Multiple Semi-Parametric Imputation

David Judkins[1], Andrea Piesse[1], and Tom Krenzke[1]
[1]Westat, 1650 Research Boulevard., Rockville, Maryland 20850

Abstract
In 2007, Judkins, Krenzke, Piesse, Fan, and Haung reported on the performance of a new semi-parametric imputation algorithm designed to impute entire questionnaires with minimal human supervision while preserving important first- and second-order distributional properties. In this paper, we report on procedures for post-imputation variance estimation to be used in conjunction with the semi-parametric imputation algorithm.

Keywords: Post-imputation variance estimation, Hot-deck imputation

1. Review of Past Research and Research Schema

In Judkins et al. (2007), it was demonstrated that semi-parametric multivariate imputation procedures are competitive with Bayesian procedures designed for the same purpose. Competitiveness was judged on the basis of three criteria: preservation of unusual data features, preservation of multivariate normal data features, and coverage of confidence intervals. The tested semi-parametric procedure clearly outperformed the tested Bayesian procedure (IVEware: Raghunathan, Solenberger, and Hoewyk, 2002) with respect to the first criterion. The two alternatives were tied with respect to the second criterion. The Bayesian procedure was the winner on the third criterion, not necessarily because it generated confidence intervals with superior coverage, but because we were unable to provide information about the performance of multiple semi-parametric imputation as a technique for post-imputation variance estimation. The current paper is meant to fill that gap.

During the time between our 2007 and current year's presentations, related research was published. Siddique and Belin (2008), and Little et al. (2008) both explore the use of multiple semi-parametric imputation as a variance estimation technique with favorable results. The current research complements this work by applying the idea of replicated imputations to different imputation algorithms. Also, the replication scheme we tested is simpler than theirs: we did not include a bootstrap step.

The schema for our research was to apply the 2007 algorithm to simulated datasets multiple times, apply Rubin's formula for multiple-imputation based post-imputation variance estimation (Rubin, 1996), use the estimated variances to construct confidence intervals, and then evaluate the results in terms of biases in variance estimates, lengths of confidence intervals, and coverage of confidence intervals.

We make no assertion that the semi-parametric procedure we tested is a proper procedure in the sense of Rubin (1996) or Schafer (1997). We simply hoped that with appropriate settings of software tuning parameters we could achieve reasonable coverage using confidence intervals constructed under the theory appropriate for proper imputation. We considered adding a bootstrap step as was done by Siddique and Belin but ran out of time.

In Section 2, we review the semi-parametric algorithm and the history of its development. We also indicate how multiple imputations were used to estimate post-imputation variances. In Section 3, we describe the simulation framework. In Section 4, we lay out the evaluative criteria we applied. Results are given in Section 5. Section 6 contains some closing remarks on implications for applications and further research.

2. A Semi-Parametric Algorithm

The needs of a data publisher are typically different than those of a secondary analyst whose research may involve a limited number of variables and who may therefore be willing to invest substantial time and energy on maximum likelihood or optimal Bayesian estimation of model parameters. Typically, the publisher must impute all missing data at low expense to support a variety of unforeseeable analyses. In the attempt to meet these goals, we have developed software called AutoImpute which blends ideas from Gibbs sampling, data mining, predictive mean matching, and hot-deck imputation. AutoImpute was designed to impute entire complex questionnaires with a single job submission while preserving questionnaire skip patterns (no pregnant men), other strong bivariate patterns (only a few Yiddish-speaking Eskimos), and essential features of all marginal distributions (first- and second-order moments, ranges, and discontinuities).

To answer a question that was raised by Xiao Li Meng during our presentation, the conceptual model underpinning the automated semi-parametric approach embodied in this software is along the following lines. We believe that there are interesting discoveries to be made from the database at hand, but we do not have the resources or responsibility for making those discoveries. Instead, our job is to simplify the discovery process for users of the published data. We want to rectangularize (i.e., complete) the data while disturbing them as little as possible. The value of semi-parametric multivariate imputation was demonstrated strongly for at least some settings in Marker, Judkins, and Winglee (2001).

Use of the semi-parametric approach has drastically reduced the time and cost to conduct item imputation (Piesse, Judkins, and Fan, 2005). Under simulated strongly informative missing data mechanisms, the analysis of data imputed using the semi-parametric algorithm resulted in smaller bias and variance on marginal means and smaller bias in correlations than the analysis of complete case data alone (Krenzke, Judkins, and Fan, 2005). As noted above, further favorable properties of the algorithm were established by Judkins et al. (2007) in terms of preserving odd marginal and conditional features as well as relationships among unordered multinomial variables.

With regard to the history and origins of the software, we three are the primary designers; the original programming was by Zizhong Fan in 2004, the current head programmer is Wen-Chau Haung; the underlying hot-deck engine was programmed by Katie Hubbell in the early 1990s; the basic idea for cyclical application of hot-decks came to David Judkins in 1993 (Judkins, Hubbell, and England, 1993); he was inspired by Joe Schafer's work in data augmentation (Schafer, 1993); who in turn, was inspired by Arthur Kennickell's 1991 paper. Judkins coined the phrase "cyclic n-partition hot-decks" in 1997 (Judkins, 1997).

The algorithm for AutoImpute is described in detail in Judkins et al. (2007). In brief, the algorithm primarily uses hot-decks with partitions defined by skip patterns and predicted values of target variables as in predictive mean matching (a phrase coined by Little, 1988, for a method first published in Rubin, 1986). Prediction models are linear with only main effects. These models are built with a forward search through all variables that are applicable to the set of cases to which the target variable applies. Previously imputed values of other variables are used in the models. Predicted values can be coarsened at the option of the user in such a way as to divide the total eligible sample into a user-specified number of equal-sized groups.

For unbounded continuous variables, the algorithm hot-decks empirical residuals to the model-based predictions in order to improve performance at the tails of the distributions of predictor variables.

For unordered multinomial variables, independent models are constructed for all the levels of the target variable. Cases are then clustered based on the multivariate prediction vectors, and the clusters are used as partitions to hot-deck the target variable.

A simple hot-deck is used to initialize the process so that all variables can be used to predict all other variables defined on comparable sets of cases, regardless of the complexity of the overall missing data pattern. Each variable is then re-imputed in turn. After all the variables have been re-imputed once, we say that one sweep through the database has been completed. We assess convergence through the R-squared statistics of the prediction models for all ordered target variables. When none of the R-squared measures is still increasing, we say that the algorithm has converged. We also allow a limit to be placed on the number of sweeps. For the current research, the limit was ten sweeps.

To estimate post-imputation variances, multiple imputations were obtained by repeating this entire process. Each of the multiple imputations was obtained by re-starting the software with a different pseudo-random number generator seed. Since the hot-deck engine at the core of the algorithm randomly matches donors and beggars within cells of the partition, this results in a seemingly reasonable dispersion of results unless the cells of the partition are defined too narrowly.

The primary aim of the current research was to determine whether that dispersion could be used to estimate post-imputation variances using Rubin's standard formula (Rubin, 1996) even though the imputation process is not proper. A secondary aim was to see whether the number of coarsening groups was related to the quality of the confidence intervals produced by the approach. We hypothesized that stronger coarsening might result in better actual coverage of nominal confidence intervals at the same time that it might widen those intervals.

3. Simulation Framework

We returned to three of the four test scenarios in Judkins et al. (2007). One was "strange pop," another was "checkerboard," and the third was a plain vanilla "bivariate normal."

In strange pop there are two variables with range restrictions and/or discontinuities, as well as a very unusual dependency. It is easiest to describe the pair by construction and by graphs. Let $X \sim \mathrm{U}(-1,1)$. Let $e_Y \sim \mathrm{N}(0,1/25)$. Let

$$Y = \begin{cases} e_Y + \max\left\{-6, \min\left[6, 2\Phi^{-1}\left(\frac{X+1}{2}\right)\right]\right\} & \text{if } |X| > 0.5; \\ e_Y & \text{otherwise.} \end{cases}$$

Figures 1, 2, and 3 show the marginal distributions of *X* and *Y*, and the conditional distribution of *Y* given *X*. The following are some of the essential features of this population: *X* is bounded by cliffs; *Y* has a large concentration near zero, with a substantial gap either side of zero in which values are highly unlikely; the conditional distribution of *Y* near the center range of *X* is flat; and the conditional distribution of *Y* near the extremes of *X* is exponential. Clearly, we would be surprised to find a pair of variables like this in the survey setting, but this scenario was designed to demonstrate the ability of our software to handle the unexpected without human intervention.

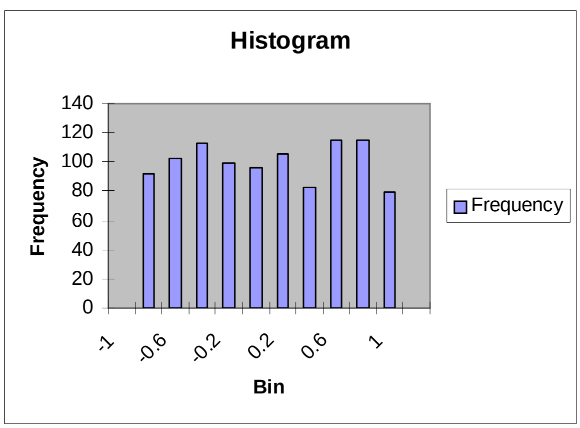

Figure 1: Marginal distribution of *X* in strange pop

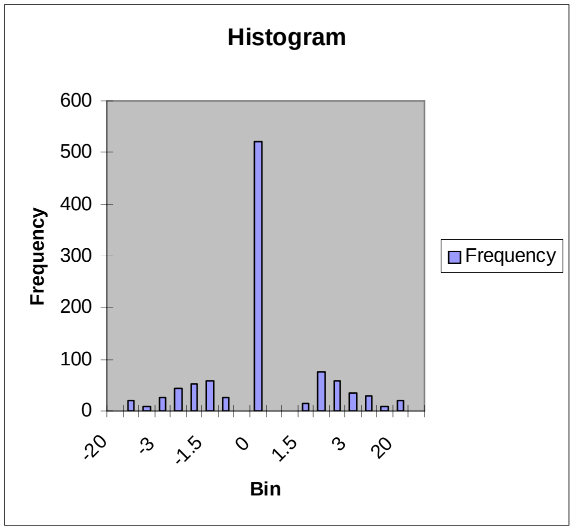

Figure 2: Marginal distribution of *Y* in strange pop

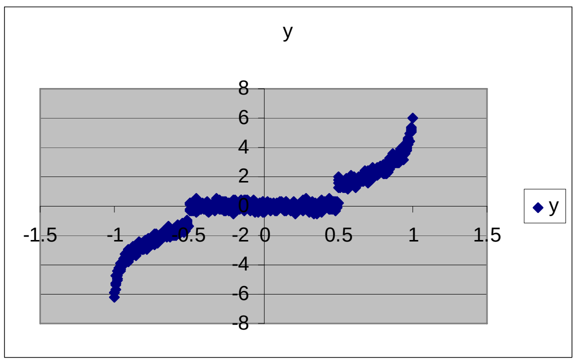

Figure 3: Conditional distribution of *Y* given *X* in strange pop

For this scenario, we chose the following statistics of interest:

$$\mathrm{E}(X) = 0\,,$$
$$\mathrm{E}(Y) = 0\,,$$
$$\mathrm{E}(Y \mid X \in [-0.5, 0.5]) = 0\,,$$
$$\mathrm{E}(Y \mid X \in (0.5, 0.75]) = 1.790\,, \text{ and}$$
$$\mathrm{E}(Y \mid X > 0.75) = 3.287\,.$$

Scenario #2 involves one binary variable and two multinomial variables where the cell frequencies follow a reversing checkerboard pattern. The full joint distribution for *X*, *Y*, and *Z* is given in Table 1. The corresponding log-linear model involves both two- and three-way interactions.

Table 1: Three-Way Table with Checkerboard Pattern

X = 1	*Y = 1*	*Y = 2*	*Y = 3*
Z = 1	0.0067	0.0467	0.0067
Z = 2	0.0467	0.0067	0.0467
Z = 3	0.0067	0.0467	0.0067
Z = 4	0.0467	0.0067	0.0467

X = 2	*Y* = 1	*Y* = 2	*Y* = 3
Z = 1	0.0367	0.0767	0.0367
Z = 2	0.0767	0.0367	0.0767
Z = 3	0.0367	0.0767	0.0367
Z = 4	0.0767	0.0367	0.0767

For this scenario, the chosen statistics of interest were:

$$\Pr(X = 1) = 0.3100\,,$$
$$\Pr(Z = 1) = 0.2100\,,$$
$$\Pr(Y = 1) = 0.3333\,,$$
$$\Pr(Y = 2) = 0.3333\,,$$
$$\Pr(X = 1 \wedge Z = 1) = 0.0600\,,$$
$$\Pr(X = 1 \wedge Y = 1) = 0.1067\,,$$
$$\Pr(X = 2 \wedge Z = 1) = 0.1500\,,$$
$$\Pr(X = 2 \wedge Y = 1) = 0.2267\,,$$
$$\Pr(X = 1 \wedge Z = 1 \wedge Y = 1) = 0.0067\,,$$
$$\Pr(X = 1 \wedge Z = 1 \wedge Y = 2) = 0.0467\,,$$
$$\Pr(X = 2 \wedge Z = 1 \wedge Y = 1) = 0.0367\,, \text{ and}$$
$$\Pr(X = 2 \wedge Z = 1 \wedge Y = 2) = 0.0767\,.$$

Scenario 3 consists of a pair of bivariate normal variables:

$$X \sim N(60, 9)$$
$$Y \sim N(120 + X, 1)$$

Note that for the regression of Y on X, the R-squared statistic is 0.5. For this scenario, the statistics of interest were:

$$\mathrm{E}(X) = 60\,,$$
$$\mathrm{E}(Y) = 120\,, \text{ and}$$
$$\beta_{Y|X} = 1/3\,.$$

For all three test scenarios, we generated 500 data sets of 1,000 observations each. We tested two different levels of item nonresponse: 30 percent and 90 percent. The 90 percent level was chosen to simulate uses where data are missing by design. Missingness was then created for each item independently across observations. The missing data mechanism was completely at random (MCAR).

Five multiple imputations were drawn for each of the 500 replicated samples.

Variable-specific tuning parameters informed AutoImpute that: *X* and *Y* in strange pop were ordered but not continuous unbounded variables; *X*, Y, and *Z* in checkerboard were unordered multinomial variables; and *X* and *Y* in the plain vanilla scenario were continuous unbounded variables. For all three scenarios, we varied the number of coarsening groups. Other than that, the software was given no human guidance.

4. Evaluation Criteria

We evaluated the results in terms of:

- Bias of point estimate;
- Variance of point estimate;
- Bias of variance estimate;
- Half-width of nominal 95 percent confidence interval; and
- Coverage of nominal 95 percent confidence interval.

5. Results

There are five tables of results for each scenario, corresponding to the five evaluation criteria listed above. Variances and biases in variance estimates are multiplied by 10,000 for ease of display. Each table contains results for three different settings of the coarsening parameter and two different levels of item nonresponse.

5.1 Strange Pop

Considering strange pop first, we see in Table 2 that biases in point estimates are negligible for the marginal means and for the conditional mean of *Y* in the central region of *X*, for all levels of coarsening, if the item nonresponse rate is 30 percent. For the conditional means of Y in the mid-high and very high regions of X, bias is negligible only if 200 groups are used for coarsening of predicted values and if the item nonresponse rate is 30 percent. For 90 percent missing data, hardly any of the means are well estimated.

Considering next the variances for strange pop estimated parameters, we see from Table 3 that variances are much higher with 90 percent missing data than with 30 percent missing data, as would of course be expected. More interestingly, we note that less coarsening of predicted values appears to lead to somewhat lower variances.

We expected to generally underestimate variances since we failed to include a bootstrap step in the imputation process, but the data in Table 4 show that the opposite was frequently true. We have no good hypotheses for why these positive biases occurred: almost all are statistically significant, considering the variance in the variance estimates across the 500 replications.

Turning now to the half-widths of nominal 95 percent confidence intervals in Table 5, we note that they are wider for 90 percent missingness than for 30 percent missingness. Also as expected, there is some shortening of half-widths with less coarsening.

In conclusion for strange pop, we see from Table 6 that actual coverage of nominal 95 percent confidence intervals is generally very poor for 90 percent missing data, while coverage results for 30 percent missing data are mixed. For the combination of 30 percent missingness and 200 coarsening groups, the coverage levels are almost perfect.

5.2 Checkerboard

Tables 7-11 contain parallel results for the checkerboard scenario. Since there were 12 statistics of interest, we averaged results to condense displays. We show averaged results for the four targeted marginal means, the four targeted two-way means, and the four targeted three-way means.

Although we tested and present performance statistics for all three coarsening levels, the number of coarsening groups generally had little effect on the results. This is because the coarsening applies only to predictions of ordinal and binary variables, and so only affected the imputation of *X* in this scenario.

Biases in point estimates are generally negligible, although they are higher for marginal means than for higher-order means. Variances on estimated parameters and half-widths of confidence intervals rise with the item missing rate, as expected. Biases in variance estimates are modest with 30 percent missing data and substantial with 90 percent missing data. Biases are generally negative which is not surprising given the lack of a bootstrap step. Actual coverage of nominal 95-percent confidence intervals is unacceptable for 90 percent missing data but perhaps tolerable for 30 percent missing data, particularly for marginal means.

5.3 Plain Vanilla Bivariate Normal

Tables 12-15 contain parallel results for the third scenario. Again we tested all three levels of coarsening, but since the coarsening for this scenario affected only the hot-decking of empirical residuals and because the true errors are homoscedastic, the number of coarsening groups had very little impact on the results.

For this very simple population, the biases in the marginal means are negligible for both levels of missingness, but the regression coefficient is strongly attenuated given 90 percent missing data. Variances and half-widths of confidence intervals increase with the missing data rate. Biases in variance estimates are modest for 30 percent missing data and very large for 90 percent missing data, but as with strange pop, the biases in the variance estimates are not always negative.

Actual coverage of confidence intervals on marginal means is good for 30 percent missing data. Coverage for confidence intervals on the regression coefficient is not as good given the same level of missingness, but perhaps acceptable. Coverage is poor for 90 percent missing data.

6. Discussion and Further Research

We are moderately encouraged by these results. As long as the missing data rate is not above 30 percent, actual coverage of nominal 95 percent confidence intervals was generally good for marginal means and not too bad for more complex statistics. On the other hand, coverage was quite poor under some circumstances. This leads us to think that users should consider verifying the appropriateness of their tuning parameters through realistic simulation studies.

Since many of the biases in variance estimates were positive, poor coverage of confidence intervals for 90 percent missing data might be due more to biases in the point estimates than to problems with the variance estimates. This suggests that increasing the maximum allowed number of sweeps may be more important than adding a bootstrap step or increasing the number of multiple imputations.

Perhaps more than 10 sweeps were needed for the 90 percent item nonresponse rate. We did not monitor how many of the imputation runs ended in convergence before reaching the limit of 10 sweeps. However, the fact that biases in point estimates of regression coefficients and conditional means were much smaller with 30 percent missing data than with 90 percent missing data, suggests that the algorithm might eventually recover from the initial naïve hot-deck if allowed to run long enough. Siddique and Belin (2008) make similar observations.

Unfortunately, this sort of simulation study was straining the limits of the computer resources available to us. To allow 500 replications of 5 multiple imputations with 10 sweeps per imputation often caused system crashes. We will need to either improve the efficiency of the software or get faster computers to resolve the issue.

Nonetheless, we are reasonably comfortable in recommending the use of multiple semi-parametric imputations with our system provided that item nonresponse rates are not too high.

References

Judkins, D.R. (1997). Imputing for Swiss cheese patterns of missing data. *Proceedings of Statistics Canada Symposium* 97, New Directions in Surveys and Censuses, 143-148.

Judkins, D.R., Hubbell, K.A., and England, A.M. (1993). The imputation of compositional data. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 458-462.

Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3211-3218.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation (with discussion pp 21-23). *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 1-10.

Krenzke, T., Judkins, D., and Fan, Z. (2005). Vector imputation at Westat. Presented at Statistical Society of Canada Meetings, Saskatoon, Saskatchewan.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287-296.

Little, R.J., Yosef, M., Cain, K.C., Nan, B., and Harlow, S.D. (2008). A hot-deck multiple imputation procedure for gaps in longitudinal data on recurrent events. *Statistics in Medicine,* 27(1), 103-120.

Marker, D.A., Judkins, D.R., and Winglee, M. (2001). Large-scale imputation for complex surveys, in *Survey Nonresponse*, Eds. R. M. Groves, D. A. Dillman, E. L. Eltinge, and R. J. A. Little. New York: Wiley.

Piesse, A., Judkins, D., and Fan, Z. (2005). Item imputation made easy. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3476-3479.

Raghunathan, T.E., Solenberger, P.W., and Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software Users Guide*. Ann Arbor: Institute for Social Research, University of Michigan.

Rubin, D.B. (1986). Statistical matching and file concatenation with adjusted weights and multiple imputation. *Journal of Business & Economic Statistics*, 4, 87-94.

Rubin, D.B. (1996). Multiple imputation after 19+ years. *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1993). Multiple imputation of missing data in NHANES III. Annual Research Conference, U.S. Bureau of the Census, Arlington, VA.

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

Siddique, J. and Belin, T.R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, 7(1), 83-102.

Table 2: Biases in point estimates for parameters of interest in Scenario 1, Strange Pop

Parameter	*True value*	*Bias in point estimate*					
Number of coarsening groups		*5*		*10*		*200*	
Missing rate		*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
$\mathrm{E}(X)$	0	0.000	-0.001	0.000	-0.002	0.000	-0.004
$\mathrm{E}(Y)$	0	-0.002	-0.023	-0.001	-0.020	-0.002	-0.030
$\mathrm{E}(Y \mid X \in [-0.5, 0.5])$	0	-0.001	-0.017	-0.001	-0.006	-0.001	-0.016
$\mathrm{E}(Y \mid X \in (0.5, 0.75])$	1.790	0.017	-0.809	-0.110	-0.935	-0.012	-1.296
$\mathrm{E}(Y \mid X > 0.75)$	3.287	-0.267	-1.905	-0.041	-1.687	-0.014	-2.488

Table 3: Variances on estimated parameters of interest in Scenario 1, Strange Pop

Parameter	*Variance of point estimate (x10,000)*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
$\mathrm{E}(X)$	3.8	27.7	3.7	27.9	3.7	30.4
$\mathrm{E}(Y)$	42.0	352.1	41.3	349.0	39.2	390.2
$\mathrm{E}(Y \mid X \in [-0.5, 0.5])$	8.2	451.7	3.6	517.2	1.8	471.2
$\mathrm{E}(Y \mid X \in (0.5, 0.75])$	111.3	3,586.7	31.7	2,803.3	18.0	1,616.1
$\mathrm{E}(Y \mid X > 0.75)$	87.9	6,387.8	76.7	9,459.1	76.1	2,811.7

Table 4: Biases in variance estimates on estimated parameters of interest in Scenario 1, Strange Pop

Parameter	*Bias in variance estimate (x10,000)*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
$E(X)$	0.0	-16.8	0.0	-15.1	0.1	-18.8
$E(Y)$	0.4	-231.1	0.2	-206.5	2.5	-259.1
$E(Y \mid X \in [-0.5, 0.5])$	0.5	225.8	2.1	397.9	1.8	234.9
$E(Y \mid X \in (0.5, 0.75])$	37.7	955.6	26.7	2,704.9	11.5	3,190.3
$E(Y \mid X > 0.75)$	17.5	799.9	7.4	786.2	30.9	3,043.1

Table 5: Half-widths of nominal 95% confidence intervalson parameters of interest in Scenario 1, Strange Pop

Parameter	*Half-width of nominal 95% confidence interval*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
$E(X)$	0.038	0.062	0.038	0.067	0.038	0.064
$E(Y)$	0.127	0.206	0.126	0.222	0.127	0.214
$E(Y \mid X \in [-0.5, 0.5])$	0.057	0.469	0.046	0.539	0.033	0.487
$E(Y \mid X \in (0.5, 0.75])$	0.236	1.210	0.148	1.328	0.092	1.256
$E(Y \mid X > 0.75)$	0.199	1.512	0.178	1.780	0.193	1.386

Table 6: Coverage of nominal 95% confidence intervals on parameters of interest in Scenario 1, Strange Pop

Parameter	*Coverage (%) of nominal 95% confidence interval*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
$E(X)$	94.4	72.4	94.8	75.4	94.6	72.6
$E(Y)$	95.0	70.4	95.0	73.0	95.6	69.0
$E(Y \mid X \in [-0.5, 0.5])$	94.6	93.2	97.6	94.4	96.4	92.2
$E(Y \mid X \in (0.5, 0.75])$	97.6	72.2	74.8	73.6	95.4	45.2
$E(Y \mid X > 0.75)$	23.0	36.6	91.8	63.2	95.4	8.8

Table 7: Average Biases in Point Estimates for Parameters of Interest in Scenario 2, Checkerboard

Parameter	*Bias in point estimate*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
1-way means	0.002	0.004	0.003	0.004	0.003	0.004
2-way means	0.001	0.001	0.001	0.001	0.000	0.001
3-way means	0.000	0.000	0.000	0.000	0.000	0.000

Table 8: Average variances on estimated parameters of interest in Scenario 2, Checkerboard

Parameter	*Variance of point estimate (x10,000)*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
1-way means	2.9	22.2	2.9	22.2	2.9	22.1
2-way means	1.6	20.4	1.6	20.5	1.6	20.7
3-way means	0.6	9.6	0.6	9.6	0.6	9.6

Table 9: Average biases in variance estimates on estimated parameters of interest in Scenario 2, Checkerboard

Parameter	*Bias in variance estimate (x10,000)*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
1-way means	-0.4	-15.2	-0.4	-15.0	-0.4	-14.8
2-way means	0.0	0.6	-0.1	1.0	-0.1	0.7
3-way means	0.0	1.7	0.0	2.1	0.0	2.0

Table 10: Average half-widths of nominal 95% confidence intervals on parameters of interest in Scenario 2, Checkerboard

Parameter	*Half-width of nominal 95% confidence interval*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
1-way means	0.031	0.050	0.031	0.050	0.031	0.050
2-way means	0.024	0.082	0.024	0.083	0.024	0.083
3-way means	0.014	0.055	0.014	0.056	0.014	0.055

Table 11: Average coverage of nominal 95% confidence intervalson parameters of interest in Scenario 2, Checkerboard

Parameter	*Coverage (%) of nominal 95% confidence interval*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
1-way means	92.5	70.4	92.6	69.8	92.7	70.3
2-way means	93.8	89.3	93.5	89.3	93.3	89.7
3-way means	90.3	78.2	89.9	78.3	90.1	78.0

Table 12: Biases in point estimates for parameters of interest in Scenario 3, Plain Vanilla Bivariate Normal

Parameter	*True value*	*Bias in point estimate*					
Number of coarsening groups		*5*		*10*		*200*	
Missing rate		*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
$\mathrm{E}(X)$	0	-0.002	-0.021	-0.002	-0.018	-0.001	-0.013
$\mathrm{E}(Y)$	0	-0.003	-0.008	-0.003	-0.010	-0.002	-0.007
$\beta_{Y\|X}$	1/3	0.000	-0.142	0.000	-0.156	0.000	-0.240

Table 13: Variances on estimated parameters of interest in Scenario 3, Plain Vanilla Bivariate Normal

Parameter	*Variance of point estimate (x10,000)*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
E(X)	116.6	902.6	115.8	892.2	120.9	925.1
E(Y)	25.3	214.6	24.6	210.2	24.8	219.1
$\beta_{Y\|X}$	1.9	125.7	1.9	113.3	2.1	31.3

Table 14: Biases in variance estimates on estimated parameters of interest in Scenario 3, Plain Vanilla Bivariate Normal

Parameter	*Bias in variance estimate (x10,000)*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
E(X)	-11.7	-631.0	-11.0	-589.6	-12.7	-594.2
E(Y)	-1.8	-152.0	-1.3	-139.3	-0.9	-146.2
$\beta_{Y\|X}$	-0.3	62.2	-0.4	62.2	-0.4	46.9

Table 15: Half-widths of nominal 95% confidence intervals on parameters of interest in Scenario 3, Plain Vanilla Bivariate Normal

Parameter	*Half-width of nominal 95% confidence interval*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
E(X)	0.200	0.308	0.200	0.323	0.203	0.337
E(Y)	0.095	0.147	0.095	0.156	0.096	0.158
$\beta_{Y\|X}$	0.024	0.243	0.024	0.239	0.025	0.161

Table 16: Coverage of nominal 95% confidence intervals on parameters of interest in Scenario 3, Plain Vanilla Bivariate Normal

Parameter	*Coverage (%) of nominal 95% confidence interval*					
Number of coarsening groups	*5*		*10*		*200*	
Missing rate	*30%*	*90%*	*30%*	*90%*	*30%*	*90%*
E(X)	93.8	68.6	92.6	70.0	93.0	72.0
E(Y)	94.2	66.0	93.2	66.0	93.2	69.0
$\beta_{Y\|X}$	91.0	80.8	91.4	77.8	89.6	11.2