Practical suggestions on rounding in multiple imputation

Recai M. Yucel and Alan M. Zaslavsky
Recai M. Yucel
Department of Biostatistics and Epidemiology,
School of Public Health and Health Sciences
University of Massachusetts, Amherst, 01003, MA

Key Words: Missing data, multivariate normal, multiple imputation, categorical data imputation, missing data software.

Abstract:

In the last decade, substantial progress has been made on methods for imputation of missing data. Modern imputation methods have become widely available for practitioners through software products such as S-Plus 6.0 (Schimert, Schafer, Hesterberg, Fraley, and Clarkson 2000), SAS PROC MI (SAS 2001), and SOLAS (2001) . The key idea underlying most of these methods is to impute missing values by random draws from the conditional distribution of the missing data given the observed data. In practice, many of these methods (e.g. the "norm" module of the "missing" library in S-Plus 6, SAS PROC MI) impose a multivariate normal distribution on the incompletely observed variables. When these variables are not normally distributed but rather categorical (binary or ordinal), practitioners are often advised to round the imputed value, typically drawn from a continuous multivariate normal, to the nearest integer (or category) that is within defined region. In this paper we provide some practical suggestions for rounding with binomial or ordinal variables that will enable practitioners to use commonly available software while providing imputed datasets that will lead to inferences that are less biased, in the sense that marginal distributions are more accurate.

1. Introduction

Modern methods and software have been increasingly popular for analyzing incomplete datasets by multiple imputation inference. The key idea underlying these methods is to impute missing values by random draws from the conditional distribution of the missing data given the observed data. For convenience, some methods (e.g. the "norm" function of the "missing" library in S-Plus 6, SAS PROC MI) impose a fully parametric multivariate normal distribution on the variables that are incompletely observed.

In problems involving categorical data with binary or ordinal scale, practitioners are often advised to employ methods assuming a multivariate normal distribution as an approximation and to round the resulting values to nearest observed values (Schafer 1997). This approach often leads to biased inference. In applications with ordinal variables that take many values and have nearly symmetrical distributions, bias due to rounding might be negligible (Schafer 1997). However, where the distributions of such are far from symmetry or oddly shaped (e.g. multimodal), naïve rounding could lead to biased inferences for marginal distributions and correlation structure.

In this paper, we illustrate a more principled approach to using normal approximations in imputing incomplete datasets containing binary or ordinal variables. Our goal is to make use of methods that are well established in the missing data literature and widely used in missing data problems. Our objective for imputation methods with categorical data is to create MAR data (Little and Rubin 1987) and reimpute so that the distribution of the imputed data will be similar to that of original data. Similarity is defined in relevant ways including the marginal distribution of the categorical variables and joint distributions with the other variables. Specifically, we seek to make the marginal distributions resemble the correct (and known) marginal distribution.

A brief background of previous work, including previous techniques for imputing binary variables along with commonly used statistical software packages for imputation, is provided in Section 2. Section 3 outlines our strategy for imputing binary or ordinal variables. Section 4 summarizes results obtained from a limited simulation study. Finally, Section 5

Recai M. Yucel is Assistant Professor of Biostatistics, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA. Alan M. Zaslavsky is Professor of Health Care Policy (Statistics), Department of Health Care Policy, Harvard Medical School, Boston, MA. This research was supported by a research program by National Center for Health Statistics and American Statistical Association.

discusses the advantages and limitations of our work.

2. Previous work

Literature on missing data has developed tremendously in the last two decades with the advent of advanced statistical computation techniques. The pioneering work by Rubin (1978) and Little and Rubin (1987) initiated the wide spread use of missing data techniques, especially inference by multiple imputation Rubin (1978). These techniques have also been implemented in a variety of general purpose statistical software packages (e.g. S-Plus 6.0 (Schimert, Schafer, Hesterberg, Fraley, and Clarkson 2000), SAS PROC MI (SAS 2001)) enabling many researchers actively use such missing data techniques.

Some of the studies focused on the imputation techniques for binary data. Rubin (1987) described multiple imputation (MI) for an incompletely observed binary variable through logistic regression. Rubin (1987) and Little (1988) also suggested a "predictive mean matching method". Schafer (1997) suggested a variety of methods including an approximation by normal distribution and rounding to the nearest integer. He also described pure categorical data methods, although these lead to practical problems such as zero-count cells when the number of categorical variables is high. These methods are implemented in the S-Plus 6.0 library 'missing' (Schimert, Schafer, Hesterberg, Fraley, and Clarkson 2000), including modules corresponding to continuous normal, categorical and mixed incomplete data.

Binary variables can be imputed using techniques for continuous data. Rubin and Schenker (1986) suggested conducting a fully normal imputation and then dichotomizing with a cut-off value that is MLE of the Bernoulli probability. Horton, Lipsitz, and Parzen (2003) suggested normal imputation of the binary variable without rounding to obtain an unbiased estimate of the mean (probability of success) of the binary variable. These studies only considered a single incompletely observed variable. More recently, Bernaards, Belin, and Schafer (2005) evaluated the robustness of multivariate normal approximation for imputation of binary data, suggesting a rule for calculating a "cut-off" value based on a normal latent variable distribution underlying the binomial variable.

Several statistical software packages specifically target missing data problems, either estimating particular parameters in the presence of missing data or serving a more general purpose through multiple imputation. SOLAS software supports a predictive mean model, using the closest observed value to the predicted value and propensity score models for missing continuous variables and discriminant models for missing binary and categorical variables. Often in practice, predictive mean matching is used for binary and ordinary variables. NORM (Schafer 2000), also available as a module in "library missing" of S-Plus 6 and SAS PROC MI (with MCMC option), assumes a multivariate normal distributions as the underlying imputation model. Users of these programs are often advised to perform a preliminary analyses to assess distributional structures of the data and perform necessary transformations to improve "normality" assumption. For binary or ordinal data, multivariate normal approximation is used and as a post-imputation procedure, i.e. imputed values that are out of the range are rounded off to the nearest observed value.

Our study is driven by the desire to use existing and well-established software for imputation under the normal model for imputation of binary or ordinal data by simple data manipulations.

3. Methods

3.1 Notation

Our notation is a variation that is used commonly in the missing data literature. Let (X_i, Y_i) denote a vector of characteristics for cases $i = 1, \ldots, n$, where Y_i is a binary variable; the complete data matrix is (X, Y). Let Y_{obs} and Y_{mis} denote the observed and missing portions of Y and assume that Y is ordered to separate observed and missing values so that $Y = (Y_{obs}, Y_{mis})$, and $Y_{obs} = (y_1, \ldots, y_{n_{obs}})$; $Y_{mis} = (y_{n_{obs}+1}, \ldots, y_n)$. The ultimate goal in inference by multiple imputation is to replace Y_{mis} by simulated or "imputed" values of Y_{mis}, drawn from their underlying joint posterior distribution $P(Y_{mis} \mid Y_{obs})$. We let $\mathbf{X}$ denote an $n \times p$ data matrix, fully observed or incomplete.

In the following section we outline our strategy on how to round imputed values in Y under a multivariate normal imputation performed on y and possibly on $\mathbf{X}$.

3.2 Methods for binary variables

Our basic strategy is to perform a multivariate normal imputation performed on y_{mis} (and possibly on $\mathbf{X}$, if it is incompletely observed), and then to round. Such imputations assume that each (X_i, Y_i) is sampled from a multivariate normal distribution with mean μ and covariance Σ. Imputations are obtained by iterative methods that alternate between two steps: (1) draw a value of μ^*, Σ^* from their pos-

terior distribution (typically from a Normal-Wishart family, given normal-inverse-Wishart prior distributions), and (2) fill in the missing values (X_{mis}, Y_{mis}) under their distribution conditional on observed data (X_{obs}, Y_{obs}) and the parameters μ^*, Σ^* . Repeating these steps results in imputed values that are draws from $P(Y_{mis}, X_{mis} \mid Y_{obs}, X_{obs})$.

We use simulation to obtain the cutoff value that will be used in rounding the imputed Y_{mis} in such a way that observed marginal distribution of y and its relationship with other variables are preserved. The imputed variable y_i^* has the following form:

$$y_i^* = \begin{cases} 1 & y_i^{**} > y_c,\ i \in mis \\ 0 & \text{otherwise,} \end{cases} \qquad (1)$$

where y_i^{**}, $(i \in mis)$ is the imputed value under the continuous multivariate normal distribution and y_c is the *cut-off* value. We determine y_c by executing the following steps:

1. Create MAR data by duplication: Let $X_{dup} = \{X, X\}$ and $Y_{dup} = (Y_{obs}, Y_{mis}, Y_{obs(mis)}, Y_{mis})$ denote the duplicated datasets; $Y_{obs(mis)}$ means Y_{obs} with observed values turned to missing.

2. Impute $\{Y_{mis}, Y_{obs(mis)}, Y_{mis}\}$ (and possibly X_{mis}) under a multivariate normal model.

3. y_c is calculated in such a way that the fraction of imputed 1s in $Y^*_{obs(mis)}$ equals the fraction of 1s in the observed data Y_{obs} for the same cases.

4. Proceed with imputing Y_{mis} using y_c following (1).

These steps are depicted in Figure (1) with a simplified example where a binary variable has %70 ratio of "1"s in the observed data. Imputations are created following (1)–(4), the imputed values are ordered in the $Y_{obs(mis)}$ (shown as $y_{imp(duplicate)}$) for notational convenience, and finally the values in $Y_{mis(imp)}$ that are bigger than y_c are imputed as 1's.

Duplication of the data is a strategy for calibration of the cutoff for rounding that is guaranteed to replicate the observed marginal distribution of y. Therefore, the performance of the procedure should be evaluated by looking at the higher level relations (e.g. correlations). Employing a well-established multivariate normal imputation technique is advantageous from this standpoint as it preserves two-way relationships.

3.3 Extension to ordinal variables

Now let $Y = (Y_{obs}, Y_{mis})$ denote the observed and missing parts of n measurements on an ordinary

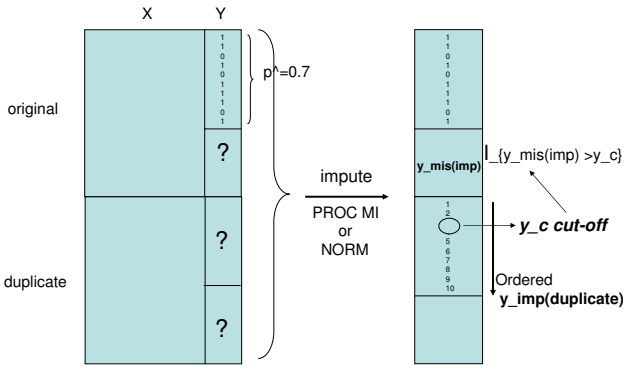

Figure 1: Computation of the *cut-off* value: y_c

variable y, where y_i may take values $1, 2, \ldots G$. Furthermore let $p_{ig} = P(Y_i = g)$. We now have

$$y_i^* = g \quad \text{if} \quad y_{c,g-1} < y_i^{**} < y_{cg},\ i \in mis \qquad (2)$$

where $i = 1, 2, \ldots, n$. (We define $y_{c0} = -\infty$, $y_{cG} = +\infty$.) Determining the cut-off points $y_{c1}, y_{c2}, \ldots, y_{c,G-1}$ proceeds in a fashion similar to Section 3.2. Evaluation criterion is also similar (via duplicating and association measures).

4. Simulation Study

Our simulation study was designed to assess the performance of the methods suggested in Section 3. The scenario of the simulation was the following:

1. The variables are a single continuous variable X and binary variable Y.

2. Generate data

$$\begin{aligned} X_i &\sim N(0,1), \\ Y_i \mid X_i = x_i &\sim \text{Bernoulli}(\text{logit}^{-1}(\alpha + \beta x_i)) \end{aligned}$$

3. The missingness mechanism is MAR:

$$r \sim \text{Bernoulli}(\text{logit}^{-1}(\alpha_r + \beta_r x_i))$$

   The parameters α_r, β_r are chosen to be 0.3 and -0.3, respectively, so that on average 57% is set to be missing on Y.

4. Draw $10,000$ bivariate data points for each set of simulation parameter values.

5. Vary α, β to obtain Y variable with means given in Table 1.

Table 1: Simulation conditions (values of α, β)

α	β	μ_Y
1.4	3.1	0.90
1.4	2.5	0.85
1.1	1.9	0.80
1.1	1.5	0.75
1.1	1.5	0.70
1.2	0.95	0.65
1.0	0.6	0.60
1.4	0.45	0.55
1.4	0.2	0.50
1.4	-0.09	0.45
1.4	-0.35	0.40
1.3	-0.9	0.30
1.4	-1.6	0.20
1.4	-2.1	0.15
0.7	-2.8	0.10

Each incomplete data across the simulations were imputed using three methods:

- The original multivariate normal imputation, without rounding;
- "Naïve" rounding using cut-off point 0.5;
- Our method, in which the cutpoint is estimated to calibrate the marginal mean of Y for the observed cases.

For each method, we compared the means of the imputed Y values to those of the simulated missing values (Figure (2)). We also compared the correlations of the imputed Y values with X to the corresponding correlations for the simulated missing values (Figure (3)).

Imputed- and missing-data means agree closely with our method, reflecting the success of our calibration approach. The same is true for normal imputation, which is nearly unbiased although the imputed values are not representative of the potential values of the binary variable. The method based on naïve rounding shows some bias for values of the mean that are close to 0 or 1. For the cases where μ_y was set to be 0.9, 0.85, 0.15 and 0.07 for example, naive method had means of 0.95, 0.89, 0.18 and 0.05, respectively, whereas our method estimated the mean as 0.91, 0.87, 0.16 and 0.07, respectively.

It is more difficult to anticipate the patterns in the correlations, since our method does not explicitly calibrate this aspect of the data. However, our method performed reasonably well capturing the true correlations even in the extreme cases. Unrounded imputations did just as good, the correlations were on the target with true correlations. Finally the naive method resulted with correlation estimates that were somewhat misleading in some of the cases.

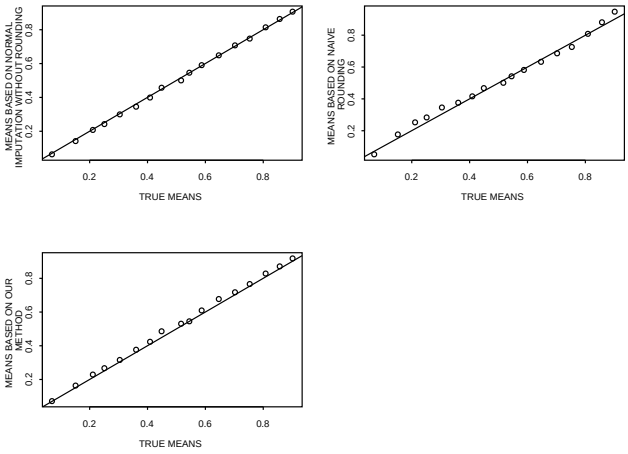

Figure 2: Comparison of the means of the Y_{mis} with estimates from imputed data under three methods.

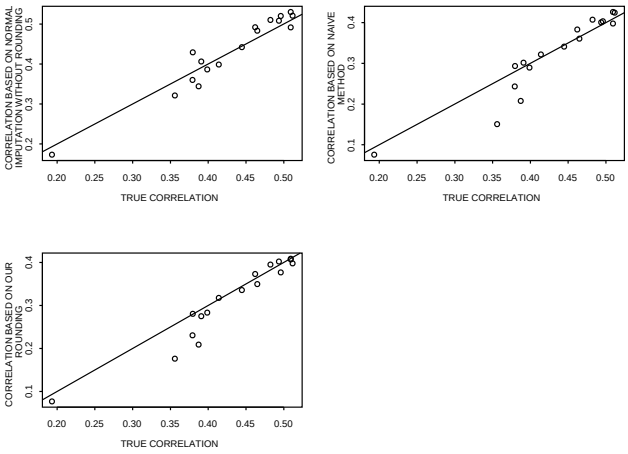

Figure 3: Comparison of the correlations of the Y_{mis} with X with estimates from imputed data under three methods.

5. Discussion

Our goal in this work was to make use of well-established methods for imputing missing data under a multivariate normal distribution to obtain usable imputation for non-normal data, specifically binary or ordinal data. We believe this is an important motivation and contribution because these algo-

rithms are widely available and are used commonly for missing data problems. With minor data manipulations our methods can easily be adopted. The limited simulation study shows results that support our proposed methods. It is important to notice that commonly assumed missing data mechanism (MAR) is more plausible by including as much information as we can (by including X in creating imputations). Finally the evaluation criterion is to create MAR (duplication) and reimpute then the distribution of the imputed data will be similar to that of original data.

More systematic investigation is needed to determine how well the proposed methods work in preserving data structure (e.g. correlations), through more extensive simulation studies. We are currently working on developing more rules for determining cut points by calibrating on functions such as regression coefficients that has a potential on preserving these relationships.

A second limitation is that we only limited our work to a single non-normal variable along with a set of multivariate normal variables. Extending this to multivariate non-normal variables is an important step as most real data applications involve multivariate non-normal data. Another potentially important contribution would be to consider how to adopt the current software to handle datasets with incompletely observed unordered categorical data in a manner similar to ours.

Finally, we would like to strengthen the theoretical basis for our methods by developing the relationship of our approach to a more completely specified multivariate normal latent variable model for the mixed continuous and categorical outcomes.

References

Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2005), "Robustness of a multivariate normal approximation for imputation of incomplete binary data," *Statistics in Medicine (under review)*.

Horton, N. J., Lipsitz, S. R., and Parzen, M. (2003), "A Potential for Bias When Rounding in Multiple Imputation," *The American Statistician*, 57, 229–232.

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

Little, R. J. A. (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, 6, 287–297.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse," in *ASA Proceedings of the Section on Survey Research Methods*.

— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons.

Rubin, D. D. and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

SAS (2001), *SAS/Stat User's Guide, Version 8.2*, NC, USA: SAS Publishing.

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

— (2000), *Multiple imputation of incomplete multivariate normal data*, The Pennsylvania State University, PA, USA.

Schimert, J., Schafer, J., Hesterberg, T., Fraley, C., and Clarkson, D. (2000), *Analyzing Data with Missing Values in S-Plus*, Seattle, WA: Data Analysis Products Division, Insightful Corporation.

SOLAS (2001), *Statistical solutions, Inc.*, Cork, Ireland.