretations. We observe, for example, that the standard error is not always larger when the missingness is larger, as it logically should be (these discrepancies are in italics). This problem is especially present in small sample sizes with the "Proc MI method" (more imputations than 16 might have helped). The obvious reason for this too small variation in the remaining sample is that the method does not work properly. A more detailed look shows that the between-imputation-variance is particularly small when the nonmissingness is 73 per cent of these samples and, consequently, the difference between Rubin's method and Bjørnstad's method is not significant. In general, these two alternatives give a different picture of the imputation uncertainty.

In most cases, the "Proc MI method" gives the lowest standard errors, except for the full sample of 200,000 units, where all standard errors of this sort are relatively large, and those based on the Bjørnstad formula are naturally the largest. This is difficult to interpret but again illustrates that the properness of this automated method is problematic. The standard errors of my two methods, *GLM* and *Logit*, are fairly logical, the largest ones resulting from the *Logit*-based nearest neighbour methods. For this method the standard error between the two nonmissingness rates is also the largest. The same concerned adjustments from Rubin's formula to Bjørnstad's formula.

The overall interpretation of Table 1 is that it is difficult to find an imputation method that could be viewed as "proper," especially when the sample size is small and the missingess is large. The automated software does not give any guarantee for such a success. In this exercise, I was not able to make simulations and, consequently, get any ideal benchmark. Fortunately, the dataset gives the option to adopt an alternative perspective, that is, to construct reweights for the data and perform corresponding estimates using these.

I thus created adjusted weights following the same methodology as in the case of the binary variable experiment (Laaksonen and Chambers 2006). The adjustment model was in this case exactly the same as in the third imputation method of this subsection but, naturally, the estimated response probabilities are now used without any noise term, and only for the respondents. Table 2 presents the analogous results with the imputations of Table 1. In addition to the overall standard errors, I include in the table its first component

Table 1. Estimated standard errors for average income based on three different multiple imputations for seven different gross sample sizes and two nonmissingness rates. Notation: SAS denotes the first method (Proc MI), GLM the second and Logit the third; R denotes Rubin's, B denotes Bjørnstad's formula

Sample size	MI methods											
	R-SAS		B-SAS		R-GLM		B-GLM		R-Logit		B-Logit	
	73%	50%	73%	50%	73%	50%	73%	50%	73%	50%	73%	50%
3,000	*1,952*	*1,791*	*1,955*	*1,830*	*2,367*	*2,351*	2,447	2,574	2,405	2,814	2,496	3,370
5,000	*1,536*	*1,372*	*1,539*	*1,398*	1,840	2,034	1,899	2,363	2,140	2,724	2,280	3,472
10,000	*1,105*	*1,045*	*1,113*	*1,069*	1,233	1,355	1,252	1,479	1,426	1,951	1,497	2,434
20,000	*801*	*769*	813	826	937	1,011	973	1,155	900	1,255	924	1,546
50,000	*537*	*516*	552	568	570	750	612	916	614	651	687	759
100,000	403	458	417	509	401	443	429	501	459	518	499	628
200,000	390	375	424	455	287	336	308	394	287	327	299	381

Table 2. Estimated standard errors for average income based on the weighting of seven different gross sample sizes and two nonmissingness rates

Sample size	Based on both variance terms		Based on the first term only		Assuming simple random sampling (SRS)	
	73%	50%	73%	50%	73%	50%
3,000	4,480	5,627	3,964	4,536	2,500	3,012
5,000	3,296	4,086	2,894	3,196	1,945	2,301
10,000	2,310	2,960	2,033	2,318	1,380	1,727
20,000	1,594	2,047	1,389	1,580	976	1,194
50,000	992	1,298	862	1,005	616	744
100,000	698	925	606	716	435	520
200,000	496	651	430	506	308	367

and the simple random-based standard errors in order to gain a wider perspective for comparing the results with those presented in Table 1.

First, it is important to note that the point estimates using these reweights were very accurate, both in respect of averages and of CVs. This method can thus be considered to be very reliable. Interestingly, the standard errors of the full formula (Columns 2 and 3 of the table) are always much larger than those from any comparable MI experiment. Furthermore, even if the incomplete formula (the second column) has been used, which is often the case in practice, the standard errors are still large, compared to MI-based ones. Finally, the SRS-based sampling errors are also quite large. Thus, this would appear to provide some evidence of the superiority of a good MI method if the only criterion was the estimated standard error, but I do not draw this conclusion. Instead, this is another new drawback of the SAS MI method, especially in the case of small sample sizes. On the other hand, these results are more favourable for Bjørnstad than for Rubin.

3. Further Comments

What have we learned thus far from my experiments and Bjørnstad's article? The experiments lend some support to the argument that Bjørnstad's formula is safer, since it generates a larger standard error (and confidence interval). In some cases, when the between-imputation-variance is small, this formula does not help much, although the imputation model was badly fitting. This illogical result could be interpreted as a clear indication of the improperness of such an imputation method. Fortunately, the between-imputation-variance is more significant in most MI experiments and, consequently, the effect of the Bjørnstad formula is more essential. In these cases, the imputation method could be considered as the more proper one but I cannot, nevertheless, be convinced about the properness of these methods either. This annoyance is usual when handling NSI data.

In my view, Bjørnstad is right in his key conclusions but, as he says, this topic requires further research. I would like to see imputation research concentrate on the properness of imputation, and more practically than what has been the case up to now. This remains a big problem both for Bayesians and for non-Bayesians.

4. References

Charlton, J. (ed.) (2003). Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project. http://www.cs.york.ac.uk/euredit/. Specifically the following chapters: – Volume 2, Chapter 12.10. Method 10: Integrated Modelling Approach to Imputation and Error Localisation (IMAI) written by S. Laaksonen. – Technical Appendix: Descriptions of Datasets and Their Perturbations written by H. Wagstaff.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91, 490–498.

Laaksonen, S. (1992). Adjustment Methods for Nonresponse and Their Application to Finnish Income Data. Statistical Journal of Economic Commission for Europe of United Nations, 9, 125–137.

Laaksonen, S. (2003). Alternative Imputation Techniques for Complex Metric Variables. Journal of Applied Statistics, 1006–1021.

Laaksonen, S. (2005). Integrated Modelling Approach to Imputation and Discussion on Imputation Variance. UNECE Work Session on Statistical Editing, Ottawa, 16–18 May. http://www.unece.org/stats/documents/2005/05/sde/wp. 36.e.pdf.

Laaksonen, S. and Chambers, R. (2006). Survey Estimation under Informative Nonresponse with Follow-up. Journal of Official Statistics, 22, 81–95.

Rao, J.N.K. (1996). On Variance Estimation with Imputed Survey Data. Journal of the American Statistical Association, 91, 499–506.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

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

Received January 2006