ROBUSTNESS OF MULTIPLE-IMPUTATION TECHNIQUES TO MODEL MISSPECIFICATION

Laura C. Lazzeroni, Nathaniel Schenker, and Jeremy M.G. Taylor, UCLA
Jeremy M.G. Taylor, UCLA Dept. of Biostatistics, Los Angeles, CA 90024-1772

KEY WORDS: Missing data, nonresponse, predictive mean matching, sample surveys.

1. Introduction

Multiple imputation (Rubin 1978, 1987) is a procedure for handling missing data that allows the data analyst to assess the uncertainty due to imputing the missing values. Several, say M, values are drawn to replace each missing observation, resulting in M completed data sets. Each completed data set is analyzed using standard data-analysis techniques, and the results are combined to yield one inference.

From the Bayesian perspective, multiple imputation is motivated as follows. Let Y_{obs} and Y_{mis} denote, respectively, the observed and missing values in a data set. Then under suitable conditions, the posterior density of a population quantity Q can be expressed as

$$(1) \quad \int g(Q|Y_{obs},Y_{mis})f(Y_{mis}|Y_{obs})dY_{mis},$$

where f is the posterior predictive density of the missing values and g is the complete-data posterior density of Q; that is, the posterior of Q is the average of the complete-data posterior over the predictive distribution of the missing data. Multiple imputation simulates approximate draws from f and thus allows the data analyst to approximate the averaging specified in (1).

Typically, the specification of f in (1) involves formulating both a model for the data and a model for the missing-data mechanism. Recent evaluations have shown that if appropriate models for the missing-data mechanism and for the data are used, then multiple imputation usually performs quite well (Herzog and Rubin 1983; Rubin and Schenker 1986, 1987; Raghunathan 1987; Rubin 1987).

This paper evaluates the robustness of four multiple-imputation procedures to misspecification of the model for the data. Fully parametric techniques are contrasted with less parametric schemes that replace missing values for incomplete cases with data from fully observed cases. A particular goal is to investigate whether the less parametric techniques are more robust. Results are given from a Monte Carlo study performed in the regression setting with two explanatory variables and a dependent variable subject to ignorable nonresponse. Model misspecification is represented by using a linear mean structure for the imputation procedure when the actual data structure is nonlinear. The imputation methods are compared with respect to the performance of point estimators and confidence intervals for the marginal distribution function of the dependent variable at several points.

2. Imputation Methods Studied

Suppose that n observations on a dependent variable Y and a p-dimensional explanatory variable X are sampled, but that only n_{obs} of the Y-values are observed.

The four imputation methods studied here are all initially based on the normal-theory linear regression model, which assumes that for observation i,

$$Y_i=X_i^T\beta+\epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2),$$

where $\epsilon_1,\ldots,\epsilon_n$ are independent, and β and σ are unknown parameters.

Under a specified method, one set of imputes for the missing Y-values is drawn in two stages, as described below. Multiple imputations are created by independently repeating the two stages M times.

The first stage is identical for all four methods. First, the linear regression model is fitted by least squares to the n_{obs} complete cases. Values of the parameters β and σ are then drawn from their posterior distribution under the Jeffreys prior (Box and Tiao 1973). Specifically, σ^2 is set equal to σ^{*2}=SSE/W, where SSE is the residual sum of squares from the least squares fit and W is drawn from a χ^2 distribution with n_{obs}-p degrees of freedom; then a value β^* is drawn from the $N(\hat{\beta},\sigma^{*2}(X^TX)^{-1})$ distribution, where $\hat{\beta}$ is the least-squares estimate of β and X now denotes the design matrix.

Given the values (β^*,σ^*) drawn in the first stage, the second stage draws a set of imputes for the missing Y-values. The second-stage procedures for the four imputation methods considered here are described next. In the descriptions, the term "conditional predictive mean" for case i refers to $X_i\beta^*$, the mean of the case given β^*.

Model-Based (MB) Method

For incomplete case i, the MB method imputes a value Y_i^* from the $N(X_i^T\beta^*,\sigma^{*2})$ distribution. Thus the impute is the conditional predictive mean of the incomplete case with some Gaussian noise added on. Since this method relies completely on the normal-theory linear regression model, it is likely to be the preferred method when the model holds exactly. It is also likely to be the most sensitive to violations of that model.

Residual-Draw (RD) Method

The RD method (Kalton 1983, p. 79) imputes the value

$$Y_i^* = X_i^T\beta^*+r_0^* = Y_0+(X_i^T-X_0^T)\beta^*$$

for incomplete case i, where X_0 and Y_0 are data values observed for a complete case drawn at random (with replacement and with equal probability) from the n_{obs} complete cases and r_0^* is the residual of that case when $\beta=\beta^*$. The RD method, like the MB method, imputes the conditional predictive mean of the incomplete case and additional noise. Because the noise for the RD method is drawn from the empirical distribution, however, this method should be less sensitive to violations of the normality assumption.

Predictive-Mean-Matching (PMM) Method

For each incomplete case, the PMM method (Little 1988) draws a case randomly from a set of complete cases having conditional predictive means close to that of the incomplete case, and then imputes the value of Y from the selected case to the incomplete case. Thus the PMM method may be thought of as a "hot-deck" procedure (Ford 1983). Preliminary results showing an increase in bias for this method with larger sets of available complete-cases led to using only the three closest cases in this study. Since the PMM method uses the normal-theory linear regression model only to define the distance between cases, it is the least parametric of the techniques studied here and should be less sensitive to violations in the model than the other methods. In addition, unlike the other methods, the PMM method imputes only realistic values that have been observed in the data set.

Local-Residual-Draw (LRD) Method

The LRD method contains aspects of both the RD and PMM methods. For each incomplete case, this method imputes the value

$$Y_i^* = X_i^T\beta^*+r_0^* = Y_0+(X_i^T-X_0^T)\beta^*,$$

where X_0, Y_0, and r_0^* are defined as in the RD method except that complete

case providing the residual is now drawn at random from a set of complete cases having conditional predictive means close to the conditional predictive mean of the incomplete case. Preliminary results led to a choice of the ten nearest cases as the available set. The LRD method is a variant of a method discussed in Scheuren (1976) and Schieber (1978). By using only residuals from nearby cases, the LRD method is designed to adjust the imputations locally for lack of fit of the regression model or heterogeneity of variance. LRD imputation may also be viewed as a modification of the PMM method in which an adjustment factor $(X_i^T - X_0^T)\beta^*$ is added to the imputed Y-value. The adjustment is intended to correct for possible bias due to the distance between the predictive means of the missing observation and its complete-case match. Since this correction is model-based, however, it should be less useful when the model is misspecified.

3. Design of the Monte Carlo Study

The Monte Carlo study was designed as a 2x2x3x3x5x8 factorial experiment with five outcomes per cell.

3.1 Factors in the Design

True Model for the Data

Samples of data were simulated as follows. For each observation, $X_i=(1,X_{i1},X_{i2})^T$ was generated with X_{i1} and X_{i2} distributed independently as N(5,1); then Y_i was generated from the model

$$Y_i^{\lambda}=X_i^T\beta+\epsilon_i,\ \epsilon_i \sim N(0,\sigma^2).$$

Data from linear and nonlinear models were created by setting $\lambda=1$ and $\lambda=1/4$, respectively; for the linear model, $\beta=(10,1,1)^T$ whereas for the nonlinear model, $\beta=(0,1,1)^T$.

Variance of the Error Term

The variance of the error term was set at two levels, $\sigma^2=1$ and $\sigma^2=2$. These correspond to the two population values for the squared multiple correlation coefficient, $\rho^2=.67$ and $\rho^2=.50$, respectively.

Expected Fraction of Missing Y-Values

The expected fractions of missing Y-values considered were 20%, 50%, and 80%, with the last one being mainly of interest for understanding the methods rather than representing realistic situations.

Missing-Data Mechanism

The three missing-data mechanisms considered were (i) missing completely at random, (ii) missing at random dependent on X and positively correlated with the predicted value of Y, and (iii) missing at random dependent on X but uncorrelated with the predicted value of Y.

Quantity of Interest (Q)

The eight quantities of interest considered were the percentages of the Y population greater than the 10th, 25th, 50th, 75th, and 90th percentiles, and the regression coefficients β_0, β_1, and β_2.

Imputation Method

The four imputation methods described in Section 2 were considered with M=5 imputations per missing value. Regardless of the model used to generate the true data, the imputation methods assumed a linear mean structure; thus the imputation model was misspecified when the true model for the data was nonlinear. In addition to the analyses with missing data, an analysis was performed using each entire data set before any observations were deleted; this "no-missing-data" analysis was used to standardize some of the results for the imputation methods.

3.2 Outcomes in the Experiment

On each Monte Carlo trial, a sample of size n=250 was generated for each cell in the factorial design according to the true model for the data and the variance of the error term. Values of Y were randomly deleted as specified by the missing-data mechanism and the expected fraction of missing values. For each imputation method separately, the missing values were multiply imputed, and the estimates and t-based confidence intervals described in Rubin and Schenker (1986) were computed for the quantity of interest. Even when an incorrect model was used in creating the imputes (corresponding to f in (1)), the correct model was used in the analysis (corresponding to g in (1)) so that the effect of misspecifying the imputation model could be isolated.

Two thousand Monte Carlo replications were simulated, and five outcomes were examined. The outcomes were the bias, variance, and mean squared error of the point estimator, and the coverage rate and average width of its associated nominal 95% confidence interval.

4. Monte Carlo Results

This section discusses the Monte Carlo results for estimating the percentages of the Y population greater than the five percentiles listed in Section 3.1, with Y-values missing completely at random at expected rates of 20% and 50%. Other results from the study, which are qualitatively similar to those presented here, are available from the authors.

The results are summarized in the table below. The outcomes for error variances of σ^2=1 and σ^2=2 have been averaged because the results for the two cases were similar. In addition, the measures of performance have been averaged over the five percentiles of the Y distribution for brevity. Monte Carlo variances, mean squared errors, and average interval widths have been standardized for each cell in the design by dividing by the value obtained from the "no-missing-data" analysis defined in Section 3.1.

Table of Monte Carlo Measures of Performance for Estimating Percentages of the Y Population with Data Missing Completely at Random

Measure of Performance	MB	RD	PMM	LRD
Linear Mean Structure				
|Bias|				
20% missing	.044	.047	.060	.055
50% missing	.078	.047	.108	.053
Variance				
20% missing	1.03	1.03	1.20	1.16
50% missing	1.23	1.22	1.81	1.65
Mean Squared Error				
20% missing	1.03	1.03	1.20	1.16
50% missing	1.23	1.22	1.81	1.65
|Error in Coverage Rate|				
20% missing	.630	.550	1.78	1.17
50% missing	1.78	1.33	4.59	2.99
Average Interval Width				
20% missing	1.09	1.08	1.07	1.08
50% missing	1.31	1.23	1.19	1.19
Nonlinear Mean Structure				
|Bias|				
20% missing	.966	.813	.079	.070
50% missing	2.45	2.05	.107	.098
Variance				
20% missing	1.06	1.05	1.21	1.17
50% missing	1.42	1.36	1.82	1.64
Mean Squared Error				
20% missing	1.28	1.19	1.21	1.17
50% missing	2.76	2.24	1.82	1.64
|Error in Coverage Rate|				
20% missing	1.66	1.22	1.09	.630
50% missing	7.06	5.85	4.70	2.58
Average Interval Width				
20% missing	1.12	1.10	1.07	1.08
50% missing	1.39	1.28	1.20	1.20

Note: Measures are averaged over the five percentiles and the two error variances. Variances, mean squared errors, and interval widths are standardized.

Bias

Under the linear population model, all four imputation methods display low bias; the largest average absolute bias is .108% for the PMM method with a 50% missing-data rate. Under the nonlinear model, however,

the biases of the MB and RD methods increase significantly, averaging over 2% in absolute value for the 50% missing-data rate; in contrast, the biases of the PMM and LRD methods remain low. Clearly, the less parametric design of the PMM and LRD methods helps to avoid biases when the population mean structure is misspecified.

Variance

The MB and RD methods typically produce lower variances than the PMM and LRD methods, with the ratio of average variances being as small as two-thirds for the linear model with 50% missing data. However, the variances of the MB and RD methods increase when the true population model is nonlinear rather than linear, whereas the performances of the PMM and LRD methods barely differ under the two models. Some features that are hidden by the averaging in the table of results include the following. All four methods have nearly constant variances across the five percentiles under the linear model, and the variances remain nearly constant for the PMM and LRD methods under the nonlinear model. The variances for the MB and RD methods, however, become unstable for the nonlinear model, ranging from .9 to 2.3 for the MB method with 50% missing data.

Mean Squared Error

Under the linear model, the variance is the dominant component of the mean squared error, so that the comparisons made above for the variances of the methods apply here as well. When the true model is nonlinear, however, bias becomes a significant component for the MB and RD methods. Once again, the results for the MB and RD methods are unstable across the five percentiles of Y; however, more often than not, the mean squared errors of the MB and RD methods are higher than those for the PMM and LRD methods.

Coverage Rate

When data are missing at a 20% rate, the nominal 95% confidence intervals for all four methods have Monte Carlo coverage rates within 2.5% of the nominal level regardless of the percentile being estimated. For 50% missing data, the MB and RD methods outperform the PMM and LRD methods when the true model is linear, although the LRD method still performs quite well. With the nonlinear population model and 50% missing data, however, the PMM and LRD methods have more accurate coverage rates on average than the MB and RD methods, with the LRD method achieving the best results in most cases.

The PMM and LRD methods tend to yield coverage rates that are lower than the nominal level, whereas the MB and RD methods produce slightly conservative intervals when the true model is linear. While the results for the PMM and LRD methods are quite consistent across the two population models, the coverage rates of the MB and RD methods are quite volatile under the nonlinear model. For example, with 50% missing data, the coverage rates for the MB method range from 82% to 98%, whereas those for the PMM method range only from 88% to 92%.

Average Interval Width

The MB and RD methods typically produce slightly wider intervals than the PMM and LRD methods. This is consistent with the fact that the MB and RD methods tend to yield conservative intervals under the linear model, whereas the PMM and LRD methods have coverage rates that were lower than nominal in most cases.

5. Discussion

The PMM and LRD methods, which were designed to be less parametric, show promise of greater robustness than the more parametric MB and RD methods in two respects. First, the

less parametric techniques maintain very low biases under both the linear and nonlinear population models, whereas the biases of the more parametric techniques increase under the nonlinear model. Second, all measures of performance for the PMM and LRD methods are very consistent across the two models; in contrast, the measures for the MB and RD methods tend to be quite volatile under the nonlinear model.

The PMM and LRD methods tend to produce less efficient estimates than the more parametric methods. In addition, the coverage rates for the PMM and LRD methods tend to be less accurate under the linear model, especially for high fractions of missing data. These results, along with the fact that the PMM and LRD methods typically produce narrower intervals despite the higher variances of the estimators, suggest that the methods need to be refined. Further research will develop different criteria for designating complete cases as close to the incomplete case in question, and will investigate alternative schemes for sampling from the chosen complete cases.

Acknowledgements

This research was supported in part by Grant 001069-7-RG from the American Foundation for AIDS Research, Joint Statistical Agreement 89-17 with the U.S. Bureau of the Census, and the IBM/UCLA Joint Project in Supercomputing.

References

Box, G.E.P., and Tiao, G.C. (1973), Bayesian Inference in Statistical Analysis, Reading: Addison Wesley.

Ford, B.L. (1983), "An Overview of Hot-Deck Procedures," in Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies, W.G. Madow, I. Olkin, and D.B. Rubin (eds.), New York: Academic Press, 185-207.

Herzog, T.N., and Rubin, D.B. (1983), "Using Multiple Imputations to Handle Nonresponse in Surveys," in Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies, W.G. Madow, I. Olkin, and D.B. Rubin (eds.), New York: Academic Press, 185-207.

Little, R.J.A. (1988), "Missing-Data Adjustments in Large Surveys," J. Bus. Econ. Statist., 6, 287-301.

Raghunathan, T.E. (1987), Large Sample Significance Levels from Multiply-Imputed Data, Ph.D. Thesis, Department of Statistics, Harvard University.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse," Proc. Survey Res. Meth. Sect., Amer. Statist. Assoc., 20-34.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, New York: Wiley.

Rubin, D.B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," J. Amer. Statist. Assoc., 81, 366-374.

Rubin, D.B., and Schenker, N. (1987), "Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes," J. Off. Statist., 3, 375-387.

Scheuren, F.J. (1976), "Preliminary Notes on the Partially Missing Data Problem - Some (Very) Elementary Considerations," working paper, Social Security Administration Methodology Group, Washington, DC.

Schieber, S. (1978), "A Comparison of Three Alternative Techniques for Allocating Unreported Social Security Income on the Survey of Low-Income Aged and Disabled," Proc. Survey Res. Meth. Sect., Amer. Statist. Assoc., 212-218.