Multiple Imputations Using Sequential Semi and Nonparametric Regressions

Irina Bondarenko[1], Trivellore Raghunathan[2]
[1]Department of Biostatistics, University of Michigan, 2029 Washington Heights, SPH-II, Ann Arbor, Michigan, 48109
[2]Department of Biostatistics, University of Michigan, 2029 Washington Heights, SPH-II, Ann Arbor, Michigan, 48109-2029

Abstract

Multiple imputation is a general purpose method for analyzing data with missing values. Under this approach the missing set of values is replaced by several plausible sets of missing values to yield completed data sets. Each completed data set is then analyzed separately and the results (estimates, standard errors, test statistics etc) are combined to form a single inference. It is fairly well established that the imputations should be draws from a predictive distribution of the missing values and should condition on as many covariates as possible. A sequential regression imputation method uses a Gibbs sampling style iterative process of drawing values from a predictive distribution corresponding to a sequence of conditional regression models to impute the missing values in any given variable with all other variables as predictors. The conditional regression models are usually parametric. In practice, however, many variables have distribution that very difficult to classify or transform to satisfy standard parametric distribution assumptions. We develop and evaluate a modification of this method. We construct propensity score for missing the given variable and the predicted value of that variable. We stratify the sample based on these two scores and then within each stratum, we use approximate Bayesian Bootstrap or Tukey's gh distribution to impute the missing values conditional on the observed values. We illustrate proposed method using actual and simulated data sets.

KEY WORDS: Missing Data, Response Propensity, Predictive Mean Matching

1. Introduction

Multiple imputation is becoming an increasingly popular approach for analyzing incomplete data. In this approach the missing set of values are replaced by more than one plausible set of values to yield several completed data sets. Each completed data is analyzed and the inferential statistics such as estimates, standard errors, test-statisitics etc are combined to form a single inference. It is fairly well established that the imputations have to be draws from a predictive distribution of the missing values and should condition on as many variables as possible (Little and Raghunathan (1997), Schafer et al (2000)). This is a tall order given that the data set may contain tens and hundreds of variables of varying type with complex structural and stochastic inter-relationships. Developing a model (joint distribution of the variables with missing values conditional on observed values and unknown parameters and then to obtain draws from the corresponding predictive distribution is difficult, if not impossible.

An approach particularly suited to such complex situation is a sequential regression approach. This approach involves Gibbs sampling style iterative sampling from a sequence of conditional regression models where the missing values in any one variable is drawn from the predictive distribution corresponding to the regression model and uses all other variables (including interaction terms) as predictors. This approach was first used in the Survey of Consumer Finances by Kennickell (1991) to impute missing values in continuous variables using a sequence of normal linear regression models. This approach was generalized to a variety of types of variables and incorporates complexities such as bounds on the imputations and the skip patterns (Raghunathan et al (2001)). This general approach has been implemented in stand-alone and as add on to commercial packages (SRCWARE (Standalone), IVEWARE (SAS), Raghunathan et al (1997), MICE (R-package) first described by Burren at al. (1999) and ICE (STATA) by Royston (2004). The sequential regression approach implemented in these packages use a sequence of parametric models such as normal linear model for continuous variables, logistic for binary, Poisson for count etc and some

facilitate incorporation of structure dependencies and constraints. This approach seems to work well provided parametric assumptions are approximately satisfied.

Nevertheless, variables collected in many practical situations rarely satisfy the underlying parametric assumptions and imputing them using parametric model may introduce bias. Attempts to transform a continuous variable to achieve an approximate normality may be not fruitful Rubin (1987) and He and Raghunathan (2006)). Regrouping, or imputing multilevel variable as continuous and applying rounding afterwards, may be a plausible solutions, but if for the future analysis the number of categories has to preserved, then another approach has to be found. Furthermore, imputation of missing values can be very sensitive to violation of underlying model specification. Consider as an example, data from Sacramento Area Latino Study of Aging (SALSA). It's an ongoing cohort study of 1,789 Latinos aged 60 and older in 1998-99 residing in rural and urban areas of the Sacramento Valley. Aging researches were interested in neuropsychological characteristics and prevalence of dementia in aging Latino population. The neuropsychological test battery includes Informant Questionnaire of cognitive Decline in Elderly (IQCODE). Histogram of observed values of IQCODE score is shown on the (Figure 1a). It's observed for 64% of 1786 subjects. It's very hard to transform this variable to achieve normality of the corresponding residuals. Figure 1b shows Kernel Density estimates of observed, imputed and completed values, when data are imputed assuming normal distribution for IQCODE. There is a remarkable difference in shape of the distribution of observed and imputed values that may signal a poor quality of the imputation.

In this paper we extend sequential regression approach by making it less dependent on the parametric framework, and propose the following iterative procedure. If the variable has binary or approximately normal distribution it's imputed using a parametric model. For the variables with distributions that do not fit in the parametric model assumptions we use the following procedure:

- At a given iteration, we construct two summary scores: propensity score for missingness based on a logistic regression model and the predicted values for all subjects using a regression model. All other variables, imputed or observed, are used as predictors in both models. This combines the ideas of propensity score and predictive mean matching.
- The subjects with observed and missing values are grouped into strata based on these two scores. For example, create quintiles or quartiles of propensity scores and then further subdivide each propensity score class into quintiles are quartiles of predicted values.
- Within each matched strata impute the missing set of values employing either 1) Approximate Bayesian Bootstrap (Rubin and Schenker (1986), Rubin (1987)) , or 2) or Tukey's *gh* distribution (He and Raghunathan (2006)).

The rest of paper is organized into 4 sections. Details of the two-step algorithm are presented in section 2 and imputation the IQCODE variable is revisited in section 3. Section 4 focuses on simulation study based on the National Health Interview Survey (NHIS) data. Possible development and discussion can be found in section 5.

2.Description of the Method.

Suppose, that the data set has p variables, $Y_v, v = 1, 2, ..., p$. Let R_v denote a binary response indicator with 0 for missing and 1 for observed value of Y_v. Let Y_{-v} denote the collection of all *p-1* variables except Y_v. With slight abuse of notation, partition the vector or matrix of observations on the n subjects as $Y_{com,v} = (Y_{obs,v}, Y_{mis,v})$ and $Y_{com,-v} = (Y_{obs,-v}, Y_{mis,-v})$. Suppose at iteration t, $Y^{(t)}_{com,-v}$ is the completed data on all subjects except for variable v and $Y^{(t)}_{com,v}$ is the completed data on all subjects for variable v.

We construct two efficient summaries of the covariates $Y^{(t)}_{com,-v}$ through two regression prediction:

1. Propensity of missingness, $e^{(t)}_v = \Pr(R_v = 1 | Y^{(t)}_{com,-v})$, estimated using a logistic regression model . It's is an efficient

summary of $Y_{com,-v}^{(t)}$ that can be used to balance the respondents and nonrespondents (Rosenbauum and Rubin (1983)). The predictors may include interaction terms to achieve balance between respondents and nonrespondents. We stratify $e_v^{(t)}$ into the K equal size strata.

2. Predicted value based on a regression of $Y_{com,v}^{(t-1)}$ on $Y_{com,-v}^{(t)}$ within each class. We create J equal size strata and thus forming $K \times J$ match classes. The prediction may be based on either the parametric model or semi-parametric model (for example, generalized additive model).

We consider two possible options for imputing missing values within each cell. The first approach is to use Approximate Bayesian Bootstrap method (Rubin and Schenker (1986)). Suppose r and m are the observed and missing observations, respectively. The basic approach is to draw a sample of size r with replacement from the observed set and from this sample draw a sample of size m as the imputed values. Repeat this step for all the cells and independently replicate the process to obtain multiple imputations.

The second approach uses Tukey's *gh* distribution in each cell as described in He and Raghunathan (2006). This distribution is based on a transformation of standard normal distribution to accommodate for skewness and elongation of the tails. The *g-h* method expresses random variable *Y* as a monotonic function of a standard Gaussian random variable Z:

$$Y_{gh}(Z) = \mu + \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2},$$

where μ is location parameter, σ is a scale parameter, g - reflects the skewness, and h governs elongation (heaviness) of tails. Varying parameters allows us to accommodate a wide spectrum of deviation from normality, and make gh family a flexible tool to model observed distribution parametrically.

Hoaglin (1985) developed simple methods to estimate parameters of gh distributions through empirical quintiles of observed variable Y. Specifically, let y_p and z_p be percentiles of the variable being imputed and standard normal distributions, respectively. We would like to fit the model $y_p = Y_{gh}(z_p)$.

It can be shown that

1) $\mu = y_{0.5}$,

2) g can be estimated as a slope of the regress

$$-\log\left(\frac{y_{1-p} - y_{0.5}}{y_{0.5} - y_p}\right) = g\, z_p,$$

3) σ and h can estimated as intercept and slope

$$\log \frac{g(y_{1-p} - y_{0.5})}{e^{-gz_p} - 1} = \log\sigma + h\frac{z_p^2}{2}$$

The essential steps are to draw a bootstrap sample of size r from the r observed set of observations. Use the bootstrap sample to estimate the parameters (μ, σ, g, h). Draw m independent standard normal random variables and then apply the transformation to yield imputed values.

3. Example revisited.

Revisiting the IQCODE example, we stratified IQCODE variable based on its propensity score of being missing and its predicted value into 9 strata. We imputed all other variables in the SALSA data set following the SRMI approach(Figure 1b) and imputed IQCODE using the proposed sequential approach use both ABB(Figure 1c) and GH (Figure 1d) methods. Kernel density estimates based on five imputations for imputed and true values of IQCODE are shown below. There is a remarkable improvement in the matching between marginal distribution of observed and imputed values

4. Simulation study

We conducted a simulation study to assess repeated sampling properties of multiple imputation inferences using the proposed approach. The simulation study consisted of the following steps:

1. Creation of population: A pseudo population of 200,000 records was constructed from the National Health Interview Survey for 1997 to 2003. The observations were the fully observed values of eight variables (age, gender, weight, height, years of education, income to poverty ratio, self-reported hypertension and diabetes).

2. Sample: Draw 250 independent simple random samples each of size n=1000. We will call these before-deletion samples.
3. Delete some values of age, years of education, income to poverty ratio, self-reported hypertension and diabetes using a known missing data mechanism. No missing values were imposed on gender, weight and height. Self-reported diabetes and hypertension are binary variables with 2% and 4% observations reset to missing, correspondingly. Percentages of observations set to missing across 250 random samples are shown at Table1. Histograms of observed values for age, education and income to poverty ratio are also shown .We will call the resulting 250 replicate data sets with values reset to missing as "after-deletion" samples.

We imputed each of the 250 after-deletion samples following 3 different methods: SRMI, ABB and GH. The number of imputations was fixed at 5 for all three methods. For both ABB and GH imputations were based on $K = 4, J = 4$ (16 strata).

The results were assessed in terms of bias, mean square error and the confidence coverage of several parameters of interest, means, proportions and the regression coefficients. The regression model involved the number of chronic conditions (Hypertension+Diabetes+Obesity) as outcome and demographic variables and poverty income ratio as predictors,

$$\text{logit}(\Pr(Index \le k) = \sum_k \beta_{0k} + \beta_1 Age + \beta_2 Income$$
$$+\beta_3 I_{<HS} + \beta_4 I_{HS} + \beta_5 Gender$$

The results are summarized in Tables 2 (Marginal means and proportions) and Table 3 (for regression coefficients). Both semi-parametric (GH) and non parametric (ABB) approaches produce good coverage rates across all variables. ABB imputation estimates are slightly biased and have low MSE for all imputed variables. GH imputation estimates are almost unbiased, however, its performance appears to be affected by the shape of the distribution of observed values. First, for education GH estimates have a smaller bias and MSE when compared to SRMI , CC and ABB. Second, for age (standardized) that has continuous distribution, GH and ABB estimates have similar properties with both being biased. Third, for income to poverty ratio where ABB and CC have very similar estimates GH yields estimates with larger bias and relatively high MSE. Table 3 shows that all the regression estimates have similar properties with slightly better performance of ABB estimates.

5. Discussion

Nonparametric approaches offer flexibility of handling nonstandard distributions and sequential regression approach handles complex data structure. By combining features of both approaches, we proposed two modifications that allowed us to relax parametric assumption and impute missing data as random draws from 1) unspecified posterior predictive distribution in the case of ABB, and 2) posterior predictive distribution of the missing values under flexible gh model. Both methods can be implemented using standard software packages as it involves iterations of two steps: 1) Stratifications via regression models and 2) imputation step involving random sampling. This approach shows promise of being more robust and less susceptible to model misspecification.

References

1. Little, R. J.A. and Raghunathan,T.E. (1997) “Should Imputation of Missing Data Condition on All Observed Variables?” Proceedings of the Section on Survey Research Methods, 1997 Joint Statistical Meetings, Anaheim, California.
2. Schafer, J. L. (2000) Analysis of incomplete multivariate data. London: Chapman and Hall.
3. Kennickell, A.B. Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation, ASA 1991 Proceedings of the Section on Survey Research Methods, 1-10. Alexandria: ASA, 1991.
4. Raghunathan, T.E., Lepkowski, J. E., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology,27, 89-95.

5. Raghunathan, T.E., Van Hoewyk, J. and Solenberg, P. (1997). IVEWARE: Imputation and Variance Estimation Software. http://www.isr.umich.edu/src/smp/ive
6. Burren, S. V., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood presure covariates in survival analysis. Statistics in Medicine, 18, 681-694.
7. MICE Multivariate Imputation by Chained Equations): associated S+ software at www.multiple-imputation.com
8. Royston P., (2004) Multiple imputation of missing values. The Stata Journal, 3, 227-241.
9. Rubin, D., Schenker N., (1986) Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366-374
10. Rubin, D. (1987) Multiple imputation for Nonresponse in Surveys. New York. John Wiley and Sons.
11. He Y., Raghunathan, T.E., (2006) Tukey's gh Distribution for Multiple Imputation. The American Statistician, 60, 251-256
12. Rosenbaum P., Rubin D., (1983) The central role of the propensity score in observational studies for causal effects, Biometrika 70(1):41-55
13. Hoaglin, D.C. (1985), "Summarizing Shape Numerically: The g-and-h Distributions" in Exploring Data Tables, Trends, and Shapes,eds D.C. Hoaglin, F. Mosteller, and J.W. Tukey, New York: Willey, pp.461-513
14. Tukey, J.W. (1977) Modern Techniques in Data Analysis. NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.

Figure 1: Distributions of the Observed and Imputed Data for IQCODE: Histograms and Kernel Density Estimates (KDE)

Figure 1a.
Histogram: Observed IQCODE

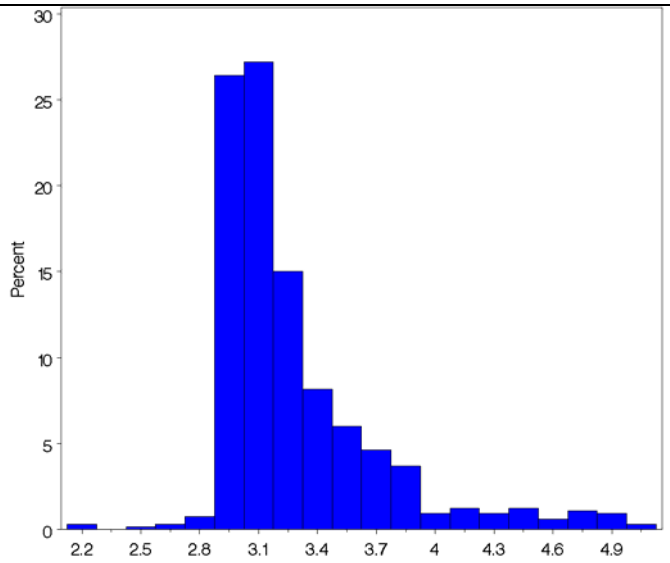


Figure 1b.
KDE of observed, imputed and completed values assuming normal distribution

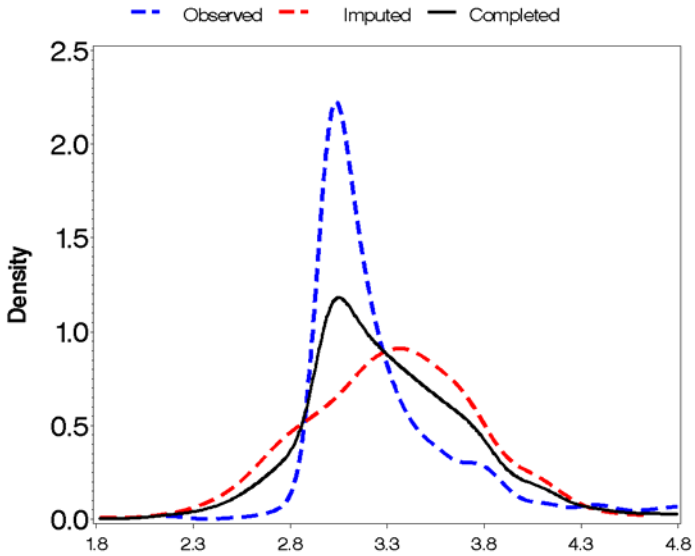


Figure 1c
Stratified ABBapproach

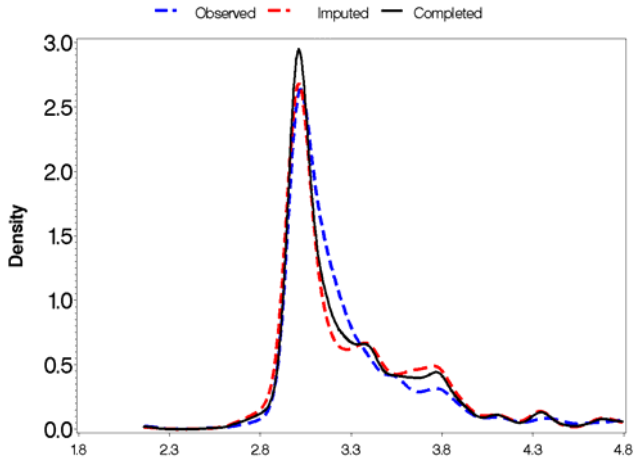


Figure 1d.
Stratified GH approach

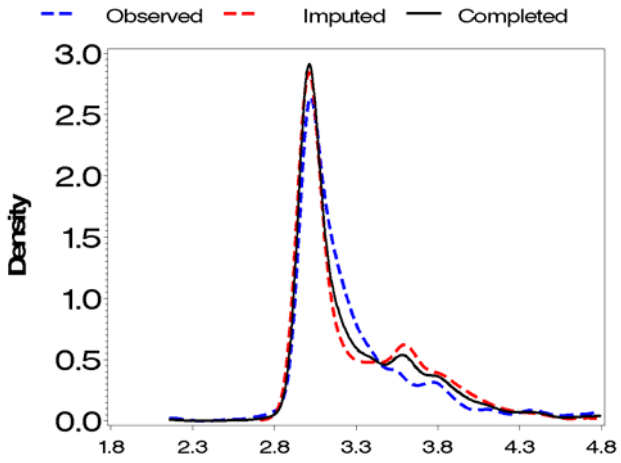


Table 1: Distribution and percent missing on three variables used in the simulation study

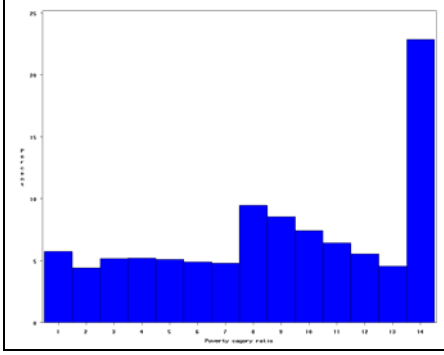
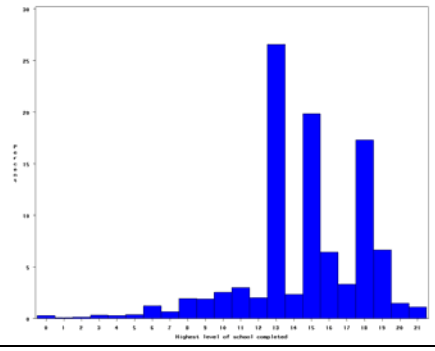
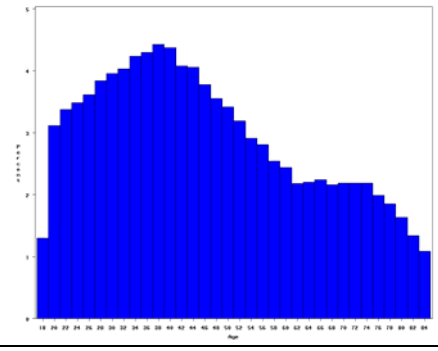
Income to poverty ratio (10% missing)	Education (21% missing)	Age (23% missing)
		

Table 2: Bias, Meansquare error and confidence coverage for marginal means and proportions.

	Bias				MSE				Coverage			
Variable	ABB	GH	SRMI	CC	ABB	GH	SRMI	CC	ABB	GH	SRMI	CC
Education												
Less than HS	-1.3	-0.9	3.1	-4.7	5.7	5.2	13.2	23.8	88	95	61	0
HS	0.4	0.8	-3.7	0.9	4.6	5.9	16.3	3.7	100	97	47	99
College degree	0.9	0.0	0.6	3.7	4.1	2.3	2.6	16.3	98	100	100	10
Age	-0.0	0.4	0.2	0.2	0.4	0.5	0.4	0.5	100	99	100	100
Income to poverty ratio												
<1	0.1	-1.1	-0.0	-0.2	1.7	2.7	1.7	1.5	100	98	100	100
>=1,<5	0.0	1.4	0.2	0.2	2.9	5.2	2.6	2.6	100	98	99	100
>=5	-0.2	-0.3	-0.1	0.0	2.3	2.7	2.0	2.0	100	100	100	100

Table 3: Simulation results for Bias, MSE, and coverage in estimation of Beta- coefficients from the proportional odds model

Predictor	*Estimate*	*ABB*	*GH*	*SRMI*	*CC*
Age	Bias (10^{-4})	-10	14	7	20
	MSE (10^{-6})	17	19	16	30
	Coverage (%)	100	71	80	80
Sex	Bias (10^{-3})	9	5	4	142
	MSE(10^{-2})	2	2	2	6
	Coverage (%)	100	94	96	89
Education					
<HS	Bias	0.07	-0.03	-0.05	-0.03
	MSE	0.08	0.08	0.06	0.12
	Coverage (%)	90	94	96	93
HS	Bias	-0.02	-0.04	-0.03	-0.07
	MSE	0.03	0.04	0.03	0.07
	Coverage (%)	100	86	89	85
Income to poverty ratio	Bias (10^{-4})	6	-78	-17	-300
	MSE (10^{-4})	4	5	4	20
	Coverage (%)	100	45	59	50