Journal of Official Statistics, Vol. 23, No. 4, 2007, pp. 477–483

Discussion

Daniel Thorburn[1]

1. Background

Missing data can cause problems for both scientific experiments and official statistics. Every contribution to the handling of incomplete data is therefore welcome. One way to attack the problem is to collect more data on the missing elements, e.g., do follow-up surveys, control studies or Hansen-Hurwitz plans. The other main approach is to find good mechanical ways to improve the estimation phase, often, but not always, using auxiliary data. Imputation is an example of this approach. Other examples are post-stratification, calibration or reweighting by estimated response probabilities. One must then make prior assumptions on the models and on the relation between the data and the nonresponse. Such models may be simple, like MAR (Missing at Random) and MCAR (Missing Completely At Random) or more complicated like those involving the assumption that the probability of no reply can be described by a logistic function or that the drop-out follows the same pattern as it did in another similar study.

It is important to choose the method that is most appropriate to the problem. Björnstad does not discuss when imputation is suitable. He studies only very nice and simple cases, assuming MAR or MCAR. His main conclusion is that it is possible to develop multiple imputation methods that do not require draws from a Bayesian posterior distribution. He does not discuss when his alternative is suitable and could be recommended. A good introduction to the treatment of missing data is De Leeuw et al. (2003). For those who want a more extensive treatment of multiple imputation Schafer (1997) can be recommended.

There exist many interpretations of the concept of multiple imputation and many procedures have been suggested. Some of these techniques are quite silly and some of them are quite ingenious. When I use the term multiple imputation it is in the original sense of Rubin, but I will also include some developments using MCMC techniques.

2. Imputation is Not Needed at All in the Examples

All the examples of Bjørnstad are fairly simple. He assumes MCAR or MAR. In all his cases he can just standardise the inclusion probabilities

$$\pi_i^* = \pi_i \frac{n_R}{\sum_{j \in R} \pi_j}$$

[1]Stockholm University, Department of Statistics, SE-106 91 Stockholm, Sweden. Email: daniel.thorburn@stat.su.se

(or a similar formula in the stratified cases). He can then forget the nonresponse and use standard estimators (e.g., Horvitz-Thompson and Yates-Grundy for variance estimation). This is by definition true for MCAR. When only MAR holds a suitable estimate that uses the auxiliary variables must be chosen. The ratio estimator of Example 2 and the (post-) stratified estimators in later examples are examples of such estimators. This would have given simpler computations and no imputation errors at all.

3. Ordinary MI Is Better from a Classical Point of View

3.1. Optimality

The properties of multiple imputation are usually derived using Bayesian techniques. Multiple imputation is thus automatically inadmissible in itself. This result requires that m tends to infinity and that the intended procedure is efficient when the complete sample, S, is known.

It is well known that Bayesian estimates with flat (vague) priors and symmetric distributions are ML estimates. General results on the optimality of Bayesian methods can for instance be found in Fergusson (1967). It is also known that ML estimates under assumption of normality are best linear unbiased estimates (BLUE). Thus for large m the standard multiple imputation based on Gaussian distributions and a flat prior gives the best non-Bayesian estimate, regardless of what the true distribution is (the distribution does not even have to exist as in finite population sampling for this result).

Bjørnstad considers only (functions of) linear estimates in Section 5. Thus his methods are inferior to ordinary multiple imputation as described by Rubin. It must be said that since the datasets are usually large in survey sampling asymptotic results can be used. The asymptotic efficiency of Bjørnstad's method is often 1 and the loss in efficiency is thus quite small. However, another disadvantage is that the computations are slightly more complicated and more checking has to be done.

That Bjørnstad's method is not optimal can also be seen from the fact that the variance $\hat{V}(\hat{\theta}(s_i^*))$ has only approximately the correct expected value. Bjørnstad notices this e.g., when he in Example 3.1 says that

$$E(\hat{\sigma}_*^2|R) = \hat{\sigma}_r^2(1-\frac{1}{n_r})(1+\frac{n_r}{n(n-1)}) \approx 1-\frac{1}{n}(\frac{n}{n_r}-\frac{n_r}{n})-\frac{1}{n(n-1)} \approx 1$$

For the large datasets of official statistics this approximation is quite acceptable but it does not hold for small sample sizes.

Thus from a non-Bayesian classical point of view Rubin's methods with normal distributions and vague priors are better for all the situations that Bjørnstad considers.

3.2. Hypothesis Testing and Confidence Intervals

One of the basic ideas of multiple imputation is that it is known what to do if there are no missing values. In order to use this, the distribution of the completed samples, $S_i, i = 1, \ldots, m$, should, given the response set, R, be as similar as possible to the distribution of the full sample, S. This means that the imputed samples, S_i, should

optimally be drawn from the distribution of S given R. The conclusions will then be close to those that could have been expected if you had known S completely.

Another basic idea is that the imputation should be repeated so many times that the law of large numbers guarantees that the mean of your estimators does not depend on the home-made randomness. From a practical point of view it is usually sufficient with eight or ten independent imputations. But to get exactly the same result if the whole multiple imputation is repeated, usually far more than 100 imputations will be needed.

The fact that Bjørnstad needs to use a number $k > 1$, shows that the variation in his imputed samples is less than in the full sample. If the variation is too small many common analyses cannot be made. For instance, one may want to perform a standard hypothesis test. In ordinary multiple imputation this is done by computing the p-value in each completed sample and then computing the average of them. This gives the true p-value for that hypothesis. But since Bjørnstad's samples have too small variation, he will reject the null hypothesis too often. It must be admitted, though, that Bjørnstad's method is much better than many other imputation methods like mean imputation.

4. Ordinary Multiple Imputation Solves Also More General Problems

Bjørnstad considers only uninformative sampling, only linear estimates, only unit nonresponse and only one variable at a time. He gives no clue how to treat informative sampling, nonlinear problems, item nonresponse or estimation of several parameters from the same data.

Let me describe one or two naturally occurring situations of each type. It would be interesting to know how Bjørnstad intends to resolve these standard multiple imputation situations.

4.1. Informative Sampling

The frame contains an auxiliary variable X. The data have been collected in two phases. The first normal data collection phase gives the response set R_1. After that a specially selected subsample R_2 is taken from the nonresponse set and complete data is obtained from that set. Using all data the response probability is estimated by the logistic expression

$$\pi(x) = \frac{\exp(\hat{a} + \hat{b}x + \hat{c}y)}{1 + \exp(\hat{a} + \hat{b}x + \hat{c}y)}$$

where c is significantly different from 0. The missing data $(S - R_1 - R_2)$ can now be imputed taking into account both the informative sampling and the uncertainty about the parameters e.g., a, b and c. The population total of Y is then easily estimated using multiple imputation.

4.2. Nonlinear Problems

Data is collected from a sample of ten-year-old children on their performance in school in five subjects and also on their behaviour and social integration into school. Data is also collected on the parents' education and the home conditions. A researcher wants to study if there is a relation between the home conditions and the performance when the influence of

the parents' education is removed. He intends to use a simple Structural Equation Model (SEM). He is also willing to assume that the nonresponse is MCAR. Using ordinary multiple imputation it is quite straightforward to impute the missing values and to use for instance LISREL on the imputed datasets analysing the appropriate SEM model. On the other hand the estimated covariance matrices in the imputed datasets of Bjørnstad are severely biased. As a consequence, a use of LISREL will give biased results. I cannot see how this problem can be remedied.

4.3. Item Nonresponse

Data is collected from a number of persons in, say, a survey on living conditions. However, some persons refused to answer some sensitive questions or gave unreadable answers or just forgot to fill in an answer. The data matrix thus has some holes. The data for the other variables can often be used to give information on the missing values for the same individual. With multiple imputation it is quite simple to use these answers to improve the imputation of the missing values. Estimates on totals will then be much better than if only the persons who respond to all questions are used.

Another example could be the school data mentioned above. In data of this type some background data are often missing or children may have been sick and have not been able to participate in all tests.

4.4. Several Parameters

Suppose that one has a register of firms with the number of employees, X. Data on the total wage costs Y is collected from a sample of the firms. Two parameters are studied, the average wage cost per firm $\frac{1}{N}\sum_U y_i/N$ and the wage cost per person in an average firm $\frac{1}{N}\sum_U y_i/x_i$. It is natural to estimate the first quantity with a ratio estimate as in Example 3.2 where Bjørnstad recommends $k = 1/(1 - f_x)$. The second quantity is more natural to estimate by an arithmetic mean as in Example 1, where Bjørnstad recommends $k = (1 - f)$. This means that different values of k ought to be used and there will be inconsistency problems. I cannot see how Bjørnstad can estimate the covariance between these two estimators using his approach. This is quite easy to do with ordinary multiple imputation.

Multipurpose surveys are quite common and whenever it is appropriate with different models *(k)* for different variables the suggested method will have troubles. An example of this is Bjørnstad's Section 4.2 on stratified sampling, where he cannot get a nice expression for a combined k. If the stratification were combined with a regression model ($y_i = a_h + bx_i + \varepsilon_i$), he would have been forced to use a common k, and I cannot see how he in a simple way could combine population and stratum estimators. With ordinary multiple imputation this is simple and one also obtains the covariance between all the estimators.

5. Some Further Comments on Multiple Imputation

5.1. MAR and MCAR

The notions MCAR and MAR are basically Bayesian notions. MCAR means that (Y, X) and the nonresponse mechanism are independent and MAR means that Y and the

nonresponse mechanism are independent conditionally on X. For example, MCAR does not hold if the response pattern is different in different strata. But in a classical setting X is considered to be a known parameter of the model and not a random variable. Since fixed constants and random variables are independent per definition a classical interpretation of probability and independence should formally imply that MCAR holds in this situtation. The approach in the article is thus not completely free from Bayesian arguments, even though Bjørnstad claims so.

5.2. Realistic Imputations

Many program packages and types of analysis require data of a certain type. If means and variances are the interesting parameters, it is quite natural to impute other values than 0 or 1 for sex, but for other uses it is not an alternative. The method of Bjørnstad does not always guarantee that realistic values are imputed (for instance in Example 3.2 where residuals are imputed). If sensible distributions and priors are used in the multiple imputation (like Poisson and a vague Gamma distribution) only possible values (integers) are automatically imputed. However, if nonnormal distributions are used the result on BLUE no longer holds, but that is not always a disadvantage (e.g., if one knows that there are both male and female individuals in the population but that the ten-person-response set consists of only males, it must be sensible to take into account that some of the missing persons may be females). The only unbiased estimator is 100% male, but that is known to be an overestimate.

5.3. Technical Aspects of Linear Estimates

If the estimates are linear in the imputed data, it is easily realised that any set of imputed values will (asymptotically in m) be efficient as long as their conditional covariance is the same. For example, in the first example of Bjørnstad, i.e., when data are "missing completely at random (MCAR)" and the design is simple random sampling (SRS), the covariance matrix of the missing values should have the diagonal elements $\hat{\sigma}_r^2(1 + 1/n_r)$ and all other elements $\hat{\sigma}_r^2/n_r$. Such samples can be generated from a multivariate normal but can also be obtained easily in other ways for example as

$$(n/(n-1))^{1/2}(y_j' - \bar{y}_r + \frac{1}{n_r}\sum_1^r y_i')$$

where y_i' are resampled with SRSWR from the response set R. In some situations it is more natural to impute in this way, since the imputed values will resemble the true values better.

The efficiency of the variance estimation depends not only on m, but also on how long tails the distribution has, i.e., on the kurtosis. If the kurtosis of the empirical distribution is higher than for the normal distribution (i.e. > 3), Bjørnstad's method has a lower efficiency in the variance estimation for fixed m. This comment is quite unimportant since using a larger m can easily compensate a loss in efficiency.

Björnstad's goal is only to estimate a particular linear function, i.e., the population average, and thus he needs only to use the correct variance of the corresponding vector. But to make it complicated, Bjørnstad does not use this covariance matrix, but a diagonal matrix with the diagonal elements $\hat{\sigma}_r^2$. He must later correct for this by replacing a 1 by the number k.

5.4. *Gibbs Sampling and MCMC-methods*

The imputed samples, S_i, are often independent given R, but they do not need to be. In the example with informative response above and many cases of item nonresponse the best way to construct them is to use an MCMC technique like Gibbs sampling. In that case the term, $1/m$, in $(1 + 1/m)$ is not correct. But if many iterations are used the second term vanishes anyway and the method can be used without any problems. When Rubin first presented multiple imputation, MCMC methods required too much computer time to be realistic but nowadays it is possible to use such methods also for surveys with normal sample sizes. There is a lot of work going on in that direction.

I do not really see how this development should fit into the multiple imputation in Bjørnstad's sense.

5.5. *Mass Imputation*

In multiple mass imputation all unobserved values in the population are imputed, even those which are not included in the intended sample. If this is done, there will formally be no sampling variation at all but the whole variance will be imputation variance (i.e., in Bjørnstad's notation: $\hat{V}(\hat{\theta}(s_i^*)) = \bar{V}^* = 0$). The methods of Bjørnstad seem to work here for linear estimates under MAR and MCAR with the same reservations as previously. One further reservation, though, is that the missing fraction f will be almost 1, and that k thus will be extremely large.

Ordinary multiple mass imputation, with or without an MCMC step, has a further advantage since it gives not only the variance of the estimators but also the full distribution. In this sense multiple mass imputation is better, than standard multiple imputation. On the other hand mass imputation is not feasible in practice for many purposes since it requires too much computer time. I cannot see how multiple mass imputation performed according to Bjørnstad can give the whole distribution.

6. Concluding Remark

It is difficult to find any advantage of Bjørnstad's methods over ordinary multiple imputation.

But it is also fair to say that Bjørnstad's methods are often better than many single imputation methods. His methods are certainly alternatives for persons, who are allergic to the word Bayes and do not want to use a classically sound method if the proof of this fact involves the word Bayesian. Also religious fundamentalists who believe that any method which in any way can be connected to Rev. Thomas Bayes must be a heresy may want to consider Bjørnstad's methods.

7. References

Fergusson, T.S. (1967). Mathematical Statistics – A Decision Theoretic Approach. New York: Academic Press.

de Leeuw, E.D., Hox, J., and Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. Journal of Official Statistics, 19, 153–176.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman and Hall.

Received January 2006