Multiple Imputation of Categorical Variables Under the Multivariate Normal Model

Paul D. Allison, University of Pennsylvania

ABSTRACT

The most widely used method of multiple imputation is the MCMC algorithm based on the multivariate normal model. While this method is often used to impute binary and polytomous data, there is a natural concern about the consequences of violating the normality and linearity assumptions. Recent work by Horton et al. suggests that the practice of rounding imputed values for binary variables may produce biased estimates of proportions, but that still leaves many questions unanswered. This paper uses simulations to address several questions: How much bias is introduced if the imputed values are left unrounded? What factors affect the degree of bias? What effect do these choices have on bias in regression coefficients for binary variables that are predictor variables? The paper also compares the MCMC method to imputation methods based on a logistic regression model or a linear discriminant model for monotonic missing data patterns. Key conclusions are that linear imputation with rounding is always inferior to linear imputation without rounding. The latter does well under most conditions, except when estimating proportions that are near 0 or 1.

INTRODUCTION

The most popular method for multiple imputation of missing data is the Markov Chain Monte Carlo (MCMC) algorithm based on the assumption of multivariate normality (Schafer 1997), which implies that valid imputations may be generated by linear regression equations. The reasons for the popularity of MCMC are not difficult to fathom. The algorithm is widely available, computationally efficient, rarely breaks down, and, most importantly, can handle arbitrary patterns of missing data. Nevertheless, because the method assumes normality and

linearity, it may not be well suited for imputing categorical variables. For a binary (0,1) variable, for example, the imputed values can be any real value rather than being restricted to 0 and 1. Although most imputed values will be within the (0,1) interval, many will fall outside that range. Several authors (e.g., Schafer 1997, Allison 2001) have recommended rounding the imputed values so that imputed values greater than or equal to .5 are set to 1 and anything less is set to 0. However, Horton et al. (2003) have shown that such rounding can produce biased estimates of proportions, especially when the true proportion is near 0 or 1.

This paper uses simulated data to evaluate various approaches to the imputation of binary variables under the multivariate normality. I begin by focusing on the estimation of proportions, comparing complete case analysis, linear imputation with rounding, linear imputation without rounding, and methods based on logistic regression and the discriminant function.

ESTIMATION OF PROPORTIONS

Horton et al. (2003) analytically investigated a case with the following conditions: there is a single dummy (0,1) variable D with an expected value p, the goal is to estimate p, there are no covariates, and some data on D are missing completely at random (MCAR). They derived formulas for the expected value of two multiple imputation estimators of p, both based on the multivariate normal model, one with rounding and one without rounding. They found that the unrounded estimator was unbiased but the rounded estimator was biased. The relative bias increased as p approached zero and as the fraction of missing data increased.

I extend their conditions in two ways. I introduce a covariate X which is moderately correlated with D and which can be used to improve the imputation of D. Second, I consider the case in which the data on D are missing at random, but not missing completely at random. Specifically, I allow the probability of missingness on D to depend on the covariate X.

For each condition, I draw 500 samples, each with 500 cases. The dummy variable *D* is drawn from a Bernoulli distribution with probability *p*. In the different conditions, *p* was .50, .20, .05, or .01. For each observation *i*, the covariate *X* is generated by the linear equation

$$X_i = -1 + D_i + \tau\varepsilon_i \qquad (1)$$

where ε is a random draw from a standard normal distribution and τ is a constant chosen to control the correlation between *X* and *D*. Although that correlation varied somewhat from condition to condition, it was always between .4 and .5. Note that equation (1) implies that the expected value of *D* given *X* is described by a logistic regression function.

I then set approximately 50 percent of the observations on *D* to be missing. For the MCAR case, I simply made a random draw from a Bernoulli distribution with probability of .5. If the drawn value was 1, the value of *D* was missing. For the MAR condition, a probability π_i was generated by the following equation:

$$\pi_i = 1/[1 + \exp(\alpha_0 + \alpha_1 X_i)] \qquad (2)$$

For each *i*, a random Bernoulli draw was made with probability π_i. Again, if the drawn value was 1, the value of *D* was missing, otherwise not. The parameters α_0 and α_1 were adjusted for each condition to keep the proportion of cases with missing data at .50 and the correlation between *X* and the missingness indicator between .30 and .45.

To estimate the proportion *p*, five different missing data techniques were applied to each sample:

1. Complete Case Analysis (Listwise Deletion). Cases with missing data on *D* were deleted from the sample and the mean of *D* was calculated. Since the probability of missingness was .5, approximately 250 cases were deleted from each sample of 500 cases.

2. Linear Imputation Without Rounding. To generate imputations under the multivariate normal model, I used PROC MI in SAS with *X* as a covariate. Five completed data sets were produced for each sample, the mean of *D* was estimated for each data set using PROC MEANS, and the results were combined using PROC MIANALYZE.

3. Linear Imputation With Rounding. This technique simply took the completed data sets produced by PROC MI in method 2, and rounded the imputed values of *D* to 0 or 1. The rule was that any value greater than or equal to .5 was assigned a value of 1 and anything less than .5 was assigned a value of 0. Using these rounded values, the mean of *D* was estimated for each data set using PROC MEANS, and the results were combined using PROC MIANALYZE.

4. Logistic Regression Imputation. This method is only available for monotonic missing data patterns in PROC MI. Using complete cases, a logistic regression is estimated by maximum likelihood. For each completed data set, a random draw is made from the posterior distribution of the parameters. Based on the resulting logistic regression equation, a probability is generated for each case with missing data and a Bernoulli draw is made for that probability, producing imputed values of 0 or 1. As in methods 2 and 3, five completed data sets were produced for each sample, the mean of *D* was estimated for each data set using PROC MEANS, and the results were combined using PROC MIANALYZE.

5. Discriminant Function Imputation. This method is only available for monotonic missing data patterns in PROC MI. The method is based on the assumption that within each category of the categorical variable, the quantitative variables have a multivariate normal distribution with means that vary across categories but a covariance matrix that is constant over categories. Although this implies a logistic regression for the dependence of the categorical variable on the covariates, the estimation method is based on estimating the means and covariance matrix for the complete cases. Compared with the logistic method, the discriminant function method is faster

and minimizes potential breakdowns in the estimation process. Again, the imputed values are necessarily 0 or 1. As with the other imputation methods, five completed data sets were produced for each sample, the mean of *D* was estimated for each data set using PROC MEANS, and the results were combined using PROC MIANALYZE.

RESULTS FOR MCAR

The SAS code for producing the imputations and calculating the statistics can be found in the Appendix. Table 1 gives the results for the MCAR conditions. For each condition and method, I report the mean of the 500 estimated proportions, the mean of the estimated standard errors, and the standard deviation of the estimates. If the standard error estimates are accurate, their mean should be close to the standard deviation. The standard deviation is also a good measure of the efficiency of the method. The last column of the table gives the proportion of nominal 95% confidence intervals that contain the true value. Obviously, if a method is performing well, that proportion should be close to .95.

When the true proportion is .50, all five methods appear to be unbiased, have about equal standard deviations, and have accurate coverage. The only surprising thing is that the standard deviation for complete case analysis is about the same as for the imputation methods. This is also the case for all the other parameter values in the table. With only half the original 500 cases, one would expect the complete case estimator to be noticeably less efficient than the imputation methods. The latter use information from the covariate to generate the imputations. The covariate is correlated with the dummy variable at around .40, which would seem large enough to give some advantage to the imputation methods. But even when I increased the correlation to .85, the standard deviations for the imputation methods were not lower than the complete case standard deviation.

Table 1. Estimates of Proportions When Data are MCAR

Method	Parameter[b]	Estimate[c]	Avg. SE[d]	Stan. Dev.[e]	Coverage[f]
Complete Case	.50	.501	.032	.032	.94
Linear No Round	.50	.501	.030	.031	.94
Linear Round	.50	.501	.029	.028	.95
Discriminant	.50	.499	.032	.031	.94
Logistic	.50	.501	.030	.031	.94
Complete Case	.20	.200	.025	.025	.96
Linear No Round	.20	.199	.024	.024	.96
Linear Round	.20	.213	.024	.026	.93
Discriminant	.20	.200	.024	.023	.97
Logistic	.20	.200	.024	.024	.94
Complete Case	.05	.050	.014	.014	.91
Linear No Round	.05	.049	.013	.014	.91
Linear Round	.05	.037	.010	.014	.61
Discriminant	.05	.050	.013	.013	.94
Logistic	.05	.050	.013	.013	.92
Complete Case	.01	.011	.006	.006	.94
Linear No Round	.01	.010	.006	.006	.91
Linear Round	.01	.006	.001	.004	.16
Discriminant	.01	.009	.005	.006	.85
Logistic[g]	.01	--	--	--	--

[b]True value of the parameter in the model producing the data.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.
[g]Method failed, presumably due to quasi-complete separation.

When the true proportion is .20, we find some evidence that linear imputation with rounding produces slightly biased estimates with a mean of .213. This is consistent with Horton et al.'s analytical results which showed that upward bias peaked when the parameter was around .20. We also see a slight worsening of coverage for this method. By contrast, linear imputation without rounding does about as well as the logistic and discriminant methods.

When the true proportion is .05, performance of linear imputation without rounding is still pretty good, while rounding seriously degrades the inferences. The estimate of the proportion is biased downward and, more seriously, coverage declines to only 61 percent. This may be a consequence of underestimation of the standard errors. Finally, when the true parameter is .01, the rounded imputations produce completely unacceptable estimates. The estimate is about 40% below the true value and the coverage is only 16 percent. Interestingly, the linear method without rounding actually does a little better than the discriminant method.

RESULTS FOR MAR

Table 2 gives results for data that are missing at random but not missing completely at random. As in the MCAR case, complete case analysis produces estimates whose standard deviations are no larger than those for the imputation methods. Now, however, the complete case estimates are severely biased downward, to the point of being unacceptable for all parameter values. When the true proportion is .50, all four imputation methods do a pretty good job. But when the true proportion is .20, both linear methods show some downward bias, as well as coverage rates that are only around .80. The rounded method is a little worse than the unrounded method.

Table 2. Estimates of Proportions When Data Are MAR But Not MCAR.

Method	Parameter[b]	Estimate[c]	Avg. SE[d]	Stan. Dev.[e]	Coverage[f]
Complete Case	.50	.392	.031	.032	.08
Linear No Round	.50	.497	.033	.035	.94
Linear Round	.50	.487	.031	.031	.94
Discriminant	.50	.498	.038	.042	.93
Logistic	.50	.498	.032	.033	.94
Complete Case	.20	.131	.021	.022	.14
Linear No Round	.20	.185	.022	.029	.83
Linear Round	.20	.176	.026	.034	.79
Discriminant	.20	.199	.033	.034	.95
Logistic	.20	.201	.028	.029	.95
Complete Case	.05	.017	.008	.008	.07
Linear No Round	.05	.033	.009	.014	.54
Linear Round	.05	.009	.003	.006	.02
Discriminant	.05	.050	.021	.023	.91
Logistic[g]	.05	--	--	--	--
Complete Case	.01	.0004	.0004	.0001	.10
Linear No Round	.01	.001	.0004	.003	.10
Linear Round	.01	.0002	4×10^{-7}	1×10^{-6}	.00
Discriminant	.01	.006	.004	.008	.47
Logistic[g]	.01	--	--	--	--

[b]True value of the parameter in the model producing the data.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.
[g]Method failed, presumably due to quasi-complete separation.

When the true proportion is .05, the problems with the linear methods worsen. The unrounded method is perhaps marginally acceptable, but the rounded method is abominable, both with respect to bias and coverage (only 2 percent for a nominal 95 percent interval). Finally, when the true proportion is .01, none of the methods is any good—even the discriminant

method has coverage below 50 percent. Note also that when the true proportion is near 0, the logistic method runs into insurmountable computational problems. (The SAS log reports an attempt to divide by 0). My guess is that this is due to complete or quasi-complete separation, which implies that the maximum likelihood estimate of the regression coefficient does not exist.

ESTIMATION OF REGRESSION COEFFICIENTS

The initial steps in the simulations for regression estimation were just like those for estimating proportions. For each condition, I drew 500 samples, each with 500 cases. In each sample, the dummy variable *D* was drawn from a Bernoulli distribution with probability *p*. In the different conditions, *p* was .50, .20, .05, or .01. For each observation *i*, the covariate *X* was generated by equation (1) above. As before, for each condition τ was chosen to keep the correlation between *X* and *D* between .3 and .4.

Once *X* and *D* had been produced, observations on a response variable *Y* were generated by the following linear equation:

$$Y_i = \beta_1 D_i + \beta_2 X_i + \sigma \upsilon_i \qquad (3)$$

where υ_i is random draw from a standard normal distribution. For the regressions reported here, σ was set equal to 3 and both β_1 and β_2 were set equal to 1.

As with estimation of proportions, I then made about 50 percent of the observations on *D* to be missing in each sample. In the MCAR condition, I simply made random draws from a Bernoulli distribution. For the MAR condition, the probability of missingness for each observation was governed by equation (2). I then applied the five missing data methods to each sample to produce estimates of β_1 and β_2.

MCAR RESULTS

Table 3 gives results for estimation of β_1, the coefficient of the dummy variable, in the MCAR condition. When *p*, the proportion of 1's on the dummy variable, is .05 or greater, all five missing data methods do pretty well, with the possible exception of linear imputation with rounding. The rounding method shows about 15-20 percent downward bias but, surprisingly, coverage is still quite good. When *p* declines to .01, all of the methods suffer some deterioration in coverage. Complete-case analysis and the two linear methods show about 15 percent downward bias. Linear imputation with rounding has terrible coverage, which is somewhat surprising given that the standard errors appear to be overestimated. The discriminant method is fairly free from bias but has less than ideal coverage.

It's important to note that, as in Table 1, the imputation methods have about the same standard errors as the complete case method. Hence, little was gained by imputing the missing data, at least for this coefficient.

Table 3. Estimates of Coefficients of Dummy Variable When Data Are MCAR.

Method	Prop.[b]	Param.[h]	Estimate[c]	Avg. SE[d]	Stan. Dev.[e]	Coverage[f]
Complete Case	.50	1.0	1.021	.437	.445	.95
Linear No Round	.50	1.0	1.004	.437	.458	.94
Linear Round	.50	1.0	.858	.408	.391	.95
Discriminant	.50	1.0	.996	.441	.447	.95
Logistic	.50	1.0	.984	.440	.444	.96
Complete Case	.20	1.0	1.012	.520	.542	.94
Linear No Round	.20	1.0	1.020	.526	.554	.92
Linear Round	.20	1.0	.809	.480	.447	.94
Discriminant	.20	1.0	.989	.522	.513	.94
Logistic	.20	1.0	1.019	.532	.486	.96
Complete Case	.05	1.0	1.035	.933	.948	.95
Linear No Round	.05	1.0	1.006	.953	.980	.93
Linear Round	.05	1.0	.875	.930	.832	.96
Discriminant	.05	1.0	.932	.929	.955	.95
Logistic[g]	.05	1.0	1.026	.944	.969	.95
Complete Case	.01	1.0	.845	2.069	1.991	.89
Linear No Round	.01	1.0	.863	1.953	2.112	.87
Linear Round	.01	1.0	.841	4.178	1.985	.07
Discriminant	.01	1.0	.985	1.871	2.066	.84
Logistic[g]	.01	1.0	--	--	--	--

[b]True value of the proportion for the dummy variable.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.
[g]Method failed, presumably due to quasi-complete separation.
[h]True value of the coefficient in the model generating the data.

Table 4. Estimates of Coefficient of Covariate When Data Are MCAR.

Method	Prop.[b]	Param.[h]	Estimate[c]	Avg. SE[d]	Stan. Dev[e]	Coverage[f]
Complete Case	.50	1.0	.991	.213	.215	.94
Linear No Round	.50	1.0	1.003	.169	.184	.93
Linear Round	.50	1.0	1.056	.158	.169	.91
Discriminant	.50	1.0	1.004	.169	.173	.95
Logistic	.50	1.0	1.017	.168	.160	.96
Complete Case	.20	1.0	.998	.212	.206	.95
Linear No Round	.20	1.0	.998	.164	.151	.96
Linear Round	.20	1.0	1.049	.155	.138	.96
Discriminant	.20	1.0	1.001	.163	.159	.95
Logistic	.20	1.0	1.000	.164	.166	.95
Complete Case	.05	1.0	1.003	.212	.204	.96
Linear No Round	.05	1.0	1.000	.155	.150	.96
Linear Round	.05	1.0	1.026	.146	.144	.94
Discriminant	.05	1.0	.997	.155	.148	.96
Logistic	.05	1.0	.993	.157	.151	.96
Complete Case	.01	1.0	1.008	.212	.216	.93
Linear No Round	.01	1.0	1.002	.143	.159	.88
Linear Round	.01	1.0	1.011	.031	.148	.05
Discriminant	.01	1.0	1.008	.139	.146	.86
Logistic[g]	.01	1.0	--	--	--	--

[b]True value of the proportion for the dummy variable.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.
[g]Method failed, presumably due to quasi-complete separation.
[h]True value of the coefficient in the model generating the data.

For the variable *X*, however, which had no missing data, Table 4 shows that the imputation methods are markedly superior to complete case analysis in estimating its coefficient. The standard errors for complete case estimates are about one-third larger than those for the imputation methods. None of the methods shows any major biases, although linear imputation with rounding yields estimates that are about 5 percent too large when *p* is .50 or .20. In the former case, coverage is also a bit low. When *p* is .01, the imputation methods deteriorate in coverage, especially linear imputation with rounding, which had coverage of only 5 percent. That's probably because the standard error estimates are way too low in this case.

MAR RESULTS

Table 5 gives results for estimating the coefficient of the dummy variable, when data are missing at random but not completely at random. Specifically, the probability of missingness on *D*

depends on the value of the covariate *X* (but not on *D* or *Y*). When the goal was to estimate a proportion in the MAR condition (Table 2), complete case analysis showed substantial bias. Here, however, the method produces approximately unbiased estimates, except in the extreme condition of p=.01. That's not surprising because it has been proven that complete case analysis produces unbiased estimates of regression coefficients so long as the probability of missingness does not depend on the response variable. As in Table 3, complete case analysis yields standard errors that are no bigger than those from the imputation methods.

As before, the worst of the five methods is linear imputation with rounding. It shows appreciable bias even when p is .2 or greater, and the coverage is terrible when p is .05 or lower. The other four methods do fine until p=.01 when they show both downward bias (around 25 percent too low) and coverage that is only around 70 percent.

Table 5. Estimates of Coefficients of Dummy Variable When Data Are MAR.

Method	Prop.[b]	Param.[h]	Estimate[c]	Avg. SE[d]	Stan. Dev.[e]	Coverage[f]
Complete Case	.50	1.0	.987	.493	.496	.94
Linear No Round	.50	1.0	.980	.494	.513	.95
Linear Round	.50	1.0	.809	.452	.421	.94
Discriminant	.50	1.0	.974	.492	.526	.94
Logistic	.50	1.0	.966	.501	.516	.95
Complete Case	.20	1.0	1.011	.629	.657	.95
Linear No Round	.20	1.0	1.018	.641	.672	.93
Linear Round	.20	1.0	.725	.546	.491	.95
Discriminant	.20	1.0	.981	.625	.678	.93
Logistic	.20	1.0	1.004	.611	.675	.92
Complete Case	.05	1.0	1.017	1.586	1.670	.94
Linear No Round	.05	1.0	.994	1.607	1.714	.93
Linear Round	.05	1.0	.908	2.416	1.611	.46
Discriminant	.05	1.0	.945	1.379	1.461	.93
Logistic[g]	.05	1.0	--	--	--	--
Complete Case	.01	1.0	.750	2.444	2.125	.72
Linear No Round	.01	1.0	.768	1.856	2.178	.71
Linear Round	.01	1.0	.721	4.614	2.088	.014
Discriminant	.01	1.0	.681	1.695	1.940	.71
Logistic[g]	.01	1.0	--	--	--	--

[b]True value of the proportion for the dummy variable.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.
[g]Method failed, presumably due to quasi-complete separation.
[h]True value of the coefficient in the model generating the data.

In Table 6, we see results for the coefficient of the covariate in the MAR condition. Here the message is pretty much the same as in Table 4 for the MCAR condition. There is little bias for any of the methods, although linear imputation with rounding has the most. The rounding method also has terrible coverage when p=.05 or .01, primarily because of poor estimation of the standard errors. When p=.01, complete case analysis still yields pretty good coverage, but coverage for the imputation methods deteriorates (especially, as noted, for imputation with rounding). Nevertheless, the imputation methods have substantially smaller standard errors than the complete case estimates.

Table 6. Estimates of Coefficients of Covariate When Data Are MAR.

Method	Prop.[b]	Param..[h]	Estimate[c]	Avg. SE[d]	Stan. Dev.[e]	Coverage[f]
Complete Case	.50	1.0	1.011	.258	.253	.95
Linear No Round	.50	1.0	1.006	.178	.177	.95
Linear Round	.50	1.0	1.068	.163	.156	.92
Discriminant	.50	1.0	1.008	.177	.176	.94
Logistic	.50	1.0	1.007	.178	.174	.94
Complete Case	.20	1.0	1.007	.240	.242	.95
Linear No Round	.20	1.0	1.040	.162	.165	.94
Linear Round	.20	1.0	1.085	.153	.151	.92
Discriminant	.20	1.0	1.011	.175	.180	.95
Logistic	.20	1.0	1.005	.175	.183	.95
Complete Case	.05	1.0	1.017	.366	.341	.96
Linear No Round	.05	1.0	1.080	.225	.237	.92
Linear Round	.05	1.0	1.118	.126	.213	.44
Discriminant	.05	1.0	1.040	.258	.254	.95
Logistic[g]	.05	1.0	--	--	--	--
Complete Case	.01	1.0	1.003	.390	.402	.94
Linear No Round	.01	1.0	1.012	.243	.309	.71
Linear Round	.01	1.0	1.031	.073	.268	.01
Discriminant	.01	1.0	1.015	.223	.271	.70
Logistic[g]	.01	1.0	--	--	--	--

[b]True value of the proportion for the dummy variable.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.
[g]Method failed, presumably due to quasi-complete separation.
[h]True value of the coefficient in the model generating the data.

CONCLUSION

Based on these results, I propose the following conclusions:

1. Linear imputation with rounding should never be used. It's usually inferior and never superior to linear imputation without rounding, which is computationally simpler.

2. For estimating proportions, the principal benefit from imputation is reduction in bias when data are MAR but not MCAR. Imputation methods have standard errors that are no smaller than those for complete case analysis. In the MCAR condition, where bias is not an issue, there is no particular benefit to imputation.

3. For estimating proportions in the MAR condition, linear imputation without rounding is inferior to the logistic and discriminant methods, but may be acceptable if the proportion to be estimated is above .20.

4. For estimating the coefficient of a dummy variable with missing data, all the methods except for linear imputation with rounding are about equally good.

5. For estimating the coefficient of the covariate that has no missing data, linear imputation with rounding is about as good as the logistic or discriminant methods. Complete case analysis has better coverage than the other methods when *p* is near zero, but it always has substantially higher standard errors.

Overall, linear imputation without rounding comes off pretty well, especially for estimating regression coefficients. Discriminant and logistic methods would be preferable, but they are only available when the missing data pattern is monotonic, a situation that rarely occurs when estimating regression models with many variables.

Like any simulation study, this one has its limitations. Both intuition and previous analytic work suggested that linear imputation methods may work well when dummy variables have a mean around .50 but may deterioriate markedly when the mean approaches 1 or 0. For

that reason, I chose the mean as the key parameter to vary across different conditions. But I did not look at the effect of variation in the other parameters, which were chosen to be fairly typical of social science data. Clearly, there is a need for further investigation of these parameters, including the regression coefficients β_1 and β_2, the correlation between *X* and *D*, the error variance σ^2 in the regression equation, and the coefficient α_2 in the logistic regression predicting missingness in the MAR case. With regard to that logistic regression, it would also be of interest to see what happens when missingness on *D* depends on *Y.* All the simulations assumed that all the continuous variables were normally distributed, and the consequences of other distributions should be investigated. Finally, I did not investigate the effects of varying sample size or the fraction of missing data.

REFERENCES

Allison, Paul D. (2001) *Missing Data*. Thousand Oaks, CA: Sage Publications.

Horton N.J., Lipsitz S.R., and Parzen, M. (2003) A potential for bias when rounding in multiple imputation. *American Statistician* 57: 229-232.

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

APPENDIX

SAS code for producing imputations and applying the missing data techniques:

```
/*Proportions, single auxiliary covariate, MCAR*/
%let cut=.50;
data dumsim;
cut=&cut;
do sample = 1 to 500;
do i=1 to 500;
d=ranuni(0)<cut;
x=-1+1*d+1*rannor(0);
miss=ranuni(0)<.5;
if miss=1 then dmiss=.; else dmiss=d;
output;
end;
end;
run;
proc corr data=dumsim; var d x miss;run;
proc mi data=dumsim out=outdum noprint;
var x dmiss;
monotone regression(dmiss=x);
by sample;
run;
/*The following macros do the analysis. Macros may be found below */
%complete
%analyze
%round
%discrim
%logistic

/*Proportions, single auxiliary covariate, MAR*/
%let cut=.01;
data dumsim;
misslope=2;
missint=2;
cut=&cut;
do sample = 1 to 500;
do i=1 to 500;
d=ranuni(0)<cut;
x=-1+1*d+.3*rannor(0);
p=1/(1+exp(-missint-misslope*x));
miss=ranuni(0)<p;if miss=1 then dmiss=.; else dmiss=d;
output;
end;
end;
run;
proc corr data=dumsim; var d x miss;run;
proc mi data=dumsim out=outdum noprint;
var x dmiss;
monotone regression(dmiss=x);
by sample;
run;
%complete
%noround
%round
%logistic
%discrim


%macro complete;
proc means data=dumsim nway noprint;
var dmiss;
class sample;
output out=a mean=mean lclm=lclm uclm=uclm stderr=se;
run;
data b;
set a;
coverage=lclm<&cut<uclm;
```

```
run;
proc means data=b;
var mean se coverage;
run;
%mend complete;

%macro noround;
%do i=1 %to 500;
proc reg data=outdum outest=a covout noprint;
where sample=&i;
model dmiss=;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  intercept;
ods output ParameterEstimates=parms&i;
run;
%end;
data parms;
set %do j=1 %to 500; parms&j %end; ;
coverage=lclmean<&cut<uclmean;
run;
ods listing;
proc means data=parms;
var Estimate stderr coverage;
run;
%mend noround;

%macro round;
data outround;
set outdum;
if dmiss>.5 then dmiss=1; else dmiss=0;
run;
%do i=1 %to 500;
proc reg data=outround outest=a covout noprint;
where sample=&i;
model dmiss=;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  intercept;
ods output ParameterEstimates=parms&i;
run;
%end;
data parms;
set %do j=1 %to 500; parms&j %end; ;
coverage=lclmean<&cut<uclmean;
run;
ods listing;
proc means data=parms;
var Estimate stderr coverage;
run;
%mend round;

%macro logistic;
%do i=1 %to 500;
proc mi data=dumsim out=outlog noprint ;
where sample=&i;
class dmiss;
var x dmiss;
monotone logistic(dmiss=x) ;
run;
proc reg data=outlog outest=a covout noprint;
model dmiss=;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  intercept;
```

```
ods output ParameterEstimates=parms&i;
run;
%end;
data parms;
set %do j=1 %to 500; parms&j %end; ;
coverage=lclmean<&cut<uclmean;
run;
ods listing;
proc means data=parms;
var Estimate stderr coverage;
run;
%mend logistic;

%macro discrim;
%do i=1 %to 500;
proc mi data=dumsim out=outlog noprint ;
where sample=&i;
class dmiss;
var x dmiss;
monotone discrim(dmiss=x) ;
run;
proc reg data=outlog outest=a covout noprint;
model dmiss=;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  intercept;
ods output ParameterEstimates=parms&i;
run;
%end;
data parms;
set %do j=1 %to 500; parms&j %end; ;
coverage=lclmean<&cut<uclmean;
run;
ods listing;
proc means data=parms;
var Estimate stderr coverage;
run;
%mend discrim;


/* Regression with dummy predictor, MAR*/
%let cut=.5;
%let b=1;
%let c=1;
data dumreg;
b=&b;
c=&c;
missint=1;
misslope=1;
sig=3;
cut=&cut;
cutmiss=.5;
do sample = 1 to 500;
do i=1 to 500;
d=ranuni(0)<cut;
x=-1+1*d+1*rannor(0);
y=b*d+c*x+sig*rannor(0);
p=1/(1+exp(-missint-misslope*x));
miss=ranuni(0)<p;if miss=1 then dmiss=.; else dmiss=d;
output;
end;
end;
run;
proc reg; model y= x d; run;
ODS LISTING;
proc corr data=dumreg;var d x y miss; run;
proc mi data=dumreg out=outreg noprint;
var x y dmiss;
monotone regression(dmiss=x y);
```

```
by sample;
run;
/*The following macros do the analysis. Macros may be found below*/
%completereg
%noroundreg
%roundreg
%discrimreg
%logreg

%macro completereg;
ods listing close;
proc reg data=dumreg;
model y=dmiss x / clb;
ods output ParameterEstimates=parms;
by sample;
run;
ods listing;
data dmiss;
set parms;
where variable='dmiss';
coverage=lowercl<&b<uppercl;
run;
proc means data=dmiss;
var Estimate StdErr coverage;
run;
data x;
set parms;
where variable='x';
coverage=lowercl<&c<uppercl;
run;
proc means data=x;
var Estimate StdErr coverage;
run;
%mend completereg;

%macro noroundreg;
%do i=1 %to 500;
proc reg data=outreg outest=a covout noprint;
where sample=&i;
model y=dmiss x;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  dmiss x;
ods output ParameterEstimates=parms&i;
run;
%end;
data parmsb;
set %do j=1 %to 500; parms&j %end; ;
where parm='dmiss';
coverage=lclmean<&b<uclmean;
run;
ods listing;
proc means data=parmsb;
var Estimate stderr coverage;
run;
data parmsc;
set %do j=1 %to 500; parms&j %end; ;
where parm='x';
coverage=lclmean<&c<uclmean;
run;
ods listing;
proc means data=parmsc;
var Estimate stderr coverage;
run;
%mend noroundreg;

%macro roundreg;
data outround;
set outreg;
```

```
if dmiss>.5 then dmiss=1; else dmiss=0;
run;
%do i=1 %to 500;
proc reg data=outround outest=a covout noprint;
where sample=&i;
model y=dmiss x;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  dmiss x;
ods output ParameterEstimates=parms&i;
run;
%end;
data parmsb;
set %do j=1 %to 500; parms&j %end; ;
where parm='dmiss';
coverage=lclmean<&b<uclmean;
run;
ods listing;
proc means data=parmsb;
var Estimate stderr coverage;
run;
data parmsc;
set %do j=1 %to 500; parms&j %end; ;
where parm='x';
coverage=lclmean<&c<uclmean;
run;
ods listing;
proc means data=parmsc;
var Estimate stderr coverage;
run;
%mend roundreg;

%macro logreg;
%do i=1 %to 500;
proc mi data=dumreg out=outlog noprint ;
where sample=&i;
class dmiss;
var x y dmiss;
monotone logistic (dmiss=x y) ;
run;
proc reg data=outlog outest=a covout noprint;
model y=dmiss x;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  dmiss x;
ods output ParameterEstimates=parms&i;
run;
%end;
data parmsb;
set %do j=1 %to 500; parms&j %end; ;
where parm='dmiss';
coverage=lclmean<&b<uclmean;
run;
ods listing;
proc means data=parmsb;
var Estimate stderr coverage;
run;
data parmsc;
set %do j=1 %to 500; parms&j %end; ;
where parm='x';
coverage=lclmean<&c<uclmean;
run;
ods listing;
proc means data=parmsc;
var Estimate stderr coverage;
run;
%mend logreg;
```

```
%macro discrimreg;
%do i=1 %to 500;
proc mi data=dumreg out=outlog noprint ;
where sample=&i;
class dmiss;
var x y dmiss;
monotone discrim (dmiss=x y) ;
run;
proc reg data=outlog outest=a covout noprint;
model y=dmiss x;
by _imputation_;
run;
ods listing close;
proc mianalyze data=a;
var  dmiss x;
ods output ParameterEstimates=parms&i;
run;
%end;
data parmsb;
set %do j=1 %to 500; parms&j %end; ;
where parm='dmiss';
coverage=lclmean<&b<uclmean;
run;
ods listing;
proc means data=parmsb;
var Estimate stderr coverage;
run;
data parmsc;
set %do j=1 %to 500; parms&j %end; ;
where parm='x';
coverage=lclmean<&c<uclmean;
run;
ods listing;
proc means data=parmsc;
var Estimate stderr coverage;
run;
%mend discrimreg;
```