Imputation of categorical variables using Gaussian-based routines

Recai M. Yucel* Yulei He† Alan M. Zaslavsky‡

Abstract

The multivariate normal (MVN) distribution is arguably the most popular parametric model used in imputation and is available in commonly used software packages (e.g. SAS PROC MI). When the incompletely-observed variables include nominal variables, practitioners often apply techniques such as creating a distinct "missing" category or disregarding the nominal variable from the imputation process, both of which may lead to biased results. In this work, we propose practical rounding rules to be used with the existing MVN-based imputation methods, allowing practitioners to obtain usable imputation with small biases. These rules are calibrated in the sense that values re-imputed for observed data have distributions similar to those of the observed data. A simulation study demonstrating the advantages of this approach is presented.

Key Words: Nominal data imputation, missing data, multiple imputation, missing data software, rounding

1. Introduction to rounding in MI

With the growing availability of software, multiple imputation (MI) under multivariate normal (MVN) distribution has emerged as a popular inferential tool in the analysis of incomplete data. Some of the most commonly used software include the missing library of S-Plus (Schimert et al. 2000), the norm library of R (Schafer, 1997), and SAS PROC MI (SAS Institute 2001). These software commonly implement specialized computational techniques for sampling from the implied posterior predictive distribution under MVN and a set of priors. Sampled data points are regarded as multiple imputations as depicted in Figure 1.

One of the major concerns in adapting the software for MI under MVN assumption is the plausibility of the normality. In many real-life problems with categorical variables measured on nominal scale, the MVN assumption is clearly violated. In the previous studies, two distinct solutions have been proposed to remedy this problem. First solution develops a strict categorical model used as a basis for imputation (see Schafer, 1997, Ch. 6). Second solution has mostly been limited to a binary case and pertains to modifying techniques used for continuous data to impute the categorical variables (Bernaards et al. 2006; Yucel, Yulei and Zaslavsky, 2008). Because we focus on using the existing MVN-based procedures to impute larger class of nominal variables (ordered or unordered), the following discussion is limited to the latter approach.

In most practices of rounding, imputed values for binary or ordinal variables are rounded off to the nearest observed values (Schafer 1997, Ch. 5). This approach can lead to biased estimates (e.g. means or correlations) when the binary or ordinal

*Recai M. Yucel is Assistant Professor of Biostatistics at the Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, SUNY, One University Place, Rensselaer, NY 12144.

†Yulei He is Assistant Professor of Biostatistics at the Department of Health Care Policy, 180 Longwood Avenue, Harvard Medical School, Boston, MA.

‡Alan M. Zaslavsky is Professor of Statistics and Health Care Policy at the Department of Health Care Policy, 180 Longwood Avenue, Harvard Medical School, Boston, MA.

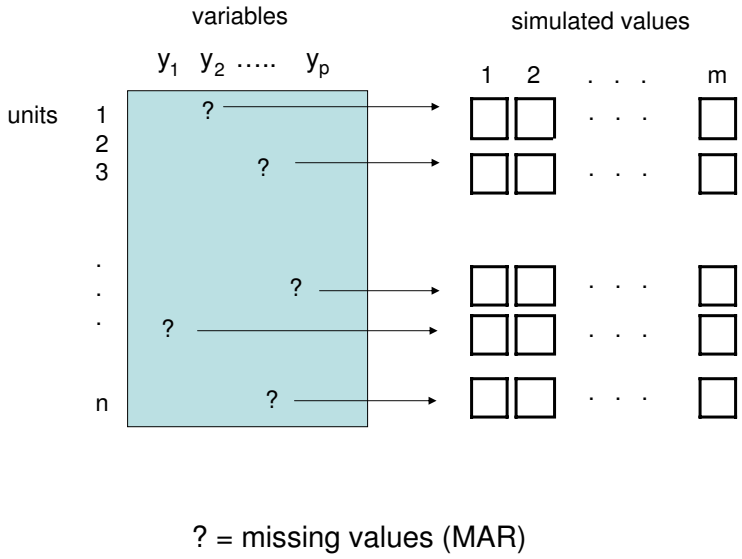

Figure 1: Depiction of MI in practice

variables whose distributions are far from symmetry or oddly shaped (e.g., multimodal). The disadvantages of such methods are likely to worsen as the amount of missing data increases. Horton et al. (2003) evaluated imputation of a Bernoulli random variable by rounding a continuous imputation to the nearest value of 0 or 1 (i.e., by a cutoff value of 1/2). They found that this caused substantial bias and suggested a normal imputation of the binary variable without rounding to obtain an unbiased estimate of the mean (probability of success). Although unbiasedness is an important property, the latter study's strategy of no rounding does not impute data of the original type as might be desired by practitioners whose intended complete-data analysis is tailored to binomial data. Bernaards et al. (2006) evaluated the robustness of the multivariate normal approximation for imputation of binary incomplete data and suggested a rule for calculating a cut-off value, based on a normal approximation to the binomial distribution. They assumed that the variance of the imputations equals the binomial variance, which might not be true if there is a strong predictor that adds variance to the predictions.

How can one use the existing MVN-based procedures to imputed non-binary variables? As in the binary case, naive rounding can lead to substantial biases, and other methods suffer from the similar shortcomings stated above. Another significant drawback of the current MVN-based methods is the inability of handling nominal variables. Although not recommended (Little and Rubin, 2002), practitioners are often forced to create qualitatively different category.

The goal of this work is to provide a widely-usable solution that facilitates the current MVN-based MI software in developing rules for rounding nominal variables. The principle of this solution is the one that extends our recent work (Yucel, Yulei and Zaslavsky, 2008, from hereon referred as YYZ) where imputed values are calibrated to resemble the observed distribution. While the MVN is the underlying assumption of the commonly-used software, the main assumption, here, is the existence of the working method to impute continuous values for the nominal variables,

hence the methods described here do not assume a particular distribution.

The remainder of this article is organized as follows. Section 2 describes our rounding strategy based on calibration of the marginal distribution and states how it can be implemented in ordered nominal or ordinal and unordered nominal cases. Section 3 summarizes our findings from a limited simulation study. Finally, Section 4 discusses the strengths and limitations of this approach.

2. Calibration-based rounding strategy

2.1 Models & assumptions

Let Y denote the nominal variable of interest subject to missingness, and let $1, 2, \ldots, G$ denote the values assumed by Y. Further, let X denote the covariates (univariate or multivariate) in the imputation model, where Y consists of observed and missing values, Y_{obs} and Y_{mis}, respectively. We assume that X either is complete or has been completed by a plausible imputation method. The goal in inference by multiple imputation is to replace Y_{mis} by random draws from its posterior predictive distribution $P(Y_{\text{mis}} \mid Y_{\text{obs}}, X)$. Another set of notation that is necessary to introduce pertains to the missingness mechanism. Following a standard notation, let R be the indicator variable for observation of Y, with $R = 1$ for observed Y and $R = 0$ for missing Y . Under a missing at random (MAR) missingness mechanism, the probability that any data value is missing may depend on quantities that are observed but not on quantities that are missing: $P(R \mid Y_{\text{obs}}, Y_{\text{mis}}, X) = P(R \mid Y_{\text{obs}}, X)$. Under a missing completely at random (MCAR) mechanism, missingness is independent of both observed and missing values: $P(R \mid Y_{\text{obs}}, Y_{\text{mis}}, X) = P(R)$.

The notation below builds upon the assumption that there is a working structure that determines the form of this posterior predictive distribution. We let Y_C^* denote a variable imputed under a continuous model $P(Y_{\text{mis}} \mid Y_{\text{obs}}, X)$ which will be rounded off to a variable that takes values in the desired scale (e.g. ordinal or unordered nominal), which will be denoted by Y^*. The rounding will proceed using a set of rules determined by calibration.

2.2 Calibration idea

The main goal of the calibration is to create imputed values with similar distribution to that of the observed values. Employment of a multivariate distribution to establish this goal can also establish the secondary goal: preserve the relationships with other potentially important variables to the (post) analyses. To implement a method with such properties, YYZ simply duplicated and appended the copy to the original data, and intentionally set Y to missing in the second copy and then generated imputations Y^* of the missing copy of Y. YYZ demonstrated this method on a binary case and derived the asymptotic biases of the method, which are largely tolerable in problems with modest missingness.

In the context of developing rounding rules, the motivation of duplication is as follows. Under a given imputation model, the sampled or imputed values of observed data, Y_{obs}^*, conditional on the covariates, X, are seen as coming from the same distribution generating Y_{obs}. Then the calibration proceeds with respect to a statistic $S(X, Y_{\text{obs}})$. In other words, we can develop rounding rules under the assumption of $S(X, Y_{\text{obs}})$ is a realization from the distribution of the statistic evaluated under the imputation model, $S(X, Y_{\text{obs}}^*)$. A natural choice in the problem

of imputation of a nominal variable is the probability of categories or its joint distribution with X.

2.3 Marginal calibration for ordinal variables

The strategy for extending the methodology of YYZ into ordinal variables uses the same principle: obtain the set of cut-off values that will be used in rounding the imputed and calibrated $Y^*_{\text{mis,C}}$ in such a way that after rounding Y_{obs} will have a marginal distribution consistent with the observed data Y_{obs}. The algorithm that implements this principle is given by

1. $X_{\text{dup}} = \{X, X\}$ and $Y_{\text{dup}} = \{Y_{\text{obs}}, Y_{\text{mis}}, Y_{\text{obs(dup)}}, Y_{\text{mis(dup)}}\}$ are the duplicated datasets; $Y_{\text{obs(dup)}}$ means Y_{obs} with observed values turned to missing
2. Impute the missing values in Y_{dup} under MVN (or any chosen continuous distribution) and denote the imputed continuous values as $Y^*_{\text{mis,C}}$, $Y^*_{\text{obs(dup),C}}$, and $Y^*_{\text{mis(dup),C}}$
3. The cut-off values $c_1, c_2, \ldots, c_G$ are determined so that the relative frequencies of each category in $Y^*_{\text{obs(dup)}}$ equals the relative frequency in Y_{obs}.
4. Use $c_1, c_2, \ldots, c_G$ to round $Y^*_{mis,C}$ to Y^*_{mis}:
$$Y^*_{\text{mis}} = g \text{ for } c_{g-1} \leq Y^*_{mis,C} < c_g, \tag{1}$$
where $c_0 = -\infty$ and $c_G = \infty$

Note that $c_1, c_2, \ldots, c_G$ can be any value between the appropriate order statistics of $Y^*_{\text{obs(dup),C}}$. To reflect uncertainty in the practice of multiple imputation, one can draw the cut-off values from a uniform distribution defined in the appropriate interval given in (1).

2.4 Marginal calibration for nominal variables

When the variable of interest Y is an unordered nominal variable (e.g. race), even approximately, current techniques assuming MVN can not be utilized. The calibration routine introduced above can be used with slight modifications to reflect the unordered nature and proceeds with a set of dummy variables and rounding rules are based on cumulative probabilities. Same algorithmic idea based on duplication, imposition of missingness and calibration apply on each of the dummy variable:

1. First step of the algorithm given above differs slightly. Here we create a set of dummy variables in Y_{obs} indicating the underlying category: $I_{Y_{\text{obs(i)}}=g} = 1$. This will thus turn the input for the MVN-based imputation procedure into a slightly different format:
$$X_{\text{dup}}, I_{Y_{\text{obs(i)}}=1}, I^{\text{copy}}_{Y_{\text{obs(i)}}=1}, \ldots, I_{Y_{\text{obs(i)}}=g-1}, I^{\text{copy}}_{Y_{\text{obs(i)}}=g-1},$$
where the subscript "copy" indicates the duplicate copy of the indicator set to missing in the second copy,
2. Impute the missing values in all of binary variables defining the nominal variable Y,
3. Compute the cut-off value using YYZ method for each of the binary variable sequentially and independently in such a way that the underlying ratio in the first copy is same as in the second copy.

2.5 Practical extensions

As noted by YYZ, calibration is a very general approach to validation of an imputation model and can be regarded as a version of the posterior predictive check (Gelman et al. 2004, pp. 159172; Gelman et al. 1996), commonly used in assessing the fit of Bayesian models. The use of calibration in the context of imputation combines the model-based imputations with an ad hoc rounding step with $g-1$ free parameters (not included in the imputation model). Thus its use is bit restrictive and negligible biases are unavoidable in some statistics. However, the idea of calibration is very attractive for practitioners with well-established tools such as MVN-based imputation software.

One potentially very useful extension of calibration is when the incomplete-data problem is further complicated by clustering. Such complications occur in multistage surveys conducted on subjects clustered within naturally occurring groups or longitudinal studies of subjects. Similar to its cross-sectional counterparts, categorical data imputation has often been done based on approximations using MVN-based routines (Schafer and Yucel, 2002 and Carpenter and Kenward, 2008). As argued above, in some instances, such approximations can be harmful to the inferences in certain circumstances.

Extending the calibration-based rounding using the MVN-based imputation in clustered designs is relatively straightforward. Consider, for example, imputing an ordinal variable when it is realized under clustered data. The algorithm given in Section 2.3 requires a working method sampling from the underlying $P(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta)$, where θ is the set of unknown parameters of the imputation model. Such sampling procedure is relatively easy under MVN, but complicated by the handling of clusters via random-effects. One can visualize this as $m-$repetition of the algorithm given in Section 2.1, leading to mean structure to be preserved across the clusters.

3. Simulation Study

We designed a limited simulation study to assess the performance of the suggested calibration-based routines allowing the users to utilize MVN-based imputation techniques. Here we report the ordinal case, more comprehensive results are currently being developed. Our simulations consisted of the following specifications:

1. Simulate random variables X and Y:

$$\begin{aligned} X &\sim N(0,1) \\ U \mid X &\sim N(\alpha + \beta X, \beta^2 + 4), \end{aligned}$$

   so that a continuous covariate X is used to obtain a continuous variable U. Using the quantile of U leads to the simulation of Y given X. We specifically chose a scenario where Y was highly skewed to emphasize the disadvantages of the current practices.

2. On average 45% of the $Y-$ values were set to missing under MAR mechanism:

$$R \sim \text{Bernoulli}(\text{logit}^{-1}(\alpha_R + \beta_R X)), \tag{2}$$

   β_R varied to assess performance under MCAR and deviations from MCAR to MAR. These varying missingness mechanism are shown in the x-axis of the plots given in Figure 2 & 3.

3. Total data points simulated was set to 10,000 to reduce the simulation error. Similar performances were also seen in higher numbers.

MVN-based imputation routines as implemented in R package norm was used to impute missing values in each incomplete data across the simulations, and three different rounding methods were compared. First method uses no rounding leading to unusable but unbiased estimates in the univariate analyses. This method was first suggested by Horton et al. (2003) in the binary case. Second method pertains to rounding to the nearest integer (Schafer, 1997). Final method is the calibration-based rounding. The performance of these methods are presented in Figures 2 and 3.

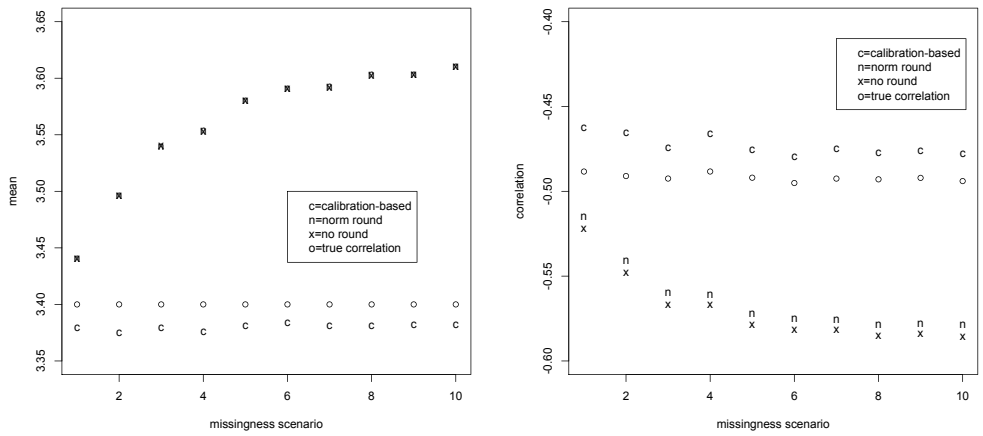

Figure 2: Performance of rounding rules with respect to mean and correlation. Missingness scenario 1 through 10 on the x-axis represent deviations from MCAR towards MAR. For example, scenario '1' represents MCAR with $\beta_R = 0$ in equation (2), scenario '2' represents MCAR with $\beta_R = 0.1$. β_R was given values ranging from 0 to 2.

The results clearly indicate that there is a significant advantage of using calibration-based rounding especially in applications where the missingness mechanism deviates from MCAR towards MAR.

4. Discussion

The motivation of our work on rounding is to provide the practitioners with a well-established, widely-available method for imputing categorical variables under continuous models to usable imputations. By definition, our methods produce imputations with similar distributional properties to the observed data by rounding imputed continuous values. Under a MCAR missingness mechanism, by construction, unbiased estimates are obtained. Relationships of imputed values to other variables are biased, engendering biases in means under an MAR missingness mechanism, however, this bias is generally not much worse than its competitiors. Our simulations as presented in Figure 2 suggest that with modest amounts of missing data, these biases are likely to be tolerable.

The idea of calibration has been recently applied in other settings, particularly on the diagnostics for multiple imputation inference or in Bayesian data analysis.

For example, we can check the model fit by comparing the statistics of the observed (or completed) data with their re-imputed copies under the model. Large differences might suggest a lack of fit of the model (Gelman et al. 2005; Abayomi et al. 2008).

In the imputation of unordered nominal variables, the performance of the calibration approach can be questionable. The method can be criticized for its incompatibility with any continuos model for the purposes of imputation. However, most practitioners follow practices such as ignoring these variables or defining a qualitatively different category representing the missing values both of which can easily lead to misleading results. Our proposal of calibration-based rounding should thus be seen as "promising" starting point in the effort of employing a reliable imputation tool with solid properties in an effort for conducting inferences in datasets with mixture of variable types.

Despite the attractive features of our methods, they should be used cautiously. They are essentially approximate methods based on normal imputation methods and are limited since they have few free parameters that can be used in post-imputation calibration. When the validity of the imputations is critical, as when there are substantial fractions of missing data, it becomes more worthwhile to impute under joint models for categorical variables or combinations of categorical and continuous variables (Schafer 1997, Ch. 9; Javaras and Van Dyk 2003). Alternatively, sequential imputation (Raghunathan et al. 2001; Van Buuren and Oudshoorn 2000) can accommodate data of miscellaneous types by imputing under a collection of univariate conditional models, at the risk of inconsistencies due to the absence of a single joint model.

REFERENCES

Abayomi, K., Gelman, A.E., and Levy, M. (2008) "Diagnostics for multivariate imputations", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**, 273-291.

Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2006), Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data, *Statistics in Medicine*, 26, 1368–1382.

Carpenter, J. and Kenward, M. (2008), *Instructions for MLwiN multiple imputation macros*, Bristol, UK: Centre for Multilevel Modelling.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis (2nd ed.)*, London: Chapman & Hall.

Gelman, A., Mechelen, I., Verbeke, G., Heitjan, D., and Meulders, M. (2005), Multiple Imputation for Model Checking: Completed-Data Plots With Missing and Latent Data, *Biometrics*, 61, 74–85.

Gelman, A., Meng, X., and Stern, H. (1996), Posterior Predictive Assessment of Model Fitness via Realized Discrepancies, *Statistica Sinica*, 6, 733–807.

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.

Javaras, K. N., and Van Dyk, D. A. (2003), Multiple Imputation for Incomplete Data with Semicontinuous Variables, *Journal of the American Statistical Association*, 98, 703–715.

Little, R. (1988), Missing-Data Adjustments in Large Surveys, *Journal of Business & Economic Statistics*, 6, 287–296.

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

Raghunathan, T. E., Lepkowski, J. M., and VanHoewyk, J. (2001), A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models, *Survey Methodology*, 27, 1–20.

Rubin, D. B. (1976), Inference and Missing Data, *Biometrika*, 63, 581–590.

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

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 Institute (2001), SAS/Stat Users Guide, Version 8.2, Carey, NC: SAS Publishing.

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

Schafer, J.L. and Yucel, R.M. (2002) "Computational strategies for multivariate linear mixed-effects models with missing values". *Journal of the Computational and Graphical Statistics*, Volume 11, Number 2, 437–457.

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

Van Buuren, S., and Oudshoorn, C. (2000), *Multivariate Imputation by Chained Equations: MICE V1.0 Users Guide*, Report PG/VGZ/00.038, TNO Preven tie en Gezondheid.

Yucel, R.M., Yulei, He and Zaslavsky, A.M. (2008) Using calibration to improve rounding in multiple imputation. *The American Statistician,* Volume 62, Number 2, 125–129.