S05 - 2008

Imputation of Categorical Missing Data: A comparison of Multivariate Normal and Multinomial Methods

Holmes Finch

Matt Margraf

Ball State University

Abstract

Procedures for the imputation of missing data have been developed for continuous data, using Markov Chain Monte Carlo techniques. This approach has been demonstrated to provide acceptable results for dichotomous data when other variables in the dataset are continuous, rather than categorical, and the resulting imputations are not rounded off so as to fit the scale of the original data. This approach has not been thoroughly studied for situations in which all variables in the dataset are categorical. A categorical analog to the standard multiple imputation method does exist, but it must typically be used with a very small set of variables due to estimation problems. Thus, if multiple imputation methods for continuous data can be demonstrated to perform acceptably for situations in which all variables are categorical, they may provide a useful alternative for many research situations. The current study used a simulation to compare the performance of the continuous and categorical methods for data imputation. The mixed results suggest that in some situations the continuous method may be appropriate for an all categorical variable scenario, while in other cases it did not perform as well as the categorical imputation approach.

Introduction

Psychometricians, data analysts and other statistics professionals are familiar with the presence of missing data in large data sets. For example, in a testing situation examinees may leave one or more items unanswered either inadvertently or because they don't know the answer and are afraid to guess. Respondents to a survey might feel inhibited in answering items dealing with a sensitive topic, leading to missing data. Much research has been conducted regarding the impact of missing data on statistical analyses in general and a variety of methods have been developed for dealing with the problem. The interested reader is encouraged to see Schafer and Graham (2002) for a comprehensive review of methods for dealing with missing data. In addition to the Schafer and Graham (2002) paper, there are a number of other comprehensive discussions regarding specific types of missing data that data analysts might see in practice (Schafer, 1997; Schafer & Olson, 1998; Bernaards & Sijtsma, 1999; Peng & Zhu, 2005; Sinharay, Stern & Russell, 2001). Data that are missing completely at random (MCAR) can be thought of as having no systematic cause; i.e. the missing data are a simple random sample of the observed data (Schafer, 1997, p. 11). When data are missing at random (MAR), the probability of a value being missing is dependent on some measurable characteristic of the individual but not on the missing value itself. Schafer (p.11) points out that for data to be MAR, the variable associated with the probability of data being missing must be observed. Finally, for values missing not at random (MNAR), the likelihood of a variable value being missing is directly related to the value of the variable itself.

Multiple Imputation for continuous data (MI)

MI has been described in very complete detail in several places (e.g. Schafer, 1997; Schafer & Graham, 2002; Leite & Beretvas, 2004; Sinharay, Stern & Russell, 2001; Schafer & Olson, 1998). The interested reader who wishes to learn more about the theory underlying the method is invited to investigate these sources. Below is a brief description to MI for continuous and categorical data. MI was first proposed by Rubin (1987), and was originally developed as an alternative to earlier approaches to imputation such as mean substitution, Hot-Deck imputation, regression based imputation and conditional distribution imputation (Madow, Nisselson & Olkin, 1983; Huisman & Molenaar, 2001). Unlike these single imputation techniques, MI accounts for the inherent uncertainty in sampling from a population by introducing a degree of randomness to the imputations and creating *m* imputed data sets, each of which can then be analyzed in standard ways. MI can incorporate information from other variables into the imputation process in order to provide more accurate values.

The use of MI requires an assumption about the probability model underlying a set of data, such as multivariate normality (frequently used for continuous variables) or a multinomial distribution (common with categorical variables). (Note that other such models are possible but are beyond the scope of this study). Once a probability model is chosen, parameter estimates are made using the Bayesian posterior distribution based upon the likelihood function of the proposed model, the observed data and a prior distribution. The Markov Chain Monte Carlo (MCMC) method of data augmentation is employed to arrive at the posterior distribution from which the imputed values can be

drawn. This imputation process is repeated *M* times (e.g., 10) to create independent data sets (Schafer & Olsen, 1998). Each of these data sets is then subjected to the analysis of interest. The results of the *M* separate analyses (e.g. parameter estimates) are then combined into a single value as

$$\overline{Q} = \frac{\sum \hat{Q}_m}{M} \quad (1).$$

The variance for these estimates is composed of two parts: between imputation variance and within imputation variance. Between imputation variance takes the form

$$B = \frac{\sum \left(\hat{Q}_m - \overline{Q}\right)^2}{M-1} \qquad (2)$$

The within imputation variance, $\overline{U}$, is the mean of estimated variances across the *m* imputations. The total variance for MI is then calculated as

$$T = \overline{U} + \left(1 + \frac{1}{M}\right)B \quad (3).$$

Multiple imputation for categorical data (MIC)

Schafer (1997) described an imputation approach specifically for categorical data (such as might characterize item responses) based on the multinomial distribution. This MIC technique relies on Dirichlet priors (as opposed to the normal priors used in MI), and models relationships among the variables using a log-linear model. MIC is carried out using the probabilities obtained from this log-linear analysis in conjunction with the multinomial distribution, from which final imputed values are obtained.

While in theory MIC is most appropriate for categorical data, Schafer (1997) pointed out that for more than a small number of variables the saturated log-linear model upon which it is based is severely degraded, making it impractical for use with most real

world problems (p. 239). Schafer went on to suggest that using the normal based approach to MI described above may work well for many categorical data problems. When the normal model is used to impute categorical data, it has traditionally been recommended that non-integer values be rounded off so that the imputed data conform to the nature of the actual data (i.e. ordinal or dichotomous integers) (Schafer, 1997; Ake, 2005). Allison (2005), however, found that when using the MI method with dichotomous data, rounding of responses should never be done, as it leads to estimation bias both for calculating proportions and for regression parameter estimates. Other researchers have shown that imputing ordinal data with 5 or more categories using the normal MI model yielded acceptable estimation results when as much as 30% of the data were missing (Leite & Beretvas, 2004; Schafer, Khare, & Ezzati-Rice, 1993).

Previous work has not explicitly compared the performance with categorical data of the theoretically more correct MIC approach for categorical data with the more accessible and flexible MI method. Given Allison's (2005) finding that rounding imputed dichotomous values may result in estimation bias of proportions, it is unclear how effective the MI approach might be when the non-rounded values are preferable (such as when conducting an item analysis). Thus, the focus of the current study was on the estimation accuracy of proportions for imputed data using both the MI (rounded and unrounded) and MIC methods.

Methods

Data were generated using a methodology described by Allison (2005). A detailed description of this method can be found in that manuscript. Specifically, dichotomous data were generated to be both MCAR and MAR with the proportion

missing varying from 0.1, 0.2 and 0.5. The MCAR data were generated so that the probability of an individual response being missing was unrelated to the response itself or to the other variables in the dataset. In the case of MAR, the probability of a data value being missing was correlated to the responses on the other categorical variables. The outcome variable of interest was the proportion of 1's under 4 conditions: (1) Listwise deletion (LD), MI rounded, MI unrounded and MIC. The target proportions were varied across 0.01, 0.05, 0.2 and 0.5. A total of four categorical variables were simulated, with one serving as the target, for which missing data were generated. For both MI conditions, the MCMC method in SAS PROC MI was used, with EM providing the initial parameter estimates on which the markov chains were based. A total of 10 imputed data sets were generated in each of the 500 simulation replications for each condition. A total of 500 subjects were simulated for each of these replications. For the MI analyses, PROC MIANALYZE was used to provide estimates of the proportions of interest. MIC was carried out using the set of CAT functions in the R software package. For each replication, 10 imputations were conducted, as with the MI.

Results

Results for the MCAR data appear in Table 1. Generally speaking, the MIC method produced the most biased estimates of the four methods studied here, when the data were MCAR. This positive bias was most pronounced when the parameter value was small (0.01 and 0.05). In no case were the MIC based estimates less biased than any of the other approaches. In addition, this estimation bias became more pronounced as the proportion of missing data increased. On the other hand, the estimates produced by the other methods were generally within 0.05 of the actual parameter values.

Table 1: Estimates of proportions for MCAR data

Proportion missing	Parameter	MIC	LD	MI NR	MI R
0.1	0.01	0.110	0.012	0.019	0.017
	0.05	0.150	0.050	0.051	0.049
	0.20	0.250	0.160	0.200	0.210
	0.50	0.470	0.476	0.495	0.482
0.2	0.01	0.217	0.009	0.018	0.016
	0.05	0.240	0.050	0.051	0.046
	0.20	0.365	0.202	0.203	0.207
	0.50	0.406	0.500	0.500	0.500
0.5	0.01	0.506	0.009	0.024	0.015
	0.05	0.471	0.053	0.057	0.052
	0.20	0.600	0.196	0.201	0.213
	0.50	0.745	0.505	0.509	0.500

MIC=MI using multinomial distribution, LD=Listwise deletion, MI NR=MI using normal distribution and nonrounded values, MI=MI using normal distribution and rounded values

Results for the MAR data appear in Table 2. Unlike with MCAR data, it appears that the performance of MIC is comparable to or slightly better than either MI method for most of the conditions simulated here. The only exception to this result is with a parameter value of 0.01 and 10% missing data. As the proportion of missing data increased, the MIC technique generally produced less biased estimates than either of the MI methods, though

these results were still positively biased. Of the methods examined here, LD produced the least biased outcomes in the MAR condition.

Table 2: Estimates of proportions for MAR data

Proportion missing	Parameter	MIC	LD	MI NR	MI R
0.1	0.01	0.080	0.009	0.010	0.001
	0.05	0.090	0.085	0.163	0.198
	0.20	0.220	0.140	0.180	0.176
	0.50	0.460	0.473	0.458	0.476
0.2	0.01	0.110	0.020	0.157	0.097
	0.05	0.090	0.011	0.097	0.100
	0.20	0.280	0.215	0.373	0.309
	0.50	0.470	0.494	0.500	0.469
0.5	0.01	0.100	0.021	0.172	0.115
	0.05	0.120	0.071	0.123	0.200
	0.20	0.230	0.236	0.311	0.370
	0.50	0.430	0.518	0.476	0.499

Conclusions

The results of this simulation study carry with them several implications for data analysis practice. First, it is unclear that the MIC approach yields clearly superior results with respect to parameter estimation bias in the case of MCAR data. Indeed, the imputation method designed for use with normally distributed continuous data, as well as listwise deletion, were associated with much lower rates of bias for MCAR. In addition,

the problems associated with the MIC method were more pronounced for a larger proportion of missing data.

A second major finding of this study is that when the data are MAR, and the variables used in the imputation of missing responses are categorical, estimates based on MIC displayed much lower levels of bias than did those based on MI. Indeed, the bias results obtained here for MI in the MAR case were very different than those reported in Allison (2005). The difference between the two studies is that Allison used a continuous variable in the imputation of the categorical target variable, while in the current study all of the variables were categorical in nature. Interestingly, the results for LD were generally as good as, or somewhat better than those for the normal based imputation methods.

The third primary result of this study is that the performance of MIC appears to be degraded when the population parameter value is small (0.01 or 0.05 in this study). Regardless of whether the data were MCAR or MAR, and across proportion of missing data, MIC consistently yielded positively biased parameter estimates. When the data were MAR, the parameter estimates associated with MIC demonstrated much less bias. However, for MCAR data, the bias remained very large, particularly as the proportion of missing data increased.

Implications for practice

The results of this study suggest that practitioners interested in imputation for dichotomous data must be very cautious regarding the method that they select. If the data are MCAR, the normal theory based approaches may well yield acceptable parameter

estimates. On the other hand, if the data are MAR and associated only with other dichotomous variables, the normal based methods may not work particularly well.

While it was designed for categorical data, the MIC method used here also presented several problems in practice. When the data were MCAR, it tended to create imputed data sets that yielded more biased results than any of the other methods studied here, including LD. On the other hand, when the data were MAR and the parameter value of interest was not small (0.01 and 0.05 in this study), MIC tended to perform better than the normal based methods, though not necessarily better than simple LD.

Taken together, these results would suggest that in many instances where only categorical variables are involved, the MIC method for data imputation may indeed be preferable to the normal based methods. However, if the incidence of the behavior or trait being studied is low, then MIC would appear to create imputations in which the trait appears too frequently. And, because it is often unclear what the actual mechanism is underlying the missing data, data analysts may not have a sense for whether MI or MIC would be the preferable alternative.

References

Ake, C.F. (2005). Rounding after multiple imputation with non-binary categorical covariates. Paper presented at the annual meeting of the Sas Users Group International, Philadelphia, PA.

Allison, P.D. (2005). Imputation of categorical variables with PROC MI. Paper presented at the annual meeting of the Sas Users Group International, Philadelphia, PA

Bernaards, C.A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data from ignorable item nonresponse. *Multivariate Behavioral Research, 34*, 277-314.

Huisman, M. & Molenaar, I.W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M.A.J. van Duijn and T.A.B. Snijders (Ed.s), *Essays on Item Response Theory* (pp. 221-244). New York: Springer.

Leite, W.L. & Beretvas, S.N. (2004). The performance of multiple imputation for Likert-type items with missing data. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Madow, W.G., Nisselson, H., & Olkin, I. (eds.) (1983). *Incomplete data in sample surveys, Vol 1: Report and case studies*. New York: Academic Press.

Peng, C.-Y.J., & Zhu, J. (2005). Comparison of two methods for handling missing covariates in logistic regression. Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman and Hall/CRC.

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177.

Schafer, J.L., Khare, M. & Ezzati-Rice, T. (1993). Multiple imputation of missing data in NHANES III. *Proceedings of the 1993 Annual Research Conference*. Washington, DC: U.S. Bureau of the Census.

Schafer, J.L., & Olsen, M.K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545-571.

Sinharay, S., Stern, H.S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317-329.

Contact information:

Holmes Finch

Department of Educational Psychology

Ball State University

Muncie, IN 47306