Multiple Imputation in Two Stages

Ofer Harel and Joseph L. Schafer

Ofer Harel, University of Washington
F-600 Health Sciences Building, Seattle, WA 98195, ofer@u.washington.edu
Joseph L. Schafer, Pennsylvania State University
325 Thomas Building, University Park, PA 16802, jls@stat.psu.edu.

Abstract

Conventional multiple imputation (MI) (Rubin, 1987) replaces the missing values in a dataset by $m > 1$ sets of simulated values. We describe a two-stage extension of MI in which the missing values are partitioned into two groups and imputed $N = mn$ times in a nested fashion. Two-stage MI divides the missing information into two components of variability, lending insight when the missing values are of two qualitatively different types. It also opens new possibilities for making different assumptions about the mechanisms producing the two kinds of missing values. Point estimates and standard errors from the N complete-data analyses are consolidated by simple rules derived by analogy to nested analysis of variance. After reviewing the theory and practice of two-stage MI, we illustrate the method with a simple analysis of binary variables from a longitudinal survey.

Introduction

Missing values in a dataset may be of two different types. In response to a sensitive question in an interview survey, for example, some participants may simply refuse to answer, whereas others may say, "I don't remember." Other examples include unit versus item nonresponse, planned missingness (e.g. as would arise from double sampling or matrix sampling) versus unplanned missing values, and mortality versus dropout for other reasons. In this article, we explore a two-stage version of Rubin's (1987) multiple imputation in which we impute the first kind of missing value m times; then, for each imputation of the first type, we impute the second type n times, treating the imputed values for the first type as if they were fixed and known.

Imputing in stages has several potential advantages. From a computational standpoint, it is sometimes possible to identify a small set of missing values which, if known, would simplify the process of imputing the rest. An application of this type is described by Rubin (2003), who first completed a monotone pattern by computationally intensive iterative procedure, then imputed the remaining missing values by a less demanding, non-iterative method. Second, two-stage MI allows us to identify the amount of uncertainty in population estimates contributed by each type of missing value, which may have important implications for interpreting current results and planning future studies.

Finally, two-stage MI may allow us to posit different assumptions about the probabilistic mechanisms generating the two types of missing values. Suppose that some values are missing by design but others are missing for reasons beyond the data collectors' control; the missing at random (MAR) condition (e.g., Little and Rubin, 1987) is known to hold for the former but not the latter. With two-stage MI, it may be feasible to impute for unplanned missing values under some alternative non-MAR methods, then impute for planned missing values under an MAR assumption. In that case, however, we would need to clarify what it means for some missing values to be MAR and others to be non-MAR. Making this notion precise requires us to extend Rubin's (1976) concepts of MAR and ignorability to mechanisms that partition the complete data into three parts (observed, missing for one reason, missing for the other reason). A full treatment of that topic has been given by Harel (2003) but is beyond the scope of this paper.

In many cases, two-stage MI can be carried out by repeatedly applying algorithms and software designed for conventional MI, such as the missing-values library in S-PLUS or SAS PROC MI. Once the $N = mn$ imputed datasets exist, they are analyzed by complete data methods, and the results are combined by simple algebraic rules derived by Shen (2000). In the remaining sections, we describe in generic terms the methods for generating the imputations and combining the results; we also discuss how to estimate the amount of missing information due to each set of missing values. We conclude with a simple example involving a binary variable from two adjacent waves of the National Crime Survey.

Imputing in Two Stages

Let $Y_{com} = (Y_{obs}, Y_{mis})$ denote the complete dataset, part of which is observed (Y_{obs}) and part of which is missing (Y_{mis}). In conventional MI, we would typically propose a parametric model for the complete data, say $P(Y_{com} \mid \theta)$, and a Bayesian prior distribution $P(\theta)$ for the unknown model parameters. The *m* multiple imputations, $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(m)}$, are independent draws from the posterior predictive distribution

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid Y_{obs}, \theta)\, P(\theta \mid Y_{obs})\, d\theta ,$$

which reflects uncertainty about θ as well as Y_{mis}. Computational algorithms for producing these draws under a variety of multivariate models are described by Schafer (1997). Software for MI is available from a variety of commercial and non-commercial sources; see Horton and Lipsitz (2001) for a review.

Now suppose that the missing data are divided into two parts, $Y_{mis} = (Y_{mis}^{A}, Y_{mis}^{B})$. To carry out two-stage MI, we first draw *m* independent values $Y_{mis}^{A(1)}, Y_{mis}^{A(2)}, \ldots, Y_{mis}^{A(m)}$ the posterior predictive distribution $P(Y_{mis}^{A} \mid Y_{obs})$. Then, for each of these, we draw *n* conditionally independent values $Y_{mis}^{B(j,1)}, Y_{mis}^{B(j,2)}, \ldots, Y_{mis}^{B(j,n)}$ from $P(Y_{mis}^{B} \mid Y_{obs}, Y_{mis}^{A(j)})$, $j = 1, \ldots, m$. Care must be taken to reflect uncertainty in the parameter θ at each stage. The resulting $N = mn$ completed datasets,

$$Y_{com}^{(j,k)} = (Y_{obs}, Y_{mis}^{A(j)}, Y_{mis}^{B(j,k)}),\ j = 1, \ldots, m,\ k = 1, \ldots n ,$$

are not independent, because each block or nest $Y_{mis}^{B(j,1)}, Y_{mis}^{B(j,2)}, \ldots, Y_{mis}^{B(j,n)}$ contains identical values for Y_{mis}^{A}. Notice, however, that gathering the first completed dataset from each nest, $Y_{com}^{(1,1)}, Y_{com}^{(2,1)}, \ldots, Y_{com}^{(m,1)}$, does give us *m* independent imputations from $P(Y_{mis} \mid Y_{obs})$. Therefore, we can perform two-stage MI by first creating *m* conventional imputations for Y_{mis}, then discarding Y_{mis}^{B} and recreating $n-1$ additional imputations of it for each imputed version of Y_{mis}^{A}.

In this discussion, we have not made use of the response indicators, the set of binary random variables that indicate for each element of Y_{com} whether the item is observed or missing. In conventional MI, we would need to condition upon the response indicators if the missing data were not MAR (Rubin, 1987). With two-stage MI, we might also need to consider the process that bifurcates Y_{mis}; that is, we might need to condition upon the set of three-level indicators that partition Y_{com} into Y_{obs}, Y_{mis}^{A} and Y_{mis}^{B}. Using an extended theory of ignorability presented by Harel (2003), one can show that these three-level indicators can ignored in two-stage MI if (a) Y_{mis} is MAR, and (b) the process that divides Y_{mis} into Y_{mis}^{A} and Y_{mis}^{B} does not depend on any portion of Y_{mis}. Condition (b) allows a data analyst classifies the missing values into groups on the basis of information gleaned from Y_{obs} or from the realized pattern of nonresponse—for example, by taking Y_{mis}^{A} to be a set of missing values that would complete a monotone pattern, as suggested by Rubin (2003). Other conditions under which part or all of the missingness mechanism can be ignored in one or both stages of two-stage MI are described by Harel (2003).

Consolidating Results from Post-Imputation Analyses

After imputation, each of the $N = mn$ imputed datasets from two-stage MI should analyzed as if it had no missing values. Estimates and standard errors from these analyses may then be consolidated using Shen's (2000) extension of Rubin's (1987) method for conventional MI. Let *Q* denote a scalar population quantity to be estimated, and let $\hat{Q}^{(j,k)}$ and $U^{(j,k)}$ denote the point and variance estimates for *Q* calculated from the (j,k) th completed dataset. The overall point estimate for *Q* is simply the grand average

$$\overline{Q}_{..} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{k=1}^{n} \hat{Q}^{(j,k)} = \frac{1}{m} \sum_{j=1}^{m} \overline{Q}^{(j,\cdot)} ,$$

where $\overline{Q}^{(j,\cdot)} = \frac{1}{n} \sum_{k=1}^{n} \hat{Q}^{(j,k)}$. The uncertainty associated with this overall estimate involves three components: the estimated complete-data variance

$$\overline{U}_{\cdot\cdot} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n} U^{(j,k)},$$

the between-nest variance

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\overline{Q}^{(j,\cdot)} - \overline{Q}_{\cdot\cdot}\right)^2,$$

and the within-nest variance

$$W = \frac{1}{m}\sum_{j=1}^{m}\ \frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{Q}^{(j,k)} - \overline{Q}^{(j,\cdot)}\right)^2.$$

The total variance is

$$T = \overline{U}_{\cdot\cdot} + \left(1-\frac{1}{n}\right)W + \left(1+\frac{1}{m}\right)B,$$

and inferences about Q are based on the Student's t-approximation $\left(Q-\overline{Q}_{\cdot\cdot}\right)/\sqrt{\mathrm{T}} \sim \mathrm{t}_{\nu}$ with degrees of freedom

$$\nu = \frac{1}{m(n-1)}\left(\frac{(1-1/n)W}{T}\right)^2 + \frac{1}{m-1}\left(\frac{(1+1/m)B}{T}\right)^2.$$

These rules arises from analysis-of-variance techniques, regarding the $\hat{Q}^{(j,k)}$'s as measurements from a balanced experiment with Y_{mis}^{A} as a random blocking factor. For the special case of $n=1$, the term involving W drops out of the total variance, and the method becomes identical to that of Rubin (1987) for conventional MI with m imputations.

Rates of Missing Information and Relative Efficiency

Shen (2000) did not consider rates of missing information, but these rates are easily estimated from the variance components. The estimated overall rate of missing information for Q is

$$\hat{\lambda} = \frac{B+(1-1/n)W}{\overline{U}_{\cdot\cdot} + B + (1-1/n)W} = \frac{\hat{r}}{1+\hat{r}},$$

where $\hat{r} = \left(B+(1-1/n)W\right)/\overline{U}_{\cdot\cdot}$ is the relative increase in variance due to nonresponse. The estimated rate of missing information due to Y_{mis}^{B} if Y_{mis}^{A} were known is

$$\hat{\lambda}^{B|A} = \frac{W}{\overline{U}_{\cdot\cdot} + W} = \frac{\hat{r}^{B|A}}{1+\hat{r}^{B|A}},$$

where $\hat{r}^{B|A} = W/\overline{U}_{\cdot\cdot}$. The difference $\hat{\lambda}^{A} = \hat{\lambda} - \hat{\lambda}^{B|A}$ estimates the amount by which the missing information would drop if Y_{mis}^{A} became known, and $\hat{\lambda}^{A}/\hat{\lambda}$ is the proportionate reduction in missing information if Y_{mis}^{A} became known. It is possible for $\hat{\lambda}^{A}$ to be negative, in which case the estimate should be set to zero.

Rubin (1987) showed that the relative efficiency of a point estimate based on *m* conventional MI's to a fully efficient one (i.e. based on an infinite number of imputations) is approximately $(1+\lambda/m)^{-1}$, where λ is the overall rate of missing information for Q. For two-stage MI, one can show that the relative efficiency of $\overline{Q}_{\cdot\cdot}$ lies between $(1+\lambda/m)^{-1}$ and $(1+\lambda/(mn))^{-1}$, reaching the upper bound when Y_{mis}^{A} carries no information about Q. The degrees of freedom for estimating the within- and between-nest variance components are approximately equal when $n=2$. In practice, we typically set $n=2$ and then choose m large enough to achieve a reasonable level of efficiency. As with conventional MI, only a few imputations are usually needed to achieve a highly efficient estimate. Unfortunately, the estimated rates of missing information can be rather noisy when *m* is small. Asymptotic results on the variability of these estimated rates are given by Harel (2003).

A Simple Example

The data in Table 1, previously analyzed by Kadane (1985) and Schafer (1997), were obtained from the National Crime Survey conducted by the U.S. Bureau of the Census. Occupants of sampled housing units were asked whether they had been

victimized by crime in the preceding half-year; six months later, occupants of the same units were asked the question again. Of the 765 sampled units, 561 (74%) provided responses at both occasions, 42 (6%) responded only the first time, 38 (5%) responded only the second time, and 115 (15%) did not respond at either time.

Table 1: *Victimization status of housing unit occupants in two successive six-month periods*

	Victimized second period?		
Victimized first period?	No	Yes	Missing
No	392	55	33
Yes	76	38	9
Missing	31	7	115

For simplicity of presentation, we will regard this as a simple random sample from the population of interest. Let Y_t denote victimization status (1=no, 2=yes) during period $t = 1,2$, and let π_{ij} denote the proportion of units in the population with $Y_1 = i, Y_2 = j$. Assuming a multinomial model with ignorable nonresponse, the loglikelihood function is

$$l = 392 \log \pi_{11} + 55 \log \pi_{12} + 76 \log \pi_{21} + 38 \log \pi_{22} + 33 \log \pi_{1+} + 9 \log \pi_{2+} + 31 \log \pi_{+1} + 7 \log \pi_{+2} + 115 \log \pi_{++},$$

where '+' denotes summation over a subscript. Maximum-likelihood (ML) estimates computed by the EM algorithm of Chen and Fienberg (1974) are $\hat{\pi}_{11} = .6971$, $\hat{\pi}_{12} = .0986$, $\hat{\pi}_{21} = .1358$ and $\hat{\pi}_{22} = .0685$ (Schafer, 1997, pp. 42-45).

Multiple imputations for missing data in cross-classified contingency tables under the multinomial model can be generated using a data augmentation (DA) algorithm, a Markov chain Monte Carlo procedure described by Schafer (1997, Chap. 7). This algorithm, which has been implemented in the missing-data library of S-PLUS (Schimert et al., 2001), can also be used for imputation in two stages. To illustrate, we defined Y_{mis}^{A} to be the missing values of Y_1 and Y_{mis}^{B} to be the missing values of Y_2. Using the default prior density

$$P(\pi_{11}, \pi_{12}, \pi_{21}) \propto \pi_{11}^{-.5} \pi_{12}^{-.5} \pi_{21}^{-.5} (1 - \pi_{11} - \pi_{12} - \pi_{21})^{-.5},$$

we generated ten imputations in two stages with $m = 5$ and $n = 2$ in the following manner. First, we generated five ordinary MI's using 100 cycles of DA between imputations. Then, for each of the five datasets, we removed the imputed values for Y_2 and re-imputed them once using another 100 cycles of DA. Frequencies for the ten imputed datasets are shown in Table 2.

Table 2: *Frequencies for victimization status in two periods after two-stage multiple imputation, with estimated odds ratios and differences*

	j = 1		*j* = 2		*j* = 3		*j* = 4		*j* = 5	
	k = 1	*k* = 2	*k* = 1	*k* = 2	*k* = 1	*k* = 2	*k* = 1	*k* = 2	*k* = 1	*k* = 2
No, No	536	531	526	523	528	526	526	530	532	527
No, Yes	70	75	75	78	71	73	75	71	72	77
Yes, No	102	92	104	101	108	102	101	107	103	102
Yes, Yes	48	58	51	54	49	55	54	48	49	50
$\hat{\alpha}$	3.60	4.46	3.44	3.58	3.37	3.89	3.75	3.35	3.52	3.36
$\log \hat{\alpha}$	1.28	1.50	1.24	1.28	1.22	1.36	1.32	1.21	1.26	1.21
$SE(\log \hat{\alpha})$	0.22	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
$\hat{\delta}$	.042	.022	.038	.030	.049	.038	.048	.048	.041	.033
$SE(\hat{\delta})$	.017	.017	.018	.018	.018	.017	.018	.018	.017	.018

Notice that in each of the blocks $j = 1,\ldots,5$, the imputed values for the marginal Y_1 frequencies are constant. Two parameters of interest are the odds ratio $\alpha = \pi_{11}\pi_{12}^{-1}\pi_{21}^{-1}\pi_{22}$ and the change in victimization rate $\delta = \pi_{2+} - \pi_{+2} = \pi_{21} - \pi_{12}$. Estimates of α, $\log\alpha$ and δ from each imputed dataset are shown in Table 2, along with standard errors calculated as

$$SE(\log\hat{\alpha}) = \sqrt{x_{11}^{-1} + x_{12}^{-1} + x_{21}^{-1} + x_{22}^{-1}}$$

and

$$SE(\hat{\delta}) = \sqrt{\frac{x_{12}}{x_{++}^2}\left(1 - \frac{x_{12}}{x_{++}}\right) + \frac{x_{21}}{x_{++}^2}\left(1 - \frac{x_{21}}{x_{++}}\right) + 2\frac{x_{12}x_{21}}{x_{++}^3}},$$

where x_{ij} is the sample frequency of $Y_1 = i, Y_2 = j$ (e.g. Agresti, 1990). Combining these quantities by Shen's rules, the overall estimate of $\log\alpha$ becomes $\overline{Q}_{..} = 1.29$ with a standard error of $\sqrt{T} = 0.23$, and the estimate of δ becomes $\overline{Q}_{..} = 0.38$ with a standard error of $\sqrt{T} = 0.19$. Although these estimates are based only on a small number of imputations, the large values for the degrees of freedom ($\nu = 345$ and $\nu = 282$, respectively) indicate that the loss of precision relative to estimates based on more imputations is slight. For comparison, we computed asymptotic standard errors by evaluating the Hessian of the actual loglikelihood at the ML estimate; this loglikelihood method gives $\log\hat{\alpha} = 1.27$ with a standard error of 0.26, and $\hat{\delta} = .037$ with a standard error of .020.

Although the estimates and standard errors from such a small number of imputations are reliable, the estimated rates of missing information are not. To estimate these rates very precisely, we repeated the procedure to create $m = 500$ blocks of $n = 2$ imputations each. The estimated parameters, standard errors and rates of missing information from this larger simulation are displayed in Table 3.

Table 3: *Estimated parameters, standard errors and rates of missing information for crime victimization data from multiple imputation with m = 500 and n = 2.*

	est	SE	$\hat{\lambda}$	$\hat{\lambda}^{B\|A}$	$\hat{\lambda}^{A}$	$\hat{\lambda}^{A}/\hat{\lambda}$
$\log\alpha$	1.27	0.25	.27	.21	.06	.24
δ	.038	.020	.21	.13	.08	.39

The rates of missing information are intuitively reasonable. First, consider the log odds ratio, which measures the strength of association between Y_1 and Y_2. Because 26% of the units had missing values for one or both of these variables, we would expect the overall rate of missing information to be approximately 0.26, and indeed it is ($\hat{\lambda} = .27$). Because the odds ratio can be regarded as a property of the conditional distribution of Y_2 given Y_1, recovery of the missing values of Y_1 should provide little additional information about this parameter, so we would expect $\hat{\lambda}^{A}$ to be much smaller than $\hat{\lambda}^{B|A}$. The difference δ, however, is a function of the Y_1 and Y_2 marginal distributions, so we might expect the overall rate of missing information for this parameter to be something like the proportion of units for which Y_1 and Y_2 are both missing, plus one-half the proportion of units for which one of the two measurements is missing. Indeed, the estimate ($\hat{\lambda} = 0.21$) is very close to .15 + (.05+.06)/2. If Y_1 were filled in, Y_2 would still be missing for 21% of the units, and indeed $\hat{\lambda} = .13$ is close to one-half of .21.

Discussion

The simple example presented here is not intended to represent a typical application of two-stage MI; we provide it merely to illustrate the method and to demonstrate the reasonableness of its results. The method was originally proposed by Shen (2000) for computational efficiency when imputation of Y_{mis}^{A} is costly but Y_{mis}^{B} given Y_{mis}^{A} is cheap. Situations of that type do exist (e.g. Rubin, 2003). However, we believe that the most valuable contribution of two-stage MI is that opens up many new possibilities for interesting and innovative data analysis. Some of these possibilities are listed below.

Questionnaires with planned missingness (Raghunathan and Grizzle, 1995) and longitudinal data-collection schedules with intentionally dropped occasions may reduce respondent burden and can be very cost-effective; with two-stage MI, one can isolate the effects of planned and unplanned missing values to inform decisions about the design of current and future studies.

Analysis of longitudinal data with nonignorably missing values has received great attention in recent years (e.g. Little, 1995). In real applications, it is often reasonable to view the missing values as a mixture, with some missing for reasons closely related to the phenomena being measured and others missed for unrelated reasons. With two-stage MI, we can potentially separate the effects of the two types of missing values. As shown by Harel (2003), however, what it means for some values to be ignorably missing and others to be nonignorably missing needs to be made precise.

Another important set of applications pertains to data with both measurement errors and missing values. Viewing observed items as imperfect measures of unobservable or latent true scores has been fundamental to the social sciences for decades, and is the central idea of item-response theory, factor analysis and latent-class modeling. Despite these models' long history, the role of missing values in the manifest items remains largely unexplored. We believe that two-stage MI will prove useful for investigating not only the relative contributions of response error and nonresponse to overall uncertainty, but how the nonresponse and measurement processes interact.

Finally, two-stage MI can also be used to measure the expected increase in information if a single missing datum or a group of missing values were suddenly observed. If this type of analysis were implemented in a survey during the data-gathering process, data collectors could identify the nonrespondents who, if converted, would yield the greatest increase in information about various population quantities of interest, and then focus their efforts for nonresponse followup on these units.

References

Agresti, A. (1990) *Categorical Data Analysis*. J. Wiley & Sons, New York.

Chen, T.T. and Fienberg, S.E. (1974) Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 32, 133-144.

Harel, O. (2003) Strategies for data analysis with two types of missing values. Ph.D dissertation, Department of Statistics, Pennsylvania State University, University Park, PA.

Horton, N.J. and Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *American Statistician*, 55, 244-254.

Kadane, J.B. (1985) Is victimization chronic? A Bayesian analysis of multinomial missing data. *Journal of Econometrics*, 29, 47-67, correction 35, 393.

Little, R.J.A. (1995) Modeling the drop-out mechanism in repeated-measure studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York.

Raghunathan, E.R. and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

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

Rubin, D.B. (2003) Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3-18.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Shen, Z.J. (2000). Nested multiple imputation. Ph.D dissertation, Department of Statistics, Harvard University, Cambridge, MA.