# Sequential nonparametric regression multiple imputations

**Irina Bondarenko and Trivellore Raghunathan**

**Department of Biostatistics, University of Michigan**

**Ann Arbor, MI 48105**

## Abstract

Multiple imputation, a general purpose method for analyzing data with missing values, involves replacing a missing set of values 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 be conditioned on as many covariates as possible. However building such predictive distributions in a practical setting can be quite daunting. As an alternative, 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 missing values in any given variable with all other variables as predictors. Many current implementation of this approach use parametric models. In practice, however, many variables have distributions that can hardly be classified or transformed to satisfy standard parametric distribution assumptions. This paper develops and evaluates a modification of this method which uses draws from nonparametric predictive distributions for imputing the missing values. For a variable to be imputed, two scores are constructed: Propensity scores for its missingness, and its predicted values. The sample is stratified based on these two scores and within each stratum, the

Approximate Bayesian Bootstrap (ABB) is used to impute the missing values. The proposed method is illustrated and evaluated using actual and simulated data sets.

**Key Words**: Approximate Bayesian Bootstrap, Missing Data, Nonresponse, Response Propensity, Predictive mean matching

## 1. Introduction.

Multiple imputation is becoming an increasingly popular approach for analyzing incomplete data. In this approach, a set of missing values is replaced by more than one plausible set of values to yield several completed data sets. Each completed data set is analyzed, and the inferential statistics, such as estimates, standard errors, test-statistics, 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 be conditioned on as many variables as possible (Little and Raghunathan [1], Paulin and Raghunathan[2], Schafer et al. [3]). This is a tall order given that the data set may contain several variables of varying types with complex structural and stochastic inter-relationships. Developing a model, a joint distribution of variables with missing values conditional on the fully observed variables with some unknown parameters, and then obtaining draws from the corresponding predictive distribution is difficult, if not impossible.

An approach particularly suited to such a complex situation is a chained equation or sequential regression multiple imputation (SRMI). This approach involves a Gibbs sampling style iterative sampling from a sequence of conditional regression models, where the missing values in any given variable are drawn from the predictive distribution corresponding to the

regression model, and uses all other variables (including interaction terms) as predictors. Specifically, suppose that $Y_1, Y_2, ..., Y_p$ are the variables with missing values and $X$ is the set of variables with no missing values. At iteration *t*, the imputations of variable, $Y_j$ are obtained as draws from the predictive distribution, $\Pr(Y_j \mid X, Y_1^{(t)}, Y_2^{(t)}, ..., Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, ..., Y_p^{(t-1)})$, where $Y^{(t)}$ denotes the variable that has either the observed value or the imputed value obtained at iteration *t*. This conditional distribution is based on a regression relating the variables being imputed and all other variables as predictors.

This approach was first used in the Survey of Consumer Finances by Kennickell [4] to impute missing values in continuous variables using a sequence of normal linear regression models. This approach has since been generalized to a variety of types of variables and incorporated complexities, such as bounds on the imputations and skip patterns[5,6]. This general approach has been implemented in both stand-alone and as an add-on to commercial packages, SRCWARE (Standalone), IVEWARE (SAS) [7], MICE (R-package) [8], and ICE (STATA) [9]. All these implementations use a sequence of parametric models, such as normal linear model for continuous variables, logistic for binary, Poisson for count, etc., and some of these packages facilitate incorporation of structure dependencies and constraints. This approach seems to work well, provided parametric assumptions are approximately satisfied. However, 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 not yield valid imputations [10].

Consider as an example, data from Sacramento Area Latino Study of Aging (SALSA)[11]. This is 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. Researches are 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). A histogram of observed values of IQCODE score is shown in Figure 1. This variable is missing for 64% of the sample subjects. Figure 2 shows Kernel Density estimates of the observed, imputed, and completed values, when data are imputed assuming normal distribution for the IQCODE. This figure indicates a remarkable difference in shapes of the distributions of observed and imputed values. None of the standard transformation techniques such as Box-Cox transformation to normality improves quality of imputations. We expect that imputed and observed data distributions to be similar only if the data are missing completely at random. However, the degree of difference between the distributions of the imputed and observed data is quite substantial, especially when our preliminary analysis did not identify covariates that are strong predictors of missingness in IQCODE.

**2. Nonparametric Sequential Regression**

We modify the sequential regression approach described above as follows:

i. **Matching:** We define two summary scores: (1) a propensity score for missingness of the variable being imputed (to balance the respondents and nonrespondents on the covariates)[12] and (2) a predicted-value-score (to match on the predictive distribution of the variable). To construct both scores we condition on all other variables, imputed or observed just as in the conditional distribution given above. We form several strata based on the joint distribution of these two scores.

ii. **Imputation:** Within each stratum, missing values are drawn from the set of observed values by applying Approximate Bayesian Bootstrap [13-14].

This approach combines the features of survey weighting, carried through grouping of respondents and nonrespondents into the strata balanced on all covariates and designed to reduce the nonresponse bias, and the predictive mean matching imputation strategies [15-16]. By matching on the covariates through response propensity score and as well as the predicted values, we are creating homogenous groups on available information and then using the nonparametric posterior predictive distribution within each group to draw the missing values.

***Matching***

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 a slight abuse of notation, partition the vector or matrix of observations on *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 models:

1. Propensity of missingness, $e_v^{(t)} = \Pr(R_v = 1 | Y^{(t)}_{com,-v})$, estimated using a logistic or generalized additive model. This is an efficient summary of $Y^{(t)}_{com,-v}$ that can be used to balance the respondents and nonrespondents [12]. The predictors may include interaction

terms to achieve balance between respondents and nonrespondents. We stratify $e_v^{(t)}$ into the K equal size strata.

Predicted-value-score is defined by regressing $Y_{obs,v}$ on $Y_{com,-v}^{(t)}$ within each class. We use estimated parameters of the regression model with dependent variable $Y_{obs,v}$ to define predicted values for $Y_{mis,v}$ Within each of the K strata we create J equal-size strata and thus forming $K \times J$ match classes.

***Impuation.***

An approximate Bayesian Bootstrap algorithm for imputation can be implemented as follows [13]. Suppose *r* and *m* are the numbers of the observed and missing observations, respectively, on the variable being imputed. First, draw a sample of size *r* with replacement from *r* observed values. Second, 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.

Revisiting SALSA study, we imputed the missing values IQCODE variable using this method. We used a total of 9 matching classes based on the propensity scores and predicted values. Kernel density estimates based on five imputations for imputed and true values of IQCODE are shown in Figure 3. There is a remarkable improvement in the matching between marginal distribution of observed and imputed values.

**3. Simulation study**

We conducted a simulation study to assess repeated sampling properties of inferences based on imputed data sets obtained using the method discussed in the previous section. 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 from 1997 to 2003. Fully observed records of eight variables (age, gender, weight, height, years-of-education, income-to-poverty-ratio, self-reported hypertension and diabetes) constituted our population.
2. 500 independent simple random samples, each of size $n$=1000, were drawn. We will refer to them as the "before-deletion" samples.
3. Some values of age, years of education, income-to-poverty-ratio, self-reported hypertension and diabetes were deleted using a known missing at random mechanism. We modeled this mechanism employing a series of logistic regression models, with probability of a variable being missing depending solely on observed values of other variables. We will call the resulting 500 replicate data sets "after-deletion" samples.

No missing values were imposed on gender, weight or height. Figure 4 gives histograms of observed values for age, education, and income-to-poverty-ratio, the descriptive statistics along with the rate of missingness across 500 replicates. Missing data mechanisms for self-reported diabetes and hypertension are binary variables yielded small amount of missing data because our concern was mainly to impute continuous type variables where normality is questionable.

As an evaluation, we compared repeated sampling properties of the following three approaches for analyzing incomplete data: (1) Complete cases (CC) analysis; (2) Parametric sequential regression as implemented in IVEWARE (SRMI); (3) ABB with propensity estimated employing logistic model, and predicted score constructed using General linear model for age and years of education, and Weibulll model for income-to-poverty-ratio. We used 'before-deletion' samples (BD) as our reference.

The binary variables, hypertension and diabetes, were imputed using logistic regression models under all imputation methods. In non-parametric sequential imputation, education, age, and income-to-poverty-ratio were imputed with subjects matched into 9 nested (3x3) strata based on the propensity score and predicted values. In the SRMI framework, we imputed education and age as continuous variables and income-to-poverty-ratio as a categorical variable.

***Simulation results***

For each method, we compared the complete-case or multiply imputed inferences to the before-deletion data inferences. We evaluated the bias, mean square error, and non-coverage rate of a nominal 95% confidence intervals for descriptive as well as analytical estimands. For each imputation method, we performed 5 imputations and computed a multiple imputation estimate and its multiple imputation variance using the standard multiple imputation combining rules [14]. The results for the population means and frequency distributions for categorical and grouped continuous variables are summarized in Table 1 and Figure 5.

As expected, analysis based on Complete Cases yielded a substantial bias, large MSE and lowest coverage rates. SRMI approach yields better inferences for population mean estimates, but not for the estimated frequency distribution. Across the three approaches considered here, ABB method produced the best results in terms of bias, MSE and coverage rates. For both the population mean and the frequency distribution estimates, ABB method introduced less bias than SRMI method. Though, there is a reduction in bias for population mean estimates, the improvement becomes more apparent for frequency estimates.

For the population means, ABB estimates have similar or slightly lower MSE when compared to SRMI estimates. For the frequency distribution estimates the abatement in MSE is most obvious for years-of-education.

We expanded the simulation study to consider whether we need to stratify on both the propensity score for missingness and the predicted values. We compared the bias, means-square error and confidence coverage for the population mean of Age variable under two scenarios: (1) Stratifying based only predicted values of the variable being imputed and (2) Stratifying based on both propensity score for missingness and predicted values. In both cases, we used the same number of strata and, hence, a finer stratification under scenario (1). The bias for the population mean under scenario (1) was 0.0746 and 0.007 under scenario (2). The mean square error under the two scenarios were 0.0236 and 0.2053, respectively. Finally, the coverage rates were 86% and 92% respectively. Thus, the simulation study seems to indicate that one should consider both the propensity score for missingness and the predicted value of the variable being imputed to create the imputation cells.

We also evaluated the methods by comparing inferences for the following proportional log-odds model regressing an ordinal outcome,- the number of chronic conditions, on demographic variables as predictors. We constructed our dependent variable as the sum of indexes for hypertension, diabetes, and obesity and chose gender, age, education, and income-to-poverty-ratio as predictors.

We categorized education as 'no HS degree', 'HS graduate', and 'college degree'. Income-to-poverty ratio was classified into 'less than 5' and '5 or more'

$$\log it(\Pr(Index <= k)) = \beta_{ok} + \beta_1 Age + \beta_2 HS + \beta_3 College \ \deg ree + \beta_4 (\frac{Income}{Poverty} >= 5) + \beta_5 Gender$$

The results are reported in Table 2. The ABB estimates are shown to have repeated sampling properties are better than the SRMI estimates. For *age* and *gender*, SRMI performs well, and ABB approach does not bring substantial reduction in bias, or MSE, or non-coverage rate. For the variables years-of-education and income-to-poverty-ratio that were grouped for the analysis purposes, there is a clear gain with ABB method. For example, for the 'HS graduate' category there is 95% reduction in bias, 50% decrease in MSE, and change in coverage rate from 78% to 95%.

## 4. Conclusion

Nonparametric sequential regression approach offers a flexibility of handling nonstandard distributions for the complex data structures where developing a joint distribution of all the variables with missing values is difficult. As we've seen from the simulation study, when the data don't satisfy parametric assumptions of the imputation model, the estimates derived from SRMI approach can be biased. The alternative, ABB approach, discussed in this paper seems to perform well under all simulation conditions considered.

The SRMI method is robust with respect to the central tendency, but imputes missing values under a particular distribution assumption, that could create problems at the tails of the distribution. ABB allows preserving a shape of an original distribution, and earns better estimates for percentiles of the distribution. Thus, for a multilevel categorical variable or continuous variable with a nonstandard distribution, ABB method produces better results. This method can be implemented using standard software packages as it involves iterations of two steps: 1) Stratifications via regression models and 2) imputation step involves random sampling. This approach shows promise of being more robust and less susceptible to model

misspecification. The proposed approach may be more useful for routine applications of sequential regression multiple imputation where performing model diagnostics and developing appropriate models can be very difficult. In this approach, the predicted values may be obtained using a nonparametric regression model instead of parametric regression model. However, several simulation studies indicate that the results are not sensitive to miss-specification of the model used to predict the missing values.

## References


1. Little RJA, Raghunathan TE. Should Imputation of Missing Data Condition on All Observed Variables? *Proceedings of the Section on Survey Research Methods*, ASA, Anaheim, CA, 1997
2. Raghunathan T.E, Paulin G.S. Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics, ASA, Dallas, TX,* 1998
3. Schafer JL, Chapman and Hall: London, 2000.*Analysis of incomplete multivariate data.*
4. Kennickell AB. Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation, *Proceedings of the Section on Survey Research Methods*, ASA, Alexandria, VA,1991.
5. Raghunathan TE, Lepkowski JE, Van Hoewyk J and Solenberg P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology* 2001; **27**: 89-95.
6. Van Burren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**: 681-694.
7. Raghunathan TE, Van Hoewyk J, Solenberg P. IVEWARE: Imputation and Variance Estimation Software, 1997. http://www.isr.umich.edu/src/smp/ive
8. Van Buuren S, Oudshoorn CGM. *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual. PG/VGZ/00.038.* TNO Preventie en Gezonheid, Leiden,2000

9. Royston P.Multiple imputation of missing values. *The Stata Journal* 2004, **3**:227-241.
10. Rubin D. A Case-Study of the Robustness of Bayesian Methods of Inference: Estimating the Total in a Finite Population Using Transformations to Normality. *Scientific Inference, Data Analysis and Robustness.* Academic Press,Inc.: New York , 1983 pp.213-244.
11. González HM, Haan MN, Hinton L. Acculturation and the Prevalence of Depression in Older Mexican-Americans: Baseline Results of the Sacramento Area Latino Study on Aging, *JAGS* 2001*;* **49(7)**, 948-53.
12. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects, *Biometrika* 1983; **70(1)**: 41-55.
13. Rubin D, Schenker N. Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association* 1986; **81**: 366-374
14. Rubin D. *Multiple imputation for Nonresponse in Surveys.* Wiley: New York, 1987, p 124
15. Little, R.J.A.. Missing data in large surveys. *Journal of Business and Economic Statistics* 1988; **6**: 287-301 (with discussion).
16. Schenker N, Taylor J.M.G. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* 1996; **22 (4)**: 425-446

Figure 1: Histogram of observed values of IQCODE from the SALSA Study

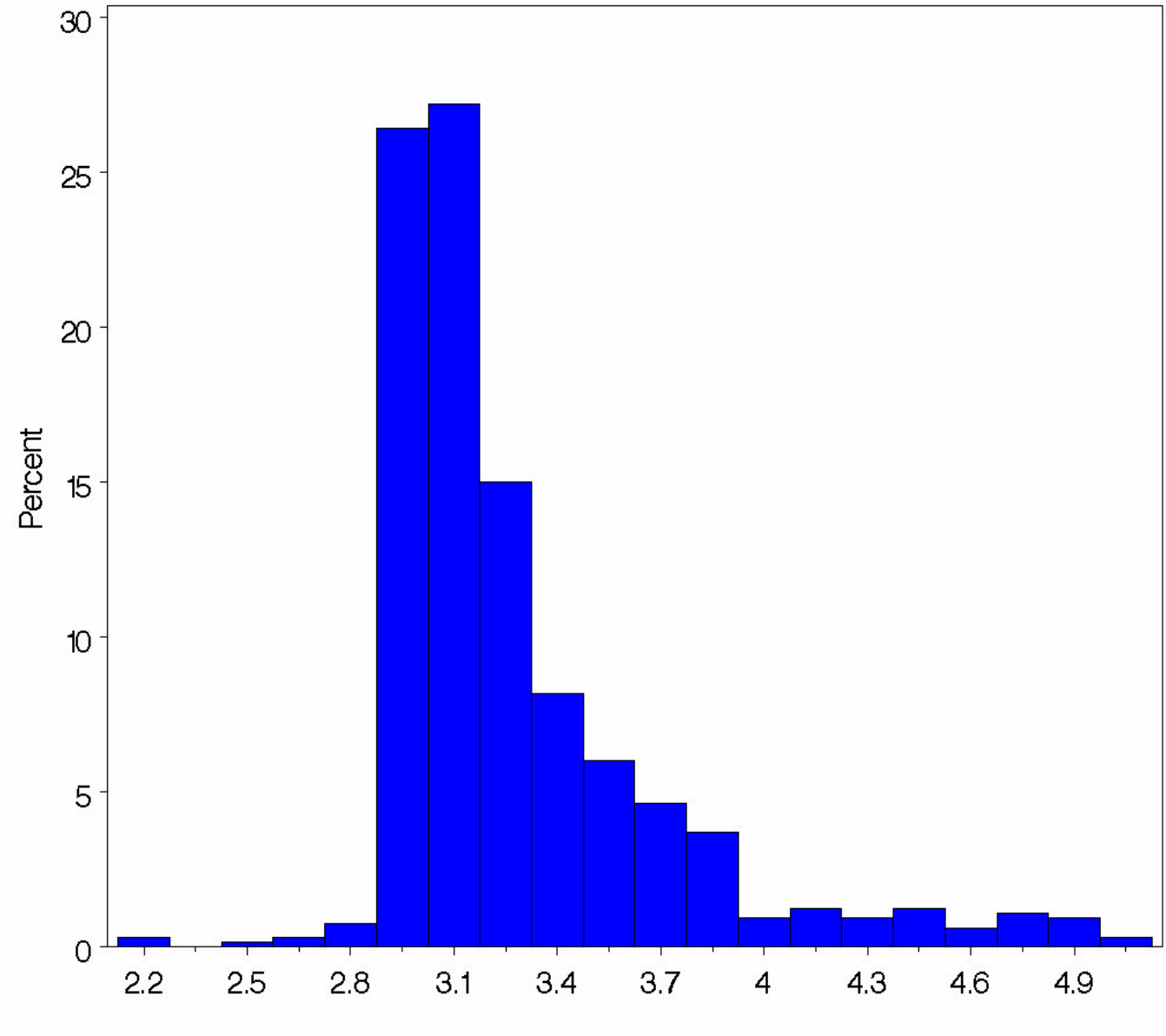


Figure 2: Kernel Density Estimates of the observed, imputed(SRMI) and completed data for variable IQCODE in the SALSA Study.

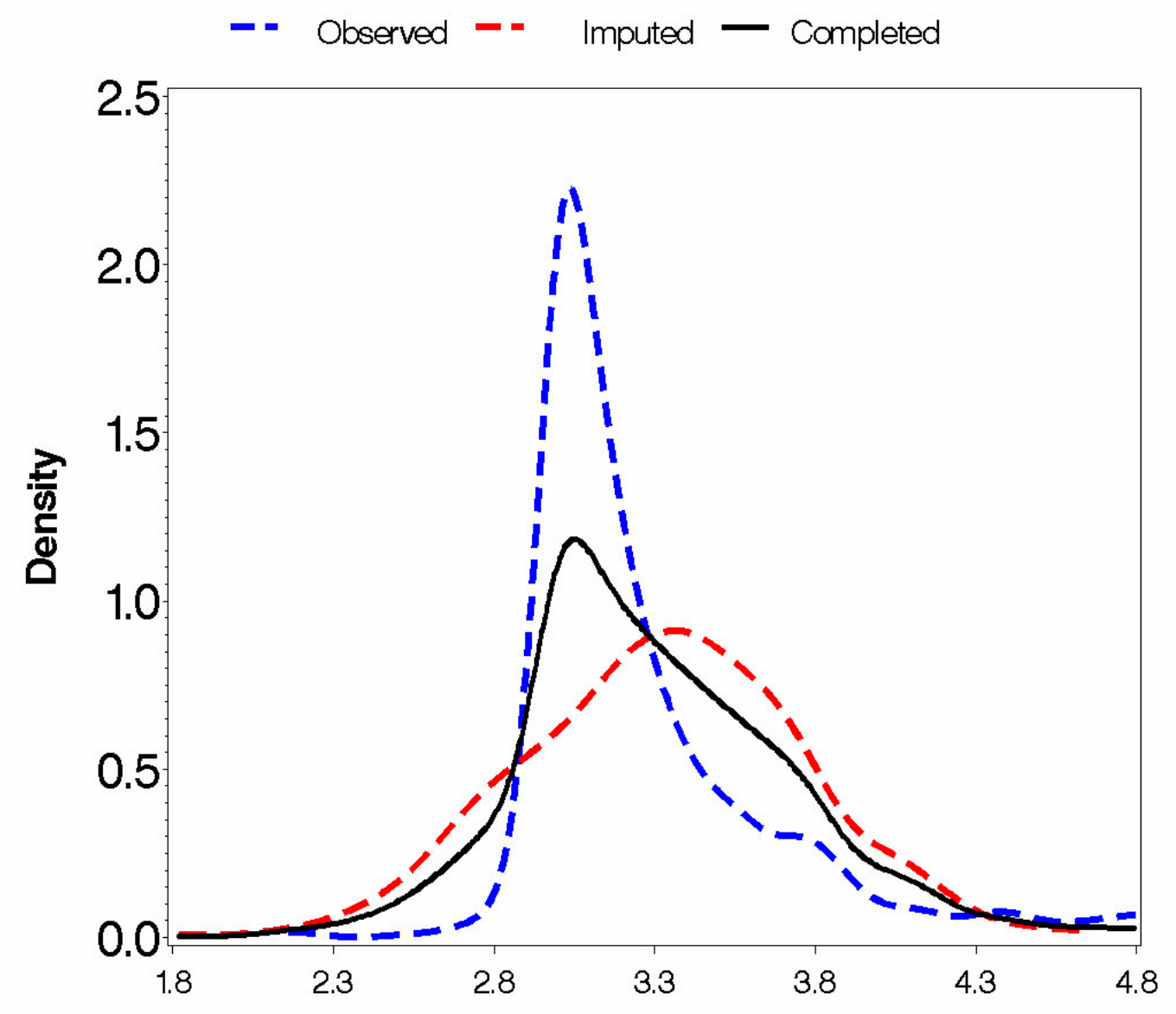


Figure 3: A comparison of Kernel Density estimates of observed, imputed (ABB) and completed data on IQCODE from the SALSA Study.

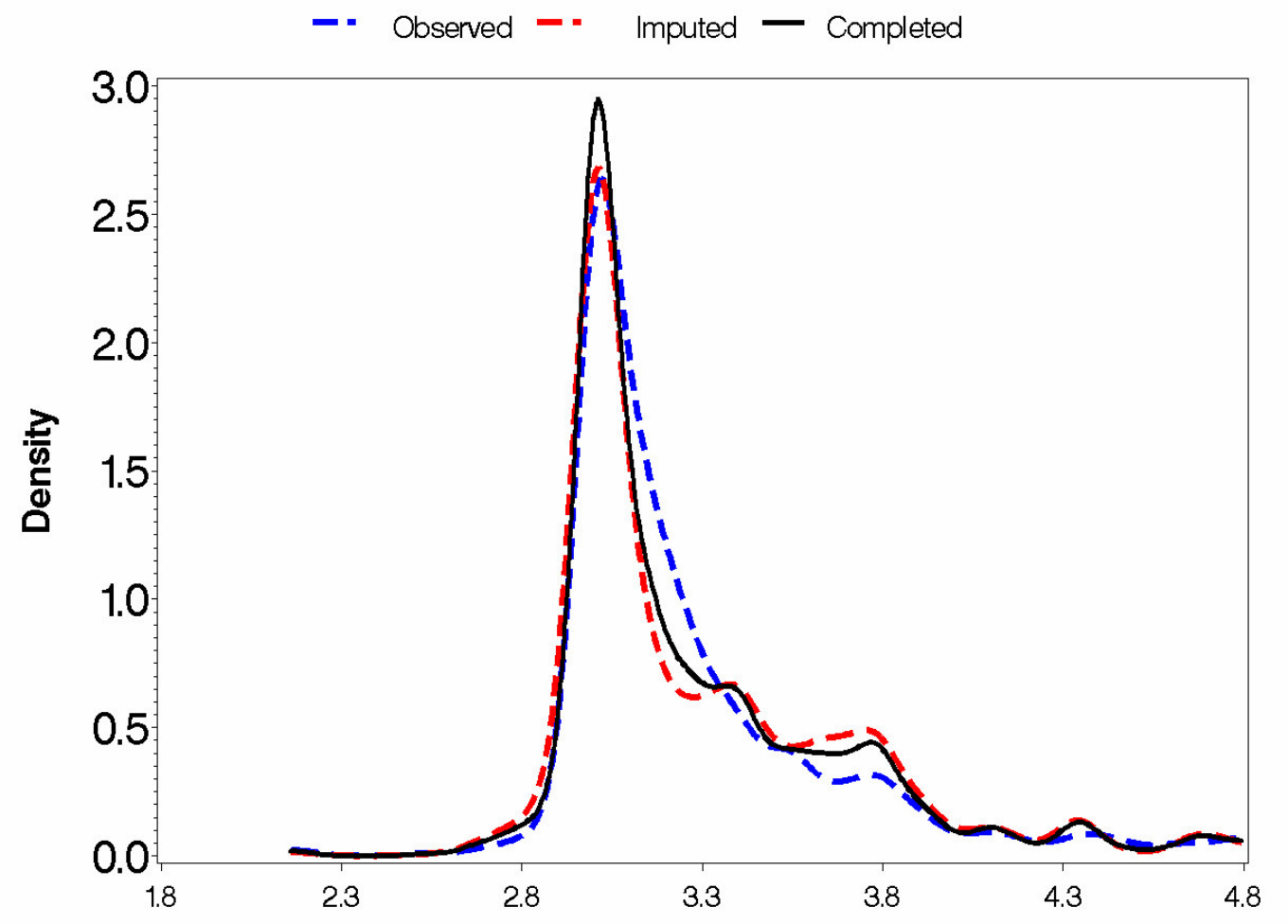


Figure 4. Histograms of observed values for age, education and income-to-poverty-ratio in the Simulation study

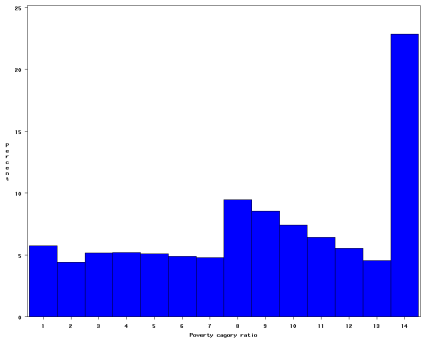
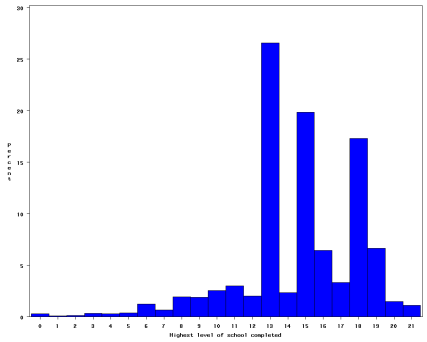
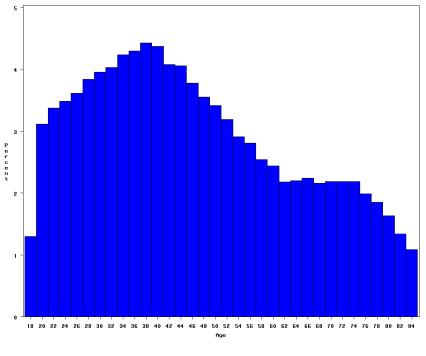
| Income-to-poverty-ratio (10% missing) | Education (21% missing) | Age (23% missing) |
|---|---|---|
|  |  |  |

Table 1. Sampling properties for means of the distributions

| Variable | | BD | CC | SRMI | ABB |
|---|---|---|---|---|---|
| Education | \|Bias\| ($10^{-2}$) | 0 | 42 | 5 | 1 |
| | MSE($10^{-2}$) | 1 | 19 | 4 | 3 |
| | Non-coverage | 7 | 87 | 12 | 7 |
| Age | \|Bias\|($10^{-2}$) | 0 | 14 | 11 | 7 |
| | MSE($10^{-2}$) | 31 | 40 | 34 | 36 |
| | Non-coverage | 5 | 7 | 6 | 6 |
| Income/poverty | \|Bias\|($10^{-2}$) | 0 | 12 | 14 | 4 |
| | MSE($10^{-2}$) | 2 | 4 | 5 | 3 |
| | Non-coverage | 6 | 14 | 10 | 6 |

Figure 5: Bias, Mean-square error and noncoverage of  based on the repeated  sampling estimates of the frequency frequency distribution of the three variables: Education, Age and Poverty-income ratio .

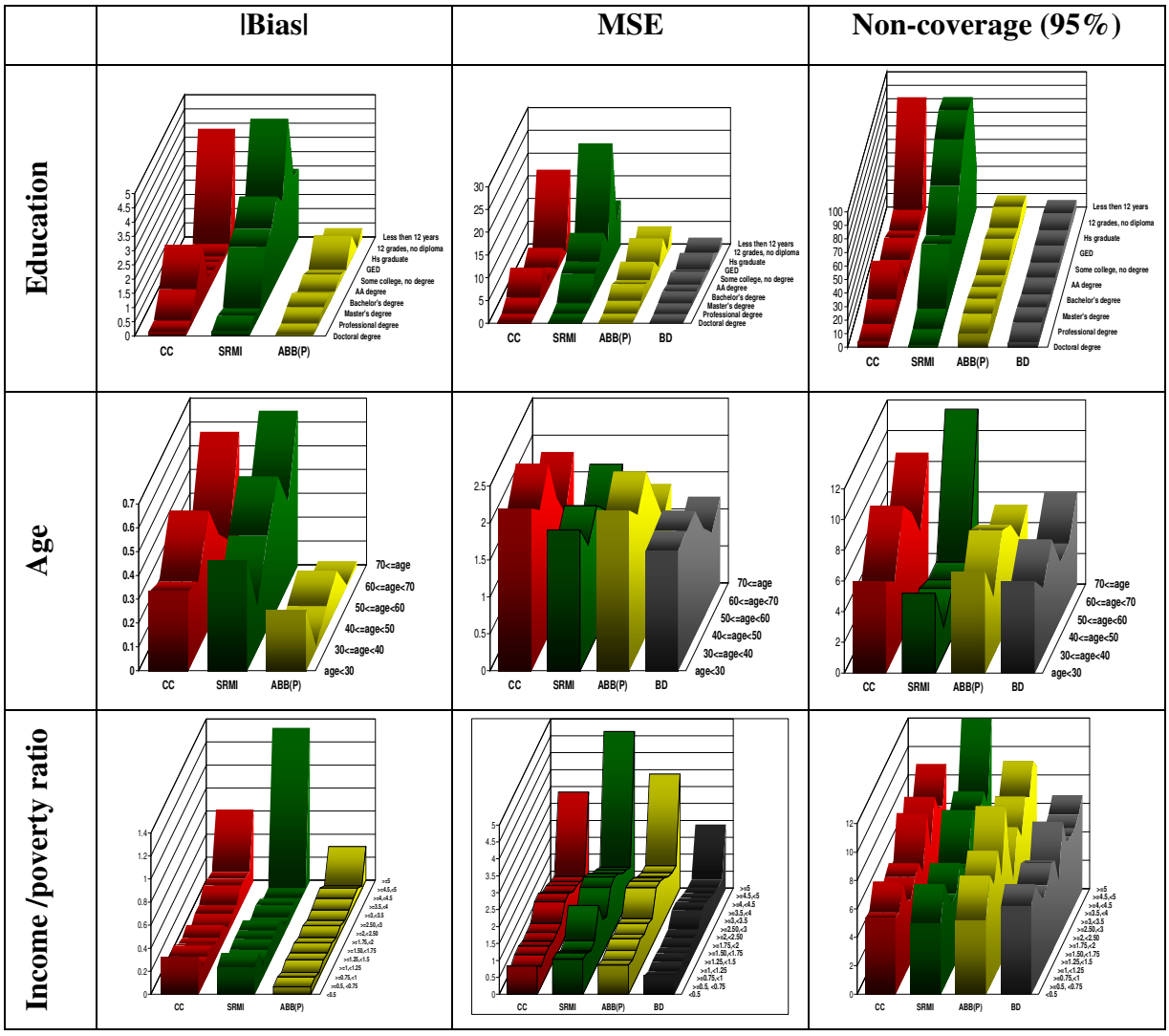


Table 2 : Bias, mean-square error and noncoverage based on the repeated sampling of Beta- coefficients in the proportional odds model

| Variable | | BD | CC | SRMI | ABB |
|---|---|---|---|---|---|
| **Age** | \|Bias\| ($10^{-3}$) | 0 | 12 | 1 | 1 |
| | MSE($10^{-5}$) | 1 | 19 | 2 | 2 |
| | Non-Coverage | 3 | 46 | 4 | 4 |
| **Gender** | Bias($10^{-2}$) | 0 | 45 | 3 | 2 |
| | MSE($10^{-3}$) | 2 | 25 | 2 | 2 |
| | Non-Coverage | 5 | 49 | 7 | 6 |
| **HS graduate** | \|Bias\|($10^{-2}$) | 0 | 7 | 29 | 1 |
| | MSE($10^{-2}$) | 3 | 9 | 11 | 5 |
| | Non-Coverage | 5 | 7 | 22 | 5 |
| **College degree** | Bias($10^{-2}$) | 0 | 1 | 29 | 3 |
| | MSE($10^{-2}$) | 5 | 13 | 16 | 9 |
| | Coverage | 5 | 4 | 13 | 6 |
| **Income/Poverty** | \|Bias\|($10^{-2}$) | 0 | 4 | 9 | 2 |
| | MSE($10^{-2}$) | 3 | 8 | 6 | 5 |
| | Non-Coverage | 5 | 5 | 5 | 3 |