Journal of Data Science **8**(2010), 361-378

Imputation Methods for Missing Categorical Questionnaire Data: A Comparison of Approaches

W. Holmes Finch
Ball State University

Abstract: Missing data are a common problem for researchers working with surveys and other types of questionnaires. Often, respondents do not respond to one or more items, making the conduct of statistical analyses, as well as the calculation of scores difficult. A number of methods have been developed for dealing with missing data, though most of these have focused on continuous variables. It is not clear that these techniques for imputation are appropriate for the categorical items that make up surveys. However, methods of imputation specifically designed for categorical data are either limited in terms of the number of variables they can accommodate, or have not been fully compared with the continuous data approaches used with categorical variables. The goal of the current study was to compare the performance of these explicitly categorical imputation approaches with the more well established continuous method used with categorical item responses. Results of the simulation study based on real data demonstrate that the continuous based imputation approach and a categorical method based on stochastic regression appear to perform well in terms of creating data that match the complete datasets in terms of logistic regression results.

Key words: Missing data, multiple imputation, stochastic regression, categorical data.

1. Introduction and Motivation

Statisticians and data analysts are often faced with the issue of missing data in surveys and questionnaires. For example, respondents may elect to leave one or more items unanswered either inadvertently or because they feel inhibited in responding to items dealing with a sensitive topic. Researchers in statistics and in the social sciences have investigated the impact of such missing data on statistical analyses in surveys (see de Leeuw, Hox and Huisman, 2003, for a review of this issue). This prior work has demonstrated that missing data can have a detrimental impact on statistical analyses based on the survey responses, including biased parameter estimates and inflated standard errors. These problems may be

particularly acute when the reason for the missing response is directly related to the missing value itself. In addition, when relatively large amounts of data are missing, the power of statistical tests can be severely compromised (de Leeuw, Hox and Huisman, 2003).

Given the potential problems that arise from the presence of missing data, a variety of methods have been suggested for data imputation. Schafer and Graham (2002) provide a comprehensive review of many of these methods, including older approaches that are no longer recommended, as well as the more modern multiple imputation and data augmentation methods. In addition to the Schafer and Graham (2002) paper, there are a number of other excellent discussions regarding the technical details of methods for handling missing data that analysts might find useful (Schafer, 1997; Schafer and Olson, 1998; Bernaards and Sijtsma, 1999; McKnight, McKnight, Sidani and Figuerdo, 2007; Sinharay, Stern and Russell, 2001; Little and Rubin, 2002).

Statisticians have traditionally characterized missing data based on the underlying process believed to have led to it. Each of these processes has unique characteristics both in terms of reasons for the missing data, and the implications of the specific type of missingness. Following is a brief description of each type of missing data. Again, interested readers are encouraged to refer to the resources listed above for a more thorough discussion of these missing data types. 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). In the context of survey responses, MCAR data might occur when a respondent simply overlooks an item, for example when neglecting to turn the page of a questionnaire booklet. Data are missing at random (MAR) when the probability of a value being missing is dependent on some measurable characteristic of the individual but not on the missing value itself. For example, male respondents might be less inclined to answer a particular survey item than females. Schafer (p.11) points out that for data to be MAR, the variable associated with the probability of data being missing must be measured, and can then be used for imputing the missing response. Thus, in the survey example, the gender of the subjects completing the instrument would need to have been recorded for the missing item response to be considered MAR. 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. As an example, respondents might be asked to indicate the number of alcoholic drinks they consumed during the previous week. A missing response would be MNAR if individuals who consumed larger amounts of alcohol during this period were more likely to leave the item unanswered rather than report their behavior.

As mentioned above, many methods have been suggested for imputing valid

responses to missing data. Among these techniques were mean substitution, in which the average value for the sample was imputed for missing observations of a particular variable, regression imputation, where the missing observation was imputed using the prediction taken from a multiple regression analysis, and Hot Deck imputation, where the missing value was replaced with that of an observed value taken from a matched observation based on the non-missing variables. Each of these single imputation methods has been found to be inadequate in terms of accurately reproducing known population parameters and standard errors (Schafer and Graham, 2002).

Given the problems inherent with these early methods, researchers have worked to develop approaches that are more appropriate for replacing the missing observations. A variety of such approaches have been described in the literature, with multiple imputation based on data augmentation being one of the premier such techniques. Indeed, a growing body of research suggests that for normally distributed continuous data, this approach is optimal (Schafer and Olson, 1998). When the data in question are categorical, however, it is not as clear what the appropriate methodology for imputing missing data should be. Schafer (1997) describes a log-linear model based multiple imputation approach for missing categorical data, though as is discussed below in more detail, it may be very limited in terms of application due to problems estimating the higher order interactions that are part of such models. Another alternative for imputing missing categorical data based on the multinomial distribution and logistic regression has recently been introduced by Sulis and Porcu (2008). This method does not appear to suffer from the same model size limitations as the log-linear based method, but yet is based on categorical, rather than continuous, data distributions. However, this stochastic regression imputation methodology has not been compared with the log-linear model approach, nor with the normal based imputation using categorical data. Thus, the primary goal of this study was to use a Monte Carlo simulation, based on actual survey data, to explicitly compare the effectiveness of these three imputation methods. Following is a brief description of the three methods for data imputation, followed by a review of existing literature examining their effectiveness in data imputation.

Multiple Imputation for continuous data (MI)

MI has been described thoroughly elsewhere, (e.g. Schafer, 1997; Schafer and Graham, 2002; Leite and Beretvas, 2004; Sinharay, Stern and Russell, 2001; Schafer and Olson, 1998) so that the interested reader desiring to learn more about the theory underlying this method is invited to investigate these sources. Following is a brief description of MI and its application to continuous and cat-

egorical data. MI was first proposed by Rubin (1987), and was originally developed as an alternative to the earlier single imputation approaches described above (Madow, Nisselson and Olkin, 1983; Huisman and Molenaar, 2001). Unlike these early techniques, MI accounts for the inherent uncertainty in sampling from a population by introducing randomness to the imputations and creating m imputed data sets, each of which is then subjected to the desired statistical analyses (e.g., regression, analysis of variance, factor analysis, etc). MI incorporates information from other variables into the imputation process in order to provide more accurate values.

The use of MI for continuous data requires an assumption that the probability distribution underlying the data is multivariate normal. (Note that other such models are possible but are beyond the scope of this paper). Based on this probability model, 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 used to obtain this posterior distribution from which the imputed values for the missing observation are drawn. This imputation process is repeated M times (e.g., 10) to create independent data sets (Schafer and Olsen, 1998), each of which is then subjected to the analysis of interest, such as the ordinal logistic regression used in this manuscript. The results of the M separate analyses (e.g. parameter estimates) are then combined into a single value as

$$\bar{Q} = \frac{1}{M} \sum_{m} \hat{Q}_m \tag{1.1}$$

where $\hat{W}_m$ is the parameter estimate of interest for imputation m. The variance for these estimates is composed of two parts: between imputation variance and within imputation variance. Between imputation variance takes the form

$$B = \frac{1}{M-1} \sum_{m} (\hat{W}_m - \bar{Q})^2. \tag{1.2}$$

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

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B \tag{1.3}$$

Multiple imputation for categorical data (MIC)

Schafer (1997) described an imputation approach for categorical variables that was similar in spirit to MI, but based on the multinomial distribution. While MI

uses normally distributed priors in obtaining the posterior distribution used for obtaining imputed values, the MIC variant relies instead on Dirichlet priors. The relationships among the categorical variables are estimated using a log-linear model, including all possible interactions among the observed variables. MIC is carried out using response category probabilities obtained from this log-linear analysis in conjunction with the multinomial distribution, from which final imputed values are obtained. One of the potential problems with this approach to imputation is that when many variables are present, the associated log-linear model will become very complex with a large number of higher order interaction terms. In order for these interactions to be estimable, a very large sample is required. For this reason, the MIC approach to imputation is often not practical in real world situations (Schafer, 1997, p.239). Schafer suggests the use of the MI approach instead, with the user rounding the imputed values to fit with the possible values of the variables (p. 148). He argues that MI should work in most situations. However, to this point, the effectiveness of using it with categorical data has not been compared with the MIC technique designed specifically for categorical variables.

Stochastic regression imputation (SRI)

Sulis and Porcu (2008) introduced an alternative method for imputation of ordinal variables that they suggest does not have the same limitations in terms of number of variables, that exist for MIC. This stochastic regression imputation (SRI) is based on two steps. In step 1, the distribution of relative frequencies of each response category for each member of the sample is created based on the observed data. Then, for each member of the sample, missing responses are replaced by random draws from the multinomial distribution with parameters set equal to the distribution of relative frequencies of the observed data for each of the discrete categories. These random draws are done once for each of the M imputed datasets. In this study, 10 such imputations were used for each missing value.

In the second step of SRI, a proportional odds logistic regression analysis is conducted for each variable for each of the M complete datasets. More specifically, in this study where instructor satisfaction survey responses are used, each item serves as the dependent variable and the other variables in the dataset serve as the independent variables in this analysis. For each dependent variable, the following logistic regression model is estimated:

$$\text{logit}[P(Y \leq k|x)] = \alpha_k + Bx, \tag{1.4}$$

where

$$\begin{aligned} \alpha_k &= \text{intercept for response category} k \\ B &= \text{slope for variable} x \end{aligned}$$

The model parameter estimates are then used in making random draws from the multinomial distribution for each missing response on the dependent variable in the logistic regression. Again, for each variable with missing data, logistic regression is conducted and the random draws from the multinomial distribution with parameters based on the resulting slope and intercept values. This procedure is conducted for each of the M imputed datasets obtained in step 1.

Previous research

As noted above, while in theory MIC is more appropriate for categorical data than is MI, 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 MI has been used to impute categorical data, it has traditionally been recommended that non-integer values be rounded so that the resulting 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 could lead to estimation bias when calculating proportions. On the other hand, other researchers have shown that imputing ordinal data with 5 or more categories using MI yielded acceptable correlation estimation results when as much as 30%of the data were missing (Leite and Beretvas, 2004; Schafer, Khare, and Ezzati-Rice, 1993).

Sulis and Porcu (2008) evaluated the performance of the SRI approach in a simulation study. They took an existing dataset and using a Monte Carlo approach, generated 5 datasets with missing values at rates of 5%, 10%, 15%, 20% and 25%, each for MAR and MCAR data. The outcomes of interest were the slope estimates obtained using logistic regression for both SRI and complete data. The results of their simulations showed that when the rate of missing data was less than 15%, SRI produced very similar estimates to those from the complete datasets. However, as the rate of missingness increased above this level, the SRI based parameter estimates began to diverge from those from the complete data. The standard errors that were obtained in the SRI condition were very similar to those in the complete case under the simulated conditions.

Previous work in the field of multiple imputation for categorical data has not explicitly compared the performance of the theoretically more appropriate MIC

approach with the more accessible and flexible MI method, or the much newer and less explored SRI approach. 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 using survey items). Furthermore, the original study examining SRI was somewhat limited in terms of the number of replications per condition (5) and in what was being directly examined. Thus, the focus of the current study was on the estimation accuracy of logistic regression parameters, standard errors and hypothesis testing results for imputed data using the MI (rounded), SRI and MIC methods. It represents one of the first studies to directly compare these three methods with one another in order to ascertain which approach might be most appropriate with survey type categorical data. In addition to the three imputation methods, this study also includes results for the complete data case, which serves as a baseline, and for the case where missing data are treated using listwise deletion; i.e., observations with any missing values are eliminated from the analyses.

2. Methods

The simulation study reported here was based on actual course satisfaction survey responses, for which missing values were generated as described below. Course evaluation data were collected from undergraduate students ($N = 2,000$) enrolled in an introductory science course at a large United States university. This particular course serves as a general science education credit for non-science majors, and thus attracts students from across the university representing a wide variety of majors and levels of interest in science. The satisfaction questionnaire (see appendix) consisted of 22 items designed to ascertain student opinions on several aspects of the class. Each item was presented in the form of a statement with which students indicated their level of agreement on a 5-point Likert scale. Note that lower value responses were associated with more positive ratings of the course and instructor.

The dependent variable for the analyses reported here was item 21, "This instructor was one of the better University Core Curriculum instructors I have had at this University". Items 2 through 5 were selected as independent variables in an ordinal logistic regression (cumulative logits model) in which item 21 served as the response. The purpose of this analysis was to determine whether, and to what extent, certain aspects of the instructor's course conduct (preparation for class, focus on student understanding, answering of questions and exam coverage) were associated with the overall rating of the instructor given by students. Thus, slopes, standard errors and results of hypothesis tests served as the outcomes of interest in the simulation study.

In order to assess the performance of the three methods for imputing missing data that were considered in this study, a random sample of respondents was selected for each sample size condition ($N = 200, 500$, and 1000). In addition, the entire sample of 2000 respondents was also used. For each sample, missing data for the 5 variables was randomly created for both the MCAR and MAR conditions using a function written for the R software package. A total of 100 replications were generated for each sample size by type of missing data combination. The three imputation methods described above were then applied to the missing data and the results of the 10 imputations for each were combined using equations 1, 2, and 3 above. For each such replication, the ordinal logistic regression described above was conducted, and the slope estimates, standard errors and hypothesis testing results were recorded.

For both MCAR and MAR data conditions, the rate of missing data was set at 25%. For the MAR data, having a missing observation was associated with the gender of the respondent, so that random missing data was generated for females. In the MCAR condition, missing data were generated through simple random selection from among all respondents. Imputation was carried out separately for each replication using the MI, MIC and SRI methods with functions available in the R software package. In the case of the MAR data, respondent gender was included in the imputation of missing responses for the individual items. For the MCAR data, imputations were based only on the item responses, excluding gender. For each of these imputed datasets as well as for those containing the missing observations (listwise deletion) and the complete data, ordinal logistic regression was conducted.

3. Results

MCAR data

Results for parameter estimation and standard errors of the four slopes for the MCAR data appear in Table 1. When compared with the results based on the complete dataset, treating the data with listwise deletion appears to have resulted in fairly minor bias. Indeed, across all sample size conditions, the listwise deletion and complete data estimates never differed by more than 0.02. However, the standard errors for these listwise deletion estimates were consistently more than twice as large as the standard errors in the complete data condition. Thus, while there appears to have been little or no bias in the former values, they were much less efficient than those based on the complete dataset.

Table 1: Parameter estimates and standard errors by missing data method: MCAR

Sample size	Complete	Normal	Categorical	Stochastic	Listwise deletion
Slope 1					
200	0.440 (0.073)	0.424 (0.074)	0.274 (0.080)	0.378 (0.073)	0.448 (0.155)
500	0.282 (0.048)	0.283 (0.050)	0.244 (0.052)	0.273 (0.047)	0.285 (0.099)
1000	0.303 (0.021)	0.296 (0.022)	0.269 (0.022)	0.296 (0.021)	0.294 (0.044)
2000	0.303 (0.021)	0.299 (0.022)	0.272 (0.022)	0.294 (0.021)	0.300 (0.044)
Slope 2					
200	0.170 (0.054)	0.180 (0.056)	0.187 (0.069)	0.175 (0.057)	0.157 (0.115)
500	0.160 (0.038)	0.161 (0.040)	0.171 (0.043)	0.169 (0.038)	0.165 (0.079)
1000	0.157 (0.017)	0.160 (0.017)	0.170 (0.019)	0.162 (0.017)	0.157 (0.035)
2000	0.157 (0.017)	0.157 (0.017)	0.170 (0.019)	0.164 (0.017)	0.156 (0.035)
Slope 3					
200	-0.043 (0.057)	-0.030 (0.059)	0.081 (0.075)	0.016 (0.058)	-0.028 (0.122)
500	0.141 (0.040)	0.145 (0.041)	0.147 (0.046)	0.151 (0.041)	0.124 (0.084)
1000	0.064 (0.017)	0.065 (0.018)	0.093 (0.019)	0.085 (0.018)	0.057 (0.036)
2000	0.064 (0.017)	0.067 (0.018)	0.094 (0.019)	0.087 (0.018)	0.065 (0.036)
Slope 4					
200	0.337 (0.077)	0.316 (0.082)	0.230 (0.084)	0.335 (0.076)	0.322 (0.171)
500	0.304 (0.049)	0.288 (0.052)	0.223 (0.050)	0.296 (0.048)	0.307 (0.103)
1000	0.358 (0.023)	0.350 (0.024)	0.297 (0.024)	0.337 (0.023)	0.463 (0.048)
2000	0.358 (0.023)	0.349 (0.023)	0.293 (0.023)	0.336 (0.023)	0.367 (0.048)

Among the data imputation methodologies examined here, MIC was notable for consistently producing the most biased slope estimates vis-à-vis those in the complete data case. The lone exception to this result occurred for slope 3 with a sample of 500. The greatest bias for analyses run with MIC generally occurred for the smallest sample size, and declined somewhat in relative magnitude as the sample size increased. The nature of bias for these estimates also warrants some discussion. Specifically, for slopes 1 and 4, the estimates for MIC was negatively biased as compared to that of the complete set of data, which in both cases was 0.30 or larger. On the other hand, for the slope estimates lower than 0.20 for the complete data set, the MIC based values demonstrated a systematic positive bias. The standard errors associated with MIC were lower than those from listwise deletion and were generally only slightly elevated when compared to the complete data analyses or the other methods of imputation. . In comparison with results for MIC, the slope estimates for both the MI and SRI methods were much closer to those from the complete dataset. In particular, the parameter estimates for the MI based approach were generally (though not always) the least biased when compared to the complete data results. However, the SRI based estimates were also closer to the complete data slopes than were those based on MIC, and only slightly more biased than those from MI. As mentioned above, the standard errors associated with the MI and SRI methods were typically lower than those from the MIC approach. Furthermore, the standard errors from the SRI data were

either equal to or slightly smaller than those from MI, and in many cases the same as those for the complete data.

Table 2: Hypothesis testing rejection rates by missing data method: MCAR

Sample size	Complete	Normal	Categorical	Stochastic	MCAR
Slope 1					
200	1.00	0.99	0.97	0.99	0.71
500	1.00	1.00	1.00	1.00	0.80
1.000	1.00	1.00	1.00	1.00	1.00
2000	1.00	1.00	1.00	1.00	1.00
Slope 2					
200	1.00	0.85	0.84	0.85	0.32
500	1.00	0.94	0.98	0.97	0.45
1000	1.00	1.00	1.00	1.00	0.99
2000	1.00	1.00	1.00	1.00	0.99
Slope 3					
200	1.00	0.06	0.12	0.11	0.02
500	1.00	0.93	0.96	0.91	0.35
1000	1.00	0.91	1.00	0.99	0.39
2000	1.00	0.91	0.99	0.98	0.47
Slope 4					
200	1.00	0.88	0.85	0.95	0.46
500	1.00	1.00	1.00	1.00	0.77
1000	1.00	1.00	1.00	1.00	1.00
2000	1.00	1.00	1.00	1.00	1.00

In terms of the hypothesis test results (Table 2), all four slopes were significantly different from 0 in the complete data case for all replications across sample sizes. For listwise deletion, the rejection rates were much lower than for the complete data when the samples were 500 or fewer for slopes 1, 2 and 4. In addition, for slope 3 the rejection rate in the listwise deletion case was well below that of the complete data even for the largest sample size condition. With respect to the imputation methods, rejection rates were comparable to those for the complete data across sample sizes for slope 1. For slopes 2, 3 and 4, the rejection rates for the three imputation methods were above 0.9 when the samples were 500 or more. For slope 3, the rejection rates for the MI data were somewhat lower than those for either MIC or SRI, across sample sizes. An examination of the slopes and standard errors in Table 1 suggests that this relatively lower rejection rate is due to the lower slope estimates in the MI case as compared to the other two imputation methods. Indeed, while these normally based slopes were closer to those obtained from the complete data, the standard errors were slightly larger,

leading to the somewhat lower rejection rates.

MAR data

In the MAR condition, missing data was simulated to be associated with the gender of the respondent. Thus, gender was included in the imputation procedures, though not in the actual logistic regression models producing the results that appear in Tables 3 and 4. As was the case with the MCAR data, for the listwise deletion condition slope estimates were biased when compared with the complete data case. In addition, the standard errors for these estimates were uniformly larger than were those from the complete data and the three imputation conditions. The degree of difference in these standard errors declined as the sample size increased. Finally, a comparison of Tables 1 and 3 reveals that the magnitude of slope estimate bias for the listwise deletion approach in the MAR condition was higher than for the MCAR data.

Table 3: Parameter estimates and standard errors by missing data method: MAR

Sample size	Complete	Normal	Categorical	Stochastic	MCAR
Slope 1					
200	0.284 (0.060)	0.284 (0.061)	0.218 (0.072)	0.289 (0.059)	0.269 (0.101)
500	0.301 (0.045)	0.320 (0.048)	0.238 (0.050)	0.303 (0.045)	0.341 (0.078)
1000	0.309 (0.031)	0.288 (0.032)	0.250 (0.034)	0.294 (0.031)	0.271 (0.053)
2000	0.303 (0.022)	0.292 (0.021)	0.259 (0.022)	0.287 (0.021)	0.289 (0.036)
Slope 2					
200	0.119 (0.043)	0.110 (0.045)	0.143 (0.062)	0.114 (0.044)	0.099 (0.074)
500	0.121 (0.037)	0.092 (0.041)	0.158 (0.045)	0.111 (0.038)	0.055 (0.066)
1000	0.114 (0.024)	0.116 (0.026)	0.154 (0.030)	0.124 (0.025)	0.105 (0.042)
2000	0.157 (0.017)	0.163 (0.017)	0.175 (0.019)	0.167 (0.017)	0.164 (0.030)
Slope 3					
200	0.151 (0.045)	0.159 (0.047)	0.155 (0.070)	0.158 (0.048)	0.186 (0.084)
500	0.130 (0.037)	0.114 (0.039)	0.139 (0.044)	0.126 (0.038)	0.109 (0.064)
1000	0.073 (0.025)	0.078 (0.026)	0.109 (0.030)	0.093 (0.025)	0.081 (0.042)
2000	0.064 (0.017)	0.070 (0.018)	0.1.00 (0.020)	0.086 (0.018)	0.067 (0.030)
Slope 4					
200	0.360 (0.064)	0.350 (0.065)	0.213 (0.069)	0.344 (0.062)	0.364 (0.108)
500	0.373 (0.045)	0.386 (0.046)	0.253 (0.047)	0.378 (0.044)	0.392 (0.077)
1000	0.387 (0.034)	0.398 (0.033)	0.261 (0.034)	0.380 (0.032)	0.437 (0.05)
2000	0.359 (0.023)	0.349 (0.023)	0.275 (0.023)	0.342 (0.022)	0.363 (0.038)

Just as was true in the MCAR case, the slope estimates for MIC were generally more biased, when compared with the complete data condition, than were those from the other two imputation techniques. The standard errors associated with these MIC results were also typically larger than for the other two methods, though this pattern did not hold in all conditions. A comparison of bias and

standard errors for MIC between the MCAR and MAR conditions revealed no systematic differences in either slope estimates or standard errors.

The results based on MI and SRI were generally similar to their results with MCAR data. Namely, the degree of bias for both methods, when compared with the complete data, was comparable. In addition, it does not appear that the magnitude of this bias differed between the MCAR and MAR conditions. In terms of standard errors, those from SRI were in many cases slightly lower than those from MI. However, these differences were not very large, being on the order of 0.001 to 0.003. As with the slope estimates, these differences were very comparable to what was found for the MCAR data.

Table 4: Hypothesis testing rejection rates by missing data method: MAR

Sample size	Complete	Normal	Categorical	Stochastic	MCAR
Slope 1					
200	0.95	0.95	0.89	0.95	0.65
500	1.00	1.00	1.00	1.00	0.95
1000	1.00	1.00	1.00	1.00	1.00
2000	1.00	1.00	1.00	1.00	1.00
Slope 2					
200	0.95	0.67	0.66	0.66	0.29
500	1.00	0.57	0.98	0.78	0.10
1000	1.00	1.00	1.00	1.00	0.74
2000	1.00	1.00	1.00	1.00	1.00
Slope 3					
200	1.00	0.86	0.66	0.85	0.62
500	1.00	0.78	0.95	0.88	0.40
1000	1.00	0.84	0.98	0.87	0.48
2000	1.00	0.97	1.00	1.00	0.64
Slope 4					
200	1.00	1.00	0.92	1.00	0.88
500	1.00	1.00	1.00	1.00	1.00
1000	1.00	1.00	1.00	1.00	1.00
2000	1.00	1.00	1.00	1.00	1.00

The rejection rates for the hypothesis tests of the slopes in the logistic regression models for the MAR data appear in Table 4. For the complete data, the null hypothesis was rejected in all replications across all sample sizes except for slopes 1 and 2 in the N=200, where the rates were 0.95. For listwise deletion the rejection rates were lower than those in the complete case for a sample of 200 subjects. On the other hand, for the largest sample condition, the MAR rejection rates were uniformly 1.0. For slopes 1 and 4, the rejection rates for $N = 500$ and

1000, the listwise deletion rejection rates were at or near 1.0, while for slopes 2 and 3 at these sample sizes, the rejection rates were much lower than those for the complete data condition. When compared to the other imputation methods, the rejection rates for these two slopes with MIC were higher than for either MI or SRI. It is important to note that the parameter estimates for MIC in this case were larger as well, except when $N = 200$. On the other hand, for slopes 1 and 4 the rejection rates for the MIC were lower than for the other methods. For each of these variables, the MIC estimates were lower than those for MIR or SRI. The rejection rates for the SRI were comparable to, or higher than those from MI across conditions simulated here.

4. Conclusions

Researchers in the social sciences who make use of surveys will almost certainly be faced with the problem of missing data from time to time. Missing item responses have been shown to be associated with problems such as parameter estimate bias, inflated standard errors, and low statistical power (de Leeuw, Hox, and Huisman, 2003). While there are a number of imputation approaches available for dealing with such missing dat, statisticians have generally settled on the optimality of what is commonly referred to as multiple imputation for normally distributed variables (Schafer and Graham, 2002). However, an analogous method for categorical variables, based on the log-linear model, may be problematic in situations involving even a modest number of variables (Schafer, 1997). For this reason, the imputation technique based on the normal distribution has often served as the only recourse for those dealing with categorical data and missing responses (Allison, 2005). Recently, an alternative method for imputing missing categorical data based on the multinomial distribution and stochastic logistic regression has been proposed (Sulis and Porcu, 2008). To date, very little if any work has been published comparing the performance of the stochastic regression, log-linear and normal based approaches to imputation. The current study, therefore, was designed to make such a direct comparison among these three methods, and to evaluate their performance based on the similarity of regression results (slopes, standard errors and hypothesis tests) with that based on the complete dataset.

The results of this study have several implications for researchers faced with missing data for categorical variables. First of all, in keeping with prior studies on missing data with normally distributed data, it seems clear that ignoring the missing values (i.e., using listwise deletion) is inappropriate, whether the data are MCAR or MAR. In both cases, the standard errors and results of hypothesis tests for the slopes varied from the complete data case to a greater extent than did the results for any of the imputation methods. In general, this divergence became less

extreme as the sample size increased, but at no point were the standard errors from the listwise deletion approach as small as those from any of the imputation techniques. The relative bias in parameter estimates was not, in general, as great for the listwise deletion approach as it was for MIC, though it was typically more marked than for either MI or SRI.

In terms of imputation for categorical missing data, these results suggest that the MIC approach may be associated with biased parameter estimates and somewhat inflated standard errors, when compared with the complete data case. When the estimates for the complete datasets were 0.30 or above, the estimates for the MIC data tended to be negatively biased, while when the complete data slopes were less than 0.20, the MIC slopes were generally positively biased. In addition, the standard errors associated with MIC were generally larger than those for the complete datasets, except in a few instances for the full sample of 2000 subjects.

While the MIC based results were characterized by biased slope estimates and inflated standard errors, the MI datasets produced estimates that generally displayed less bias and less inflation of the standard errors. This result was not uniformly true, however, and was more pronounced for the MCAR data. With respect to MAR, the MI bias was smaller than MIC for variables 1, 2 and 4, which had the largest slope estimates in the complete data case. On the other hand, the bias for MIC on slope 3 was comparable to or slightly less than that of MI in the MAR case. The standard errors for the MI data were typically less than or equal to those for MIC, and were generally slightly larger for the MCAR data, though this difference was never more than 0.01. With respect to the MAR condition, the standard errors of the MI data were even closer to those based on the complete case than was true for MCAR.

The regression results based on SRI typically displayed slightly greater estimation bias than did MI when the data were MCAR, but when the data were MAR SRI displayed comparable bias to that of MI across slopes and sample sizes. And, as was the case for MI, SRI estimation bias was less pronounced than was that for MIC data. For both types of missing data, the SRI based estimates had lower standard errors than did MI for nearly all slope by sample size combinations. In addition, the standard errors for the SRI data were comparable to or lower than those for the complete data in the MCAR case, and only slightly elevated for MAR.

Given these results, it appears that either the MI or SRI methods of imputation for missing ordinal data are preferable to MIC. The latter method was associated with both greater estimation bias and higher levels of variation in these estimates, regardless of whether the data were MCAR or MAR. Furthermore, the bias for MI based estimates in this study was fairly small, obviating some

concerns over previous findings (e.g., Ake, 2005) demonstrating that rounding the imputed values did not produce markedly biased results. It should be noted that these earlier results demonstrating bias were for dichotomous data, and were based on proportion estimates rather than regression analyses as was the case in this study.

It would appear that Schafer's (1997) recommendation for researchers to use MI for missing categorical data is supported by the results presented here. This method generally performed as well as SRI and better than MIC in terms of parameter estimation and the associated standard errors. This outcome has important implications for data analysts and others faced with missing ordinal data because the MI approach is widely available in software packages such as SAS and SPSS. Clearly, the SRI technique is also a reasonable alternative to MIC, though it is generally not as accessible to data analysts not familiar with the R software package. In addition, the limitations regarding the number of variables to be imputed that are endemic to the MIC method are not problematic for MI (Schafer, 1997), nor to SRI (Sulis and Porcu, 2008). Given the current results, researchers need not be concerned that their categorical data imputations based on MI are problematic, and can apply this method to much larger datasets than could be used with MIC.

Study Limitations and directions for future research

As with any study, there are limitations to the current effort that must be considered when interpreting the results. First of all, the simulations were based on one set of questionnaire responses. While we believe that these data are representative of similar instructor satisfaction surveys used in universities across higher education, it would be interesting to extend the current study to include other types of surveys that are commonly used in the social sciences.

A second limitation of the current study is its focus on ordinal categorical variables in the form of likert items. While this type of data is common with respect to many surveys, there are many other contexts in which dichotomous or multinomial (unordered) categorical variables are of interest. Future research should focus on the relative utility of MI, MIC and SRI methods of imputation for such non-ordered categorical variables.

Finally, the three methods of imputation examined here should be examined with respect to other types of statistical analyses. The current study focused on slopes, standard errors and hypothesis tests for ordinal logistic regression, which is very similar in form to the logistic regression model used to impute missing data in the SRI approach. While it is not clear that the relatively favorable results demonstrated for this method are due to this similarity between the models

underlying the SRI and the target analysis, further research focusing on other statistical analyses might help to elucidate this issue.

Appendix

Answer questions by selecting the number on the answer sheet which best describes your level of agreement with the statement according to the following scale.

(1) strongly agree, (2) agree, (3) undecided, (4) disagree, (5) strongly disagree

1. My instructor displays a clear understanding of course topics.
2. My instructor seems well prepared for class.
3. My instructor emphasizes conceptual understanding of the material.
4. My instructor is careful and precise when answering questions.
5. Exams cover a reasonable amount of material.
6. Exams are fair.
7. My course grade accurately reflects my knowledge of the material.
8. The grading system was clearly explained.
9. I think that the grading system in this course is appropriate.
10. My instructor returned graded materials in a timely manner.
11. The teaching strategy used in this course was appropriate.
12. I am generally pleased with the textbook in this course.
13. The assigned reading is well integrated into the course.
14. The amount of material covered in the course is reasonable.
15. The workload in this course was appropriate.
16. This course was effective in improving my understanding of the universe.
17. This course was effective in improving my understanding of natural laws.

18. This course was effective in improving my understanding of the history of science and its relationship with human civilization.

19. This course was effective in improving my awareness of the space program and its impact on mankind.

20. This course was effective in improving my understanding of the nature of science in general.

21. This instructor was one of the better University Core Curriculum instructors I have had at this University.

22. This course is an effective University Core Curriculum Course.

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. (2006). Imputation of categorical variables with PROC MI. Paper presented at the annual meeting of the SAS Users Group International, San Francisco, CA.

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

de Leeuw, E. D. , Hox, J. and Husman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics* **19**, 153-176.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement* **64**, 419-436.

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

Leite, W. L. and 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.

Little, R. J. A. , and Rubin, D. B. (2002). Statistical analysis with missing data. John Wiley.

McKnight, P. E. , McKnight, K. M. , Sidani, S. and Figueredo, A. J. (2007). *Missing Data: A gentle introduction*. The Guilford Press.

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

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

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

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

Schafer, J. L. , and 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. , and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods* **6**, 317-329.

Sulis, I. and Porcu, M. (2008). Assessing the effectiveness of a stochastic regression imputation method for ordered categorical data. Working papers, Centro Ricerche Econoiche Nord Sud.

Received October 9, 2008; accepted December 23, 2008.

Holmes Finch
Department of Educational Psychology
Ball State University
Muncie, IN 47306
whfinch@bsu.edu