5-1-2005

Multiple Imputation For Missing Ordinal Data

Ling Chen
University of Arizona

Marian Toma-Drane
University of South Carolina

Robert F. Valois
University of South Carolina

J. Wanzer Drane
University of South Carolina

Multiple Imputation For Missing Ordinal Data

Ling Chen
University of Arizona

Mariana Toma-Drane
University of South Carolina

Robert F. Valois
University of South Carolina

J. Wanzer Drane
University of South Carolina

Simulations were used to compare complete case analysis of ordinal data with including multivariate normal imputations. MVN methods of imputation were not as good as using only complete cases. Bias and standard errors were measured against coefficients estimated from logistic regression and a standard data set.

Key words: complete case analysis, missing data mechanism, multiple logistic regression

Introduction

Surveys are important sources of information in epidemiologic studies and other research as well, but often encounter missing data (Patricia, 2002). Ordinal variables are very common in survey research; however, they challenge primary data collectors who might need to impute missing values of these variables due to their hierarchical nature but with unequal intervals.

The traditional approach, complete case analysis (CC), excludes from the analysis observations with any missing value among variables of interest (Yuan, 2000). CC remains the most common method in the absence of readily available alternatives in software packages. However, using only complete cases could result in losing information about incomplete cases, thus biasing parameter

Ling Chen is a doctoral student in the Department of Statistics at the University of Missouri, Columbia. Mariana Toma-Drane, is a doctoral student at Norman J. Arnold, School of Public Health, Department of Health Promotion Education and Behavior. John Wanzer Drane is Professor of Biostatistics at USC and Fellow of the American Academy of Health Behavior. Robert F. Valois is Professor Health Promotion, Education and Behavior at USC and a Fellow of the American Academy of Health Behavior.

estimates, and compromising statistical power (Patricia, 2002). Multiple imputation (MI) procedure replaces each missing value with m plausible values generated under an appropriate model. These m multiply imputed datasets are then analyzed separately by using procedures for complete data to obtain desired parameter estimates and standard errors. Results from the m analyses are then combined for inferences by computing the mean of the m parameter estimates and a variance estimate that include both a within-imputation and a between-imputation component (Rubin, 1987).

MI has some desirable features, such as introducing appropriate random error into the imputation process and making it possible to obtain unbiased estimates of all parameters; allowing use of complete-data methods for data analysis; producing more reasonable estimates of standard errors and thereby increasing efficiencies of estimates (Rubin, 1987). In addition, MI can be used with any kind of data and any kind of analysis without specialized software (Allison, 2000). MI appears to be a more attractive method handling missing data in multivariate analysis compared to CC (King et al., 2001; Little & Rubin, 1989).

However, certain requirements should be met to have its attractive properties. First, the data must be missing at random (MAR). Second, the model used to generate the imputed values must be correct in some sense. Third, the model used for the analysis must catch up, in some sense, with the model used in the imputation

(Allison, 2000). All these conditions have been rigorously described by Rubin (1987) and Schafer (1997). The problem is that it is easy to violate these conditions in practice.

The purpose of this study was to investigate how well multivariate normal (MVN) based MI deals with non-normal missing ordinal covariates in multiple logistic regression, while there is definite violation against the distributional assumptions of the missing covariates for the imputation model.

Simulated scenarios were created for the comparison assuming various missing rates for the covariates (5%, 15% and 30%) and different missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing nonignorable (NI). The performance of MVN based MI was compared to CC in each scenario.

Methodology

The mechanism that leads to values of certain variables being missing is a key element in choosing an appropriate analysis and interpreting the results (Little & Rubin, 1987).

In sample survey context, let Y denote an $n \times p$ matrix of multivariate data, which is not fully observed. Let Y_{obs} denote the set of fully observed values of Y and Y_{mis} denote the set containing missing values of Y, i.e., $Y = (Y_{obs}, Y_{mis})$.

Rubin (1976) introduced a missing data indicator matrix R. The (i, j)th element $R_{ij} = 1$ if Y_{ij} is observed; and $R_{ij} = 0$ if Y_{ij} is missing. The notation of missing data mechanisms was formalized in terms of a model for the conditional distribution $P(R \mid Y, \zeta)$ of R given Y according to whether the probability of response depends on Y_{obs} or Y_{mis} or both, where ζ is an unkown parameter.

Data are MCAR, if the distribution of R does not depend on Y_{obs} or Y_{mis}; that is $P(R \mid Y, \zeta) = P(R \mid \zeta)$ for all Y. In this case, the observed values of Y form a random subset of all the sampled values of Y. Data are MAR if the distribution of R depends on the data Y only through the observed values Y_{obs}; that is, $P(R \mid Y, \zeta) = P(R \mid Y_{obs}, \zeta)$ for all Y_{mis}. MAR implies missing depends on observed covariates and outcomes, or missingness can be predicted by

observed information. MCAR is a special case of MAR. The missing data mechanism is ignorable for likelihood-based inferences for both MCAR and MAR (Little & Rubin, 1987). Missing NI occurs when the probability of response of Y depends on the value of Y_{mis} and possibly the value of Y_{obs} as well.

The data used in this investigation are from the 1997 South Carolina Youth Risk Behavior Survey (SCYRBS). The total number of complete and partial questionnaires collected is 5545. The survey employed a two-stage cluster sampling with derived weightings designed to obtain a representative sample of all South Carolina public high school students in grades 9-12, with the exception of those in special education schools. The survey ran from March until June 1997.

The questionnaire covers six categories of priority health-risk behaviors required by the Center for Disease Control and Prevention, and locally, two additional psychological categories of questions were added that include quality of life and life satisfaction (Valois, Zulling, Huebner & Drane, 2001). The six categories of priority health-risk behaviors among youth and young adults are those that contribute to unintentional and intentional injuries; tobacco use; alcohol and other drug use; sexual behaviors; dietary behaviors and physical inactivity (Kolbe, 1990).

The items on self-report youth risk behaviors are Q10 through Q20. The six life-satisfaction variables, Q99 through Q104, are based on six domains: family, friends, school, self, living environment and overall life satisfaction. Each of the questions has seven response options based on the Multidimensional Students' Life Satisfaction Scale (Seligson, Huebner & Valois, 2003). The response options are from the Terrible-to-Delighted Scale: 1 - terrible; 2 - unhappy; 3 - mostly dissatisfied; 4 - equally satisfied and dissatisfied; 5 - mostly satisfied; 6 - pleased; and 7 - delighted (10).

The four race-gender groups: White Females (WF, 26.7%), White Males (WM, 26.0%), Black Females (BF, 26.0%) and Black Males (BM, 21.3%) accounted for almost equal percentage in the sample. The sample was due to the belief that the relationship between life satisfaction and youth risk behaviors varies

across different race-gender groups, as demonstrated in previous research (Valois, Zulling, Huebner & Drane, 2001).

Multiple Logistic Regression Analysis

Exploring the relationship between life satisfaction and youth risk behaviors powered this study. Three covariates in ordinal scale were selected from the 1997 SCYRBS Questionnaire (see the Appendix for details). They were dichotomized as Q10: DRKPASS (Riding with a drunk driver); Q14: GUNSCHL (Carrying a gun or other weapon on school property) and Q18: FIGHTIN (Physical fighting), respectively. Each of them was coded "1" for "never" (0 time) and "2" for "ever" (equal to or greater than 1 time), with "1" as the referent level. All the six ordinal variables of life satisfaction (Q99 ~ Q104) were pooled for each participant to form a pseudo-continuous dependent variable ranging in score from 6 to 42, i.e., "Lifesat = Q99 + Q100 + Q101 + Q102 + Q103 + Q104". The score was expressed as Satisfaction Score (SS) with lower scores indicative of reduced satisfaction with life (Valois, Zulling, Huebner & Drane, 2001). SS ranging from 6 to 27 was categorized as dissatisfied. For the dichotomized outcome variable D2, the students in dissatisfied group (D2 = 1) served as the risk group and the others as the referent group (D2 = 0).

As defined, all the four variables used in logistic regression were dichotomized. DRKPASS, GUNSCHL and FIGHTIN were used as predictor variables while D2 was chosen as the response or criterion variable. The three predictor variables are each independently associated with life dissatisfaction with odds ratios (OR) ranging from 1.42 to 2.27; they are also associated with each other with odds ratios ranging from 2.22 to 4.52.

To use the sampling design in multiple logistic regression analysis, dichotomous logistic regression (PROC MULTILOG) was conducted using SAS-callable Survey Data Analysis (SUDAAN) for weighted data at an alpha level of 0.05 (Shah, Barnwell & Bieler, 1997) (See Appendix.). The analyses were done separately for the four race-gender groups, and the regression coefficient (β) and the standard error of the regression coefficient (Se (β)) for each covariate were obtained.

Simulations

Simulations were applied to compare the performance of CC and MI in estimating regression. Create a complete standard dataset. The SAS MI procedure was used to impute the very few missing values in the youth risk behavior variables (Q10 through Q20) and the six life-satisfaction variables (Q99 through Q104) in the 1997 SCYRBS Dataset *once*, because missing percentages of these variables are very low, ranging from 0.13% to 4.11%. The resulting dataset was regarded as the Complete Standard Dataset in the simulations. This dataset was considered the true gold standard and some values of the three variables related to the three predictors in logistic regression were set to be missing. The PROC MI code (see Appendix) used to create the Complete Standard Dataset was the same as that used to impute values for missing covariates except that missing values were imputed five times in the simulations. The distributions of the three ordinal covariates in the Complete Standard Dataset were also examined. The three covariates are all highly skewed instead of being approximately normal (Figure 1).

Simulating datasets with missing covariates

Three missing data mechanisms were simulated: MCAR, MAR and NI. For the case of MCAR, each simulated sample began by *randomly* deleting a certain percentage of the values of Q10, Q14 and Q18 from the Complete Standard Dataset such that the three covariates were missing at the same rate (5%, 15% and 30%).

For MAR, a certain percentage of values of Q10 were removed from the Complete Standard Dataset with a probability related to the outcome variable (D2) and the other two variables Q14 and Q18. For the NI condition, a certain percentage of values of Q10 were removed such that the larger values of Q10 were more likely to be missing, as in real datasets some covariates corresponding to sensitive matters, whether large or small, their responses are often more likely to be missing (Wu & Wu, 2001). For all the scenarios assuming MAR and NI, Q14 and Q18 were randomly removed assuming MCAR at the same rate as Q10.

Figure 1. Distribution of the three covariates in the Complete Standard Dataset.
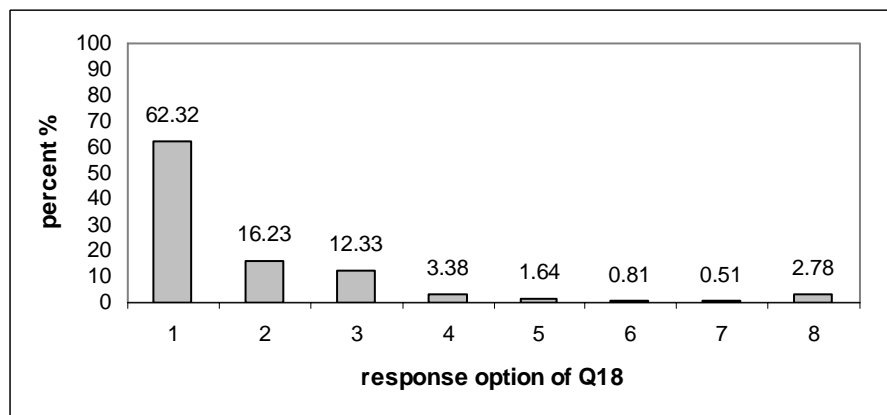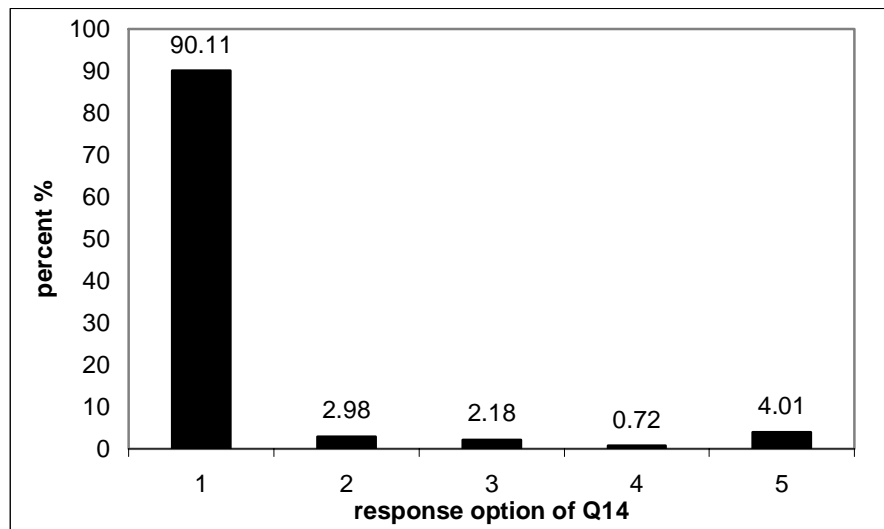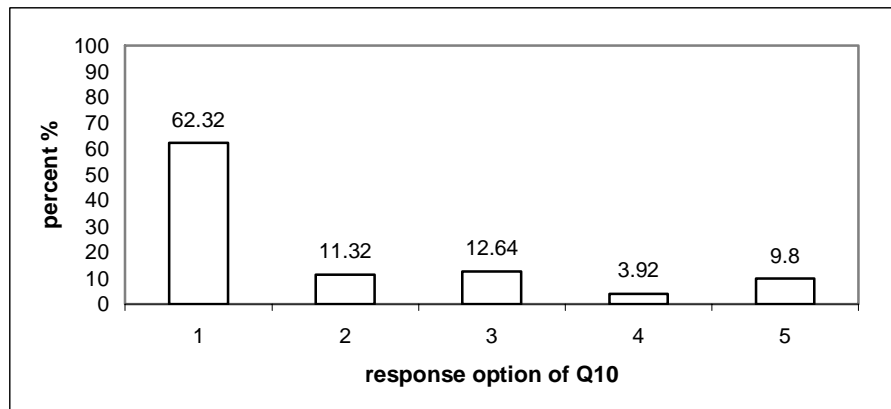
Nine scenarios were created where the covariates Q10 (DRKPASS), Q14 (GUNSCHL) and Q18 (FIGHTIN) were missing at the same rate (5%, 15% and 30%), the life-satisfaction variables (Q99 ~ Q104) were complete as in the Complete Standard Dataset, however. In each scenario 500 datasets with missing covariates were generated. Table 1 lists the missing data mechanisms for the covariates, and the average percentage of complete cases (all the three covariates complete) in the 500 datasets for each scenario. All the simulations were performed using SAS version 8.2 (2002).

Multiple Imputation

The missing covariates in each simulated dataset were then imputed five times using the SAS MI procedure (see Appendix). First, initial parameter estimates were obtained by running the Expectation-Maximization (EM) algorithm until convergence up to a maximum of 1000 iterations. Using the EM estimates as starting values, 500 cycles were ran of Markov Chain Monte Carlo (MCMC) full-data augmentation under a ridge prior with the hyperparameter set to 0.75 to generate five imputations. A multivariate normal model was applied to the data augmentation for the non-normal ordinal data without trying to meet the distributional assumptions of the imputation model.

Three auxiliary variables (Q11, Q13 and Q19) as well as the outcome variable D2 were entered into the imputation model as if they were jointly normal, to increase the accuracy of the imputed values of Q10, Q14 and Q18 (Allison, 2000; Schafer, 1997 & 1998; Rubin, 1996).

The maximum and minimum values for the imputed values were specified, which were based on the scale of the response options for the 1997 SCYRBS questions. These specifications were necessary so that the imputations were not made outside of the range of the original variables. The continuously distributed imputes for Q10, Q14 and Q18 were rounded to the nearest category using a cutoff value of 0.5.

Inferences from CC and MI

For inference from CC, multiple logistic regression analysis was performed for each of the 500 datasets with missing covariates. The estimates for β and Se (β) for CC in each scenario were the average of the 500 estimates from the 500 incomplete datasets, respectively. For inference from MI, The point estimate of β was first obtained from the five imputed dataset estimates; and Se (β) was obtained by combining the within-imputation variance and between-imputation variance from the five repeated imputations (Rubin, 1987; SAS Institute, 2002). The estimates for β and Se (β) for MI in each scenario were the average of the 500 point estimates of β and the 500 combined Se (β), respectively.

Comparison of complete case and multiple imputation model results

To compare the performance of CC and MI, biases and standard errors of point estimates were mainly considered. Each regression coefficient calculated from the Complete Standard Dataset was taken as the true coefficient and those from CC and MI in each scenario were compared to the true ones. Bias is expressed as estimate from CC or MI minus the estimate from the Complete Standard Dataset, i.e., estimated $\beta - \beta_{\text{true value}}$. The average absolute value of bias (AVB) of β for each covariate was compared between the two methods for the same race-gender group.

Results

The missing values in the risk behavior and life-satisfaction variables were imputed, and the resulting dataset was defined as the Complete Standard Dataset as if it was originally complete. Table 2 contains the estimates and standard errors of the regression coefficients from the 1997 SCYRBS dataset together with those from the Complete Standard Dataset. Given the low percentages of missing variables in 1997 SCYRBS dataset and thus the few cases omitted from the CC, the results from the two datasets are very similar.

Table 1. Simulated scenarios for datasets with missing covariates.

Scenario	Missing percentage of each covariate	Average percentage of complete cases	Missing data mechanism for each covariate		
			Q10 (DRKPASS)	Q14 (GUNSCHL)	Q18 (FIGHTIN)
1	5%	85.73%	MCAR	MCAR	MCAR
2	5%	85.42%	MAR	MCAR	MCAR
3	5%	85.55%	NI	MCAR	MCAR
4	15%	61.34%	MCAR	MCAR	MCAR
5	15%	61.19%	MAR	MCAR	MCAR
6	15%	62.54%	NI	MCAR	MCAR
7	30%	34.22%	MCAR	MCAR	MCAR
8	30%	34.30%	MAR	MCAR	MCAR
9	30%	34.10%	NI	MCAR	MCAR

Table 2. Logistic regression coefficients and standard error estimates in the 1997 SCYRBS Dataset and the Complete Standard Dataset.

Group	DRKPASS		GUNSCHL		FIGHTIN	
	β *	Se(β) †	β	Se(β)	β	Se(β)
White female	0.14	0.10	0.99	0.21	0.88	0.16
N=1359 (1361) ‡	(0.16)	(0.11)	(0.94)	(0.23)	(0.84)	(0.16)
Black female	0.03	0.14	0.69	0.28	0.36	0.15
N=1335 (1336)	(0.02)	(0.14)	(0.63)	(0.24)	(0.45)	(0.16)
White male	0.32	0.17	0.10	0.17	0.43	0.13
N=1338 (1340)	(0.25)	(0.16)	(0.32)	(0.15)	(0.53)	(0.11)
Black male	0.43	0.16	0.95	0.20	0.32	0.11
N=1119 (1119)	(0.35)	(0.14)	(0.94)	(0.23)	(0.52)	(0.11)

* β, logistic regression coefficient.

† Se (β), standard error of logistic regression coefficient.

‡ Numbers in parentheses, sample size, logistic regression coefficient and standard error of logistic regression coefficient from the Complete Standard Dataset.

An example is presented from comparing CC and MI across the nine scenarios among White Females in table 3. The histogram of the average AVB of β for each covariate in this example is shown in figure 2. To evaluate the the imputation procedure, the absolute value of bias in point estimates and coverage probability were mainly considered. The coverage probability is defined as the possibility of the true regression coefficient β being covered by the actual 95 percent confidence interval. Further, the percent AVB of β for each covariate, calculated by dividing AVB by the corresponding true β, better compares the two methods with regard to bias. Greater or equal to10% of bias is beyond acceptance.

Both CC and MI produced biased estimates of β in all the scenarios. CC showed little or no bias for all the scenarios under MCAR. The AVB of β for each covariate is consistently less than 0.05 for all the three covariates even with about 34% complete cases (30% missing for each covariate). However, CC showed larger AVB's of β in the scenarios under MAR and NI than in those under MCAR with the same missing covariate rates. Further, MI was generally less successful than CC because MI showed larger AVB's of β than CC in most of the scenarios regardless of missing data mechanism and missing covariate rate. (Results for the other three race-gender groups not shown here.)

Figure 2. Average AVB's (absolute value of bias) of logistic regression coefficients across the nine scenarios among White Females. S1 ~ S9 represent Scenario1 ~ Scenario 9, respectively.
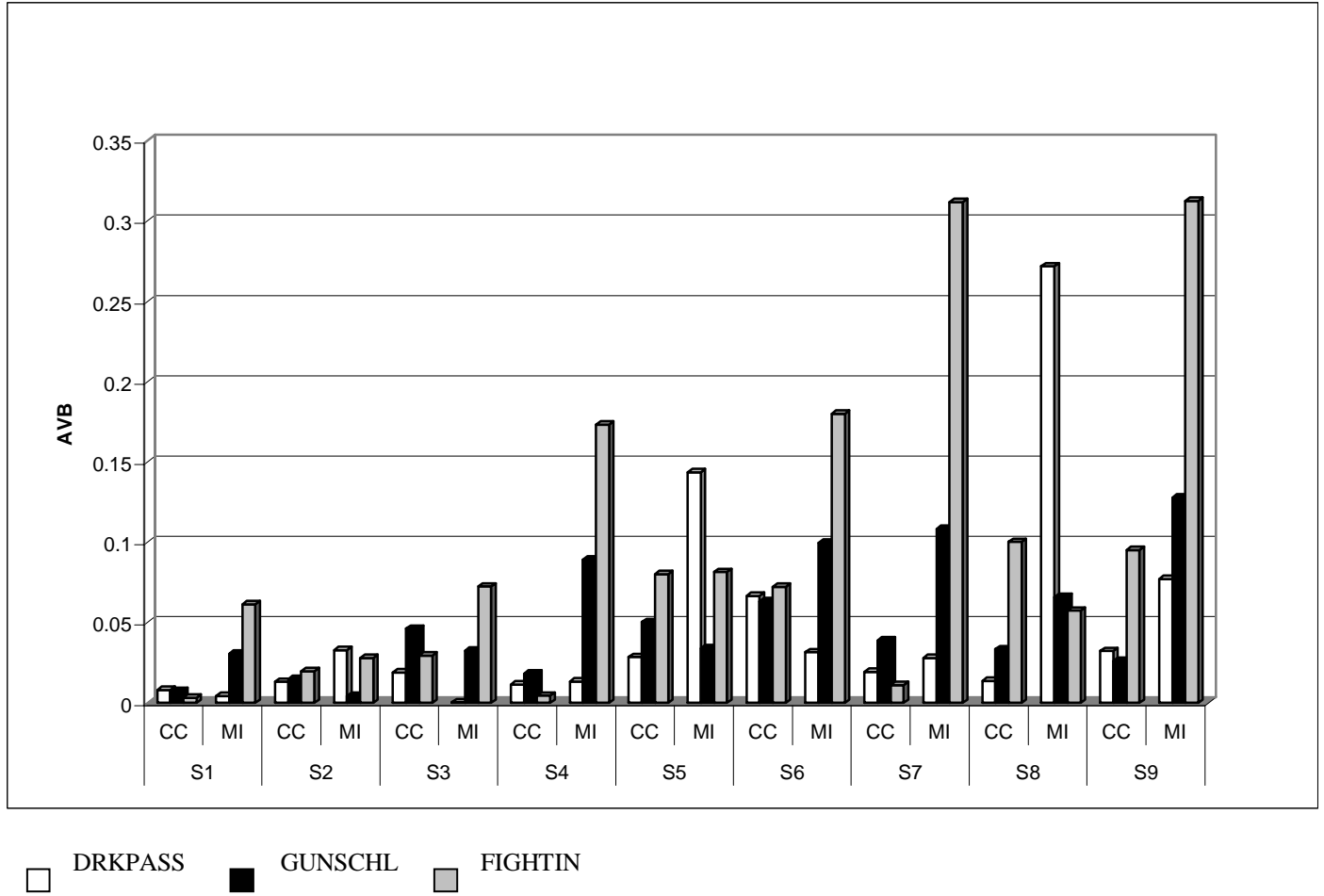
Table 3. Comparison of complete case and multiple imputation model results across the nine scenarios among White Females.

		DRKPASS β *true value = 0.16 Se (β) true value = 0.11		GUNSCHL β true value = 0.94 Se (β) true value = 0.23		FIGHTIIN β true value = 0.84 Se (β) true value = 0.16	
		AVB ‡	Se(β) †	AVB	Se(β)	AVB	Se(β)
Scenario 1	CC	0.0082	0.1178	0.0075	0.2574	0.0033	0.1740
	MI	0.0043	0.1088	0.0306	0.2414	0.0613	0.1627
Scenario 2	CC	0.0132	0.1190	0.0148	0.2725	0.0198	0.1768
	MI	0.0329	0.1094	0.0044	0.2448	0.0280	0.1602
Scenario 3	CC	0.0189	0.1162	0.0462	0.2712	0.0295	0.1759
	MI	0.0004	0.1061	0.0324	0.2470	0.0725	0.1593
Scenario 4	CC	0.0116	0.1467	0.0182	0.3302	0.0046	0.1964
	MI	0.0133	0.1151	0.0893	0.2591	0.1732	0.1574
Scenario 5	CC	0.0286	0.1521	0.0504	0.3666	0.0802	0.2039
	MI	0.1437	0.1166	0.0339	0.2610	0.0815	0.1555
Scenario 6	CC	0.0667	0.1451	0.0633	0.3517	0.0724	0.2105
	MI	0.0315	0.1137	0.0996	0.2628	0.1800	0.1556
Scenario 7	CC	0.0194	0.2097	0.0390	0.4840	0.0111	0.2478
	MI	0.0279	0.1237	0.1083	0.2704	0.3118	0.1523
Scenario 8	CC	0.0138	0.1991	0.0335	0.4312	0.1002	0.2347
	MI	0.2718	0.1227	0.0660	0.2738	0.0575	0.1500
Scenario 9	CC	0.0323	0.2227	0.0261	0.5611	0.0951	0.2661
	MI	0.0771	0.1344	0.1278	0.2828	0.3126	0.1506

* β, logistic regression coefficient.
† Se (β), standard error of logistic regression coefficient.
‡ AVB, absolute value of bias (| estimated $\beta - \beta$ true value |).

Table 4. Coverage probability in Scenarios 2 and 8 for White Females.

		DRKPASS (%)	GUNSCHL (%)	FIGHTIN (%)
Scenario 2	CC	96.8	96.4	95.0
	MI	94.2	99.0	93.8
Scenario 8	CC	95.0	94.0	87.0
	MI	77.0	90.4	88.2

Table 5. Average Correct Imputation Rate for the three covariates.

Transformation *	Scenario	Original scale (%)			Recoded (%)		
		Q10	Q14	Q18	DRKPASS	GUNSCHL	FIGHTIN
Without	2	15.94	83.20	31.40	47.81	86.77	49.20
	8	21.25	83.22	29.21	41.05	86.75	47.39
With	2	40.04	89.47	50.80	65.14	92.11	66.00
	8	52.40	89.54	50.75	65.52	91.81	63.22

Also, in most scenarios the percent AVB of β from MI is far greater than that from CC and is greater than 10% of acceptance level. This discrepancy was especially obvious for all the scenarios under MCAR (Scenarios 1, 4 and 7). Moreover, the AVB's and percent AVB's from MI increase substantially as larger proportions of the covariates were missing. Interestingly, MI showed consistently decreased Se (β) for each covariate in all the scenarios, which is not surprising, because the standard error of MI is based on full datasets (Allison, 2001).

Table 4 lists the coverage probabilities in Scenarios 2 and 8 among White Females as an example. In both scenarios, the coverage probabilities from MI are not all better than those from CC.

Clearly, the current MVN based multiple imputation did not perform as well as CC in generating unbiased regression estimates. To investigate how well the present MI actually imputed the missing non-normal ordinal covariates, Scenarios 2 and 8 were used to check the imputation efficiency, as the two scenarios have the same setting for missing data mechanism but different missing covariate rates. The Average Correct Imputation Rate is calculated as the average proportion of correctly imputed observations among the missing covariates. Correct imputation occurs when the imputed value is identical to its true value in the Complete Standard Dataset. Table 5 displays the Average Correct Imputation Rates for the three covariates in both original scales (Q10, Q14 and Q18) and recoded scales (DRKPASS, GUNSCHL and FIGHTIN).

The Average Correct Imputation Rates for Q10 and Q18 are lower than 32% in both scenarios. Recoding helped to improve imputation efficiency for all the three covariates, this can be explained by the loss of precision after recoding. Surprisingly, the Average Correct Imputation Rates for Q14 (GUNSCHL) are very close in the two scenarios. In addition, they are consistently and considerably higher than those for the other two covariates. This may be explained by the fact that a vast majority of its observations fall into one category (figure 1).

Natural logarithmic transformation on the three covariates was also attempted before multiple imputation to approximate normal variables and to fit the distributional assumptions of the imputation model. The Average Correct Imputation Rates for Q10 and Q18 in original and recoded scales in both scenarios improved as compared to before the transformation, but still not satisfactory (below 53%). Nevertheless, the majority of Q14 (above 89%) in both scales have been correctly imputed.

Also examined was the effect of rounding on imputation efficiency, because the continuously distributed imputes have been rounded to the nearest category using a cutoff value 0.5 to preserve their ordinal property. For illustration an example is presented using a random dataset with missing covariates created in Scenario 8. The $50^{th} \sim 65^{th}$ observations of Q10 in this dataset are listed in table 6 along with their five imputed values in the same manner as in the simulations but without rounding the continuous imputed values. A large proportion (34 out of 50) of the imputed values is in different categories from their true values after being rounded using the cutoff value 0.5.

The prevalence of dissatisfaction, D2 = 1, ranges from 0.58% to 6.95% among the four race-gender groups. Interestingly, even with such low frequencies of the outcome (D2 = 1), all the covariates are significantly related to the outcome with odds of dissatisfaction with the trait present ranges from 1.42 to 2.27 times the same odds when the trait is absent. The three traits DRKPASS, GUNSCHL and FIGHTIN are strongly associated with each other with odds ratios between the traits ranging from 2.22 to 4.52. The significant associations between the four variables support these four variables as objects of our study of imputations on their values and whether imputation removes biases under these conditions.

In this study, CC showed smaller bias in the scenarios assuming MCAR for each covariate than in those with MAR and NI, regardless of proportions of missing covariates. This is consistent with the study by Allison (2000). The finding that the scenarios under NI showed relatively large biases in CC as compared to the MCAR conditions is also in accordance with King et al. (2001).

Table 6. Five Imputations for missing Q10 without rounding on imputed values from one random dataset in Scenario 8.

Obs.	Q10	True value	Imputation number				
			1	2	3	4	5
50	2	2					
51	.	1	1.7823 *	1.6022 *	1.7633 *	1.8180 *	1.3918
52	2	2					
53	.	1	1.3587	2.0277	1.8274 *	1.6079 *	1.4763
54	.	2	2.1264	2.4809	2.3249	2.0099	2.0358
55	.	1	1.7062 *	1.6104 *	1.6476 *	1.5978 *	1.6790 *
56	1	1					
57	.	1	1.4641	1.8700 *	1.5210 *	1.2140	1.5401 *
58	.	1	1.9022 *	1.8579 *	1.6802 *	1.7611 *	1.5634 *
59	1	1					
60	1	1					
61	.	1	1.5195 *	1.6148 *	1.5551 *	1.9029 *	1.5423 *
62	.	1	1.6313 *	1.6186 *	1.7034 *	1.4602	1.8294 *
63	.	1	1.6788 *	1.6553 *	1.6355 *	1.6657 *	1.5695 *
64	.	2	1.7022	2.0307	1.7366	1.4447 *	1.8448
65	1	1					

Accumulating evidence suggests that MI is usually better than, and almost always not worse than CC (Wu & Wu, 2001; Schafer, 1998; Allison, 2001; Little, 1992). Evidence provided by Schafer (1997, 2000) demonstrated that incomplete categorical (ordinal) data can often be imputed reasonably from algorithms based on a MVN model. However, our study did not show consistent results with the findings from Schafer, this is mainly due to ignorance of assumption of normality.

It is known that sensitivity to model assumptions is an important issue regarding the consistency and efficiency of normal maximum likelihood method applied to incomplete data. The improved, though unsatisfactory, imputation after natural logarithmic transformation presented a good demonstration of the importance of sensitivity to normal model assumption.

Moreover, normal ML methods do not guarantee consistent estimates, and they are certainly not necessarily efficient when the data are non-normal (Little, 1992). The MVN based MI procedure not specifically tailored to highly skewed ordinal data may have seriously distorted the ordinal variables' distributions or their relationship with other variables in our study, and therefore is not reliable when imputing highly skewed ordinal data.

It was suggested that and highly skewed variables may well be transformed to approximate normality (Tabachnick & Fidell, 2000). Nevertheless, highly skewed ordinal variables with only four or five values can hardly be transformed to nearly normal variables as shown by the unsatisfactory imputation efficiencies after natural logarithmic transformation. This study gives a warning that doing imputation without checking distributional assumptions of imputation model can lead to worse trouble than not imputing at all.

In addition, rounding after MI should be further explored in terms of appropriate cutoff values. One is cautioned that rounding could also bring its own bias into regression analysis in multiple imputations of categorical variables.

Conclusion

Applied researchers can be reasonably confident in utilizing CC to generate unbiased regression estimates even when large proportions of data missing completely at random. For ordinal variables with highly skewed distributions, MVN based MI cannot be expected to be superior to CC in generating unbiased regression estimates. It is cautionary that researchers doing imputation without checking distributional assumptions of imputation model can get into worse trouble than not imputing at all.

References

Allison P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2000). Multiple imputation for missing data: a cautionary tale. *Sociological Methods & Research, 28*, 301-9.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review, 95* (1), 49-69.

Kolbe, L. J. (1990). An epidemiologic surveillance system to monitor the prevalence of youth behaviors that most affect health. *Journal of Health Education*, *21*(6), 44-48.

Little R. J. A. (1992). Regression with missing X's: a review. *Journal of American Statistical Association*, 87, 1227-37.

Little, R. J. A., & Rubin D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 18,* 292-326.

Little, R. J. A., & Rubin D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.

Patrician, P. A. (2002). Focus on research methods multiple imputation for missing data. *Research in Nursing & Health*, *25*, 76-84.

Rubin D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91 (434), 473-89.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-92.

SAS/STAT Software (2002), Changes and Enhancements, Release 8.2, Cary, NC: SAS Institute Inc.

Schafer J. L., & Olsen M.K. (2000). Modeling and imputation of semicontinuous survey variables. Federal Committee on Statistical Methodology Research Conference: Complete Proceedings, 2000.

Schafer J. L. (1998). *The practice of multiple imputation*. Presented at the meeting of the 17. Methodology Center, Pennsylvania State University, University Park, PA.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*, New York: Chapman & Hall.

Seligson J. L., Huebner E. S., & Valois R. F. (2003) Preliminary validation of the Brief Multidimensional Student's Life Satisfaction Scale (BMSLSS). *Social Indicators Research*, *61*, 121-45.

Shah B. V., Barnwell G. B., & Bieler G. S. (1997). *SUDAAN*, software for the statistical analysis of correlated data, User's Manual. Release 7.5 ed. Research Triangle Park, NC: Research Triangle Institute.

Tabachnick B.G., & Fidell L.S. (2000). *Using multivariate statistics* (4th ed.). New York: HarperCollins College Publishers.

Valois, R. F., Zulling K. J., Huebner E. S., & Drane J. W. (2001). Relationship between life satisfaction and violent behaviors among adolescents. *American Journal of Health Behavior*, *25*(4), 353-66.

Wu H., & Wu L. (2001). A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. *Statistics in Medicine*, 20, 1755-69.

Yuan, Y. C. (2000). Multiple imputation for missing data: concepts and new developments. *SAS Institute Technical Report*, 267-78.

Appendix A: 1997 SCYRBS Questionnaire items associated with the three covariates in regression analysis

Question 10 (Q10). During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?
1. 0 times
2. 1 time
3. 2 or 3 times
4. 4 or 5 times
5. 6 or more times

Question 14 (Q14). During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club on school property?
1. 0 days
2. 1 day
3. 2 or 3 days
4. 4 or 5 days
5. 6 or more days

Question 18 (Q18). During the past 12 months, how many times were you in a physical fight?
1. 0 times
2. 1 time
3. 2 or 3 times
4. 4 or 5 times
5. 6 or 7 times
6. 8 or 9 times
7. 10 or 11 times
8. 12 or more times

Appendix B: SAS Code

SAS PROC MI code for multiple imputation

```
proc mi data=first.c&I out=outmi&I seed=6666 nimpute=5
minimum=1 1 1 1 1 1 0  maximum=5 5 5 5 8 5 1 round=1 noprint;
em maxiter=1000 converge=1E-10;
mcmc impute=full initial=em prior=ridge=0.75 niter=500 nbiter=500;
freq weight;
var Q10 Q11 Q13 Q14 Q18 Q19 D2;
run;
```

Appendix C: SUDAAN Code

SUDAAN PROC MULTILOG code for multiple logistic regression analysis

```
Proc multilog data=stand filetype=sas design=wr noprint;
nest stratum psu;
weight weight;
subpopn sexrace=1  / name="white female";
subgroup D2 drkpass gunschl fightin;
levels 2 2 2 2;
reflevel drkpass=1 gunschl=1 fightin=1;
model D2 = drkpass gunschl fightin;
output beta sebeta/filename=junk_2 filetype=sas;
run;
```