A Comparison of Approximate Bayesian Bootstrap and Weighted Sequential Hot Deck for Multiple Imputation

Darryl V. Creel
RTI International

Abstract

To account for missing data, Rubin and Schenker (1986) describe a multiple imputation approach called Approximate Bayesian Bootstrap (ABB) Imputation, which is simpler and more direct computationally than Bayesian Bootstrap Imputation. Several Monte Carlo studies have investigated the properties of ABB and suggested improvements to the ABB procedure. This paper proposes an alternative to ABB for multiple imputation. We will empirically investigate the multiple imputation variance estimator for ABB, the ABB alternatives, and weighted sequential hot deck (WSHD) when the missing data mechanism is ignorable. Two different approaches to WSHD will be explored. The first approach uses WSHD to multiply impute using the same donor pool. The second approach uses a two-stage process that selects, with replacement, a new donor pool from the original set of donors and then applies WHSD to the new donor pool. The multiple imputation variance estimator will be assessed using relative bias.

Key Words: Multiple Imputation, Approximate Bayesian Bootstrap (ABB), Weighted Sequential Hot Deck (WSHD)

1. Introduction

Almost all survey data collection efforts experience some type of missing data. If there is unit nonresponse, weight adjustments are often used to minimize the potential bias of estimates from the data. If there is item nonresponse, imputation is generally used. This paper focuses on item nonresponse and multiple imputation (Rubin 1987). Specifically, it reviews Approximate Bayesian Bootstrap (ABB) and some adjustments to ABB to correct for downward biased variance estimates produced from the multiple imputation variance estimator. We extend the Monte Carlo simulation in the previous papers related to ABB to include the weighted sequential hot deck imputation methodology for multiple imputation and assess its performance using the multiple imputation variance estimator.

2. Approximate Bayesian Bootstrap

Rubin and Schenker (1986) describe a multiple imputation approach called Approximate Bayesian Bootstrap (ABB) Imputation to account for item nonresponse. To describe the ABB, let *r* be the number of respondents and *m* be the number of nonrespondents. ABB has two steps that are repeated *b* number of times, where *b* is the number of imputation will be conducted. Within each imputation class,

1. Select *r* units with replacement from the respondents to create the donor pool (Potential Donors)

2. Select *m* units with replacement from the donor pool to be actual donors

Kim (2002) investigate the multiple imputation variance estimator for ABB and found that the "relative bias is negative and its absolute value is larger for small samples and for smaller response rate" (Kim p. 474). To reduce the bias, Kim (2002) proposed an adjustment to reduce the sample size for the of the donor pool for the ABB procedure.

Parzen, Lipsitz, and Fitzmaurice (PLF) (2005) proposed a "simple correction factor applied to the multiple imputation variance estimate. The proposed correction is more easily implemented and more efficient than the procedure proposed by Kim (2002)" (PLF, p. 971).

Demirtas, Arguelles, Chung, and Hedeker (DACH) (2007) evaluated "the comparative performance of the two proposed bias-reduction techniques and their impact on precision. The results suggest that to varying degrees, bias improvements are outweighed by efficiency losses for the variance estimator. [They] argue that the original *ABB* has better small-sample properties than the modified versions in terms of integrated behavior of accuracy and precision, as measured by the root mean-square error" (DACH 2007, p. 4064).

3. Weighted Sequential Hot Deck

The weighted sequential hot deck (WSHD) (Cox 1980, Iannacchione 1982, RTI International 2008) procedure "transfers" the weighted mean of potential donors to the expected mean of recipients by including both the donor weights and the recipient weights in the procedure. The WSHD procedure is an adaptation of the *probability minimum replacement* (PMR) sequential sample selection method developed by Chromy (1979). The assignment of a selection probability to a potential donor depends both on the donor's weight and on the weights of nearby recipients. The association of donors with neighboring recipients is implemented by first sorting the file (or deck) of donors and the deck of recipients by characteristics related to the data being donated. The two decks then are interleaved by the characteristics so that donors and recipients with similar attributes are close to each other. Like other hot-deck imputation procedures, the WSHD procedure uses data from the current data set and assumes that recipients answer in a manner similar to donors with similar characteristics.

Figure 1 shows a small conceptual example of how the alignment of donors and recipients works in WSHD. Let n_r be the number of item respondents (in the example there 5 donors), w_h be the sample weight for the h^{th} respondent, n_m is the number of item nonrespondents (in the example there are 3 recipients), and s_i is the scaled weight for the i^{th} nonrespondent. The scaled weights for the recipients, i.e., nonrespondents, are on the top of the following graphic, and the weights for the donors, i.e., respondents, are on the bottom. The dashed lines are the zones set up by the scaled recipient weights. The probability of selecting a donor is based on the relative sizes of the donor weights in the zones set up by the scaled recipient weights.

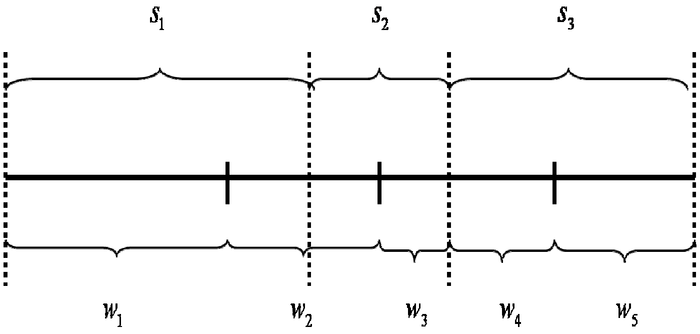

Figure 1: Alignment of Recipients and Donors for Weighted Sequential Hot Deck

The first nonrespondent has the possibility of getting the first or second respondent as the donor. The second nonrespondent has the possibility of getting the second or third respondent as the donor. The third nonrespondent has the possibility of getting the fourth or fifth respondent as the donor.

The WSHD imputation methodology was used in two ways. The first approach was to use the same donor pool, i.e., all the potential donors. That is, there was no first round of with replacement sampling to create the donor pool. The WSHD procedure simply selected the donors directly from the respondents. This was repeated *b* times. This was certainly not "proper" multiple imputation. The second approach selected the donor pool with replacement from the respondents. This is essentially the first with replacement sample in the ABB procedure. Then the WSHD procedure selected actual donors from this donor pool. This two-step process was repeated *b* times. The WSHD procedure was implemented using SUDAAN® (RTI International 2008).

4. Monte Carlo Simulation

The Monte Carlo simulation extends the Monte Carol simulation in Kim (2002), PLF (2005), and DACH (2007) by adding the two WSHD imputation approaches. The Monte Carol simulation consist of

1. Two Sample Sizes
    a. 20
    b. 100
2. Two Distributions of the Analytic Variable
    a. Normal with mean 5 and variance 1
    b. Chi-Squared with 5 degrees of freedom
3. Three Response Rates
    a. 40%

   b. 60%

   c. 80%

4. Two Values for the Number of Multiple Imputations

   a. 3

   b. 10

5. Five Imputation Methods

   a. ABB

   b. Kim – modifies donor pool size

   c. PLF – variance correction factor

   d. WSHD – same donor pool

   e. WSHDB – bootstrap to create donor pool

The five imputation methods were applied 10,000 times to each combination of the four factors, i.e., sample size, distribution, response rate, and number of multiple imputations.

5. Results

Figure 2 and **Figure 3** show the results of the Monte Carol simulation for the normal (mean = 5 and variance = 1) and Chi-Squared (5 degrees of freedom) distributions, respectively. Note that the 60% response rate is not included in the figures. The numbers plotted in the figures are the relative biases of the multiple imputation variance estimates for the different imputation methods compared to the simulation variance. The relative bias is 100 times the ratio of the difference of the variance estimate and the simulation variance divided by the simulation variance. That is, the relative bias, *relBias*, is

$$relBias = 100*((estVar - simVar)\ /simVar),$$

where *estVar* is the estimated variance and *simVar* is the simulation variance. The imputation methods are colored black for ABB, red for Kim, blue for PLF, orange for the two-step WSHD, i.e., bootstrap sample and WSHD imputation, and purple for the one-stage WSHD, i.e., only the WSHD imputation.

The results are essentially the same for both distributions. Both WSHD methods performed poorly compared to the ABB methods for the 40% response rate regardless of the sample size, and perform reasonable well but still not as good as the ABB methods for the 80% response rate regardless of the sample size.

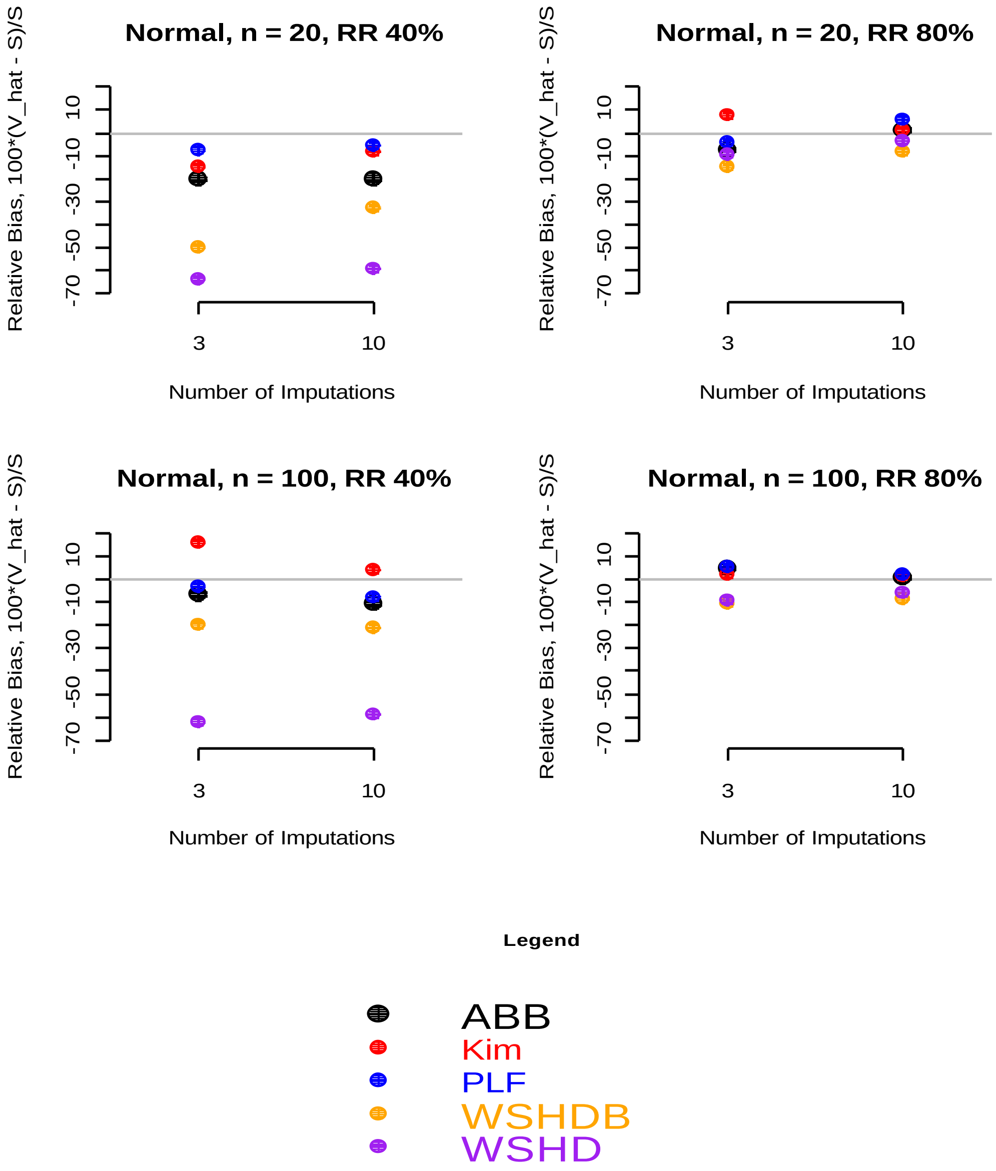

Figure 2: Plot of the Relative Variance for the Normal Distribution

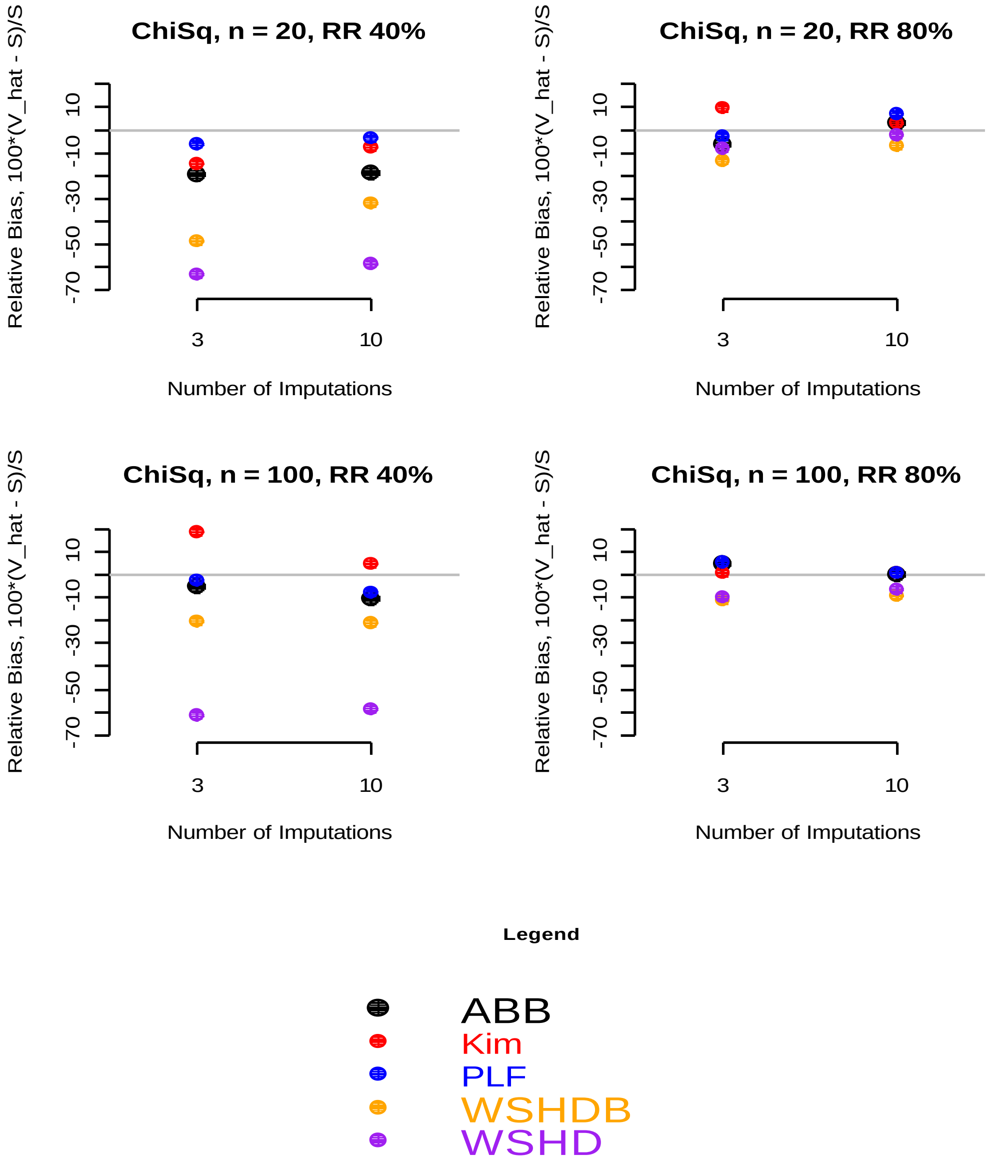

Figure 3: Plot of the Relative Variance for the Chi-squared Distribution

6. Conclusions

This Monte Carlo simulation was limited to investigating multiple imputation variance estimator for imputations from two different procedures for the WSHD. For this situation, the ABB procedures worked better. Additional empirical investigations into more complex missing data patterns and other criterion to evaluate WSHD multiple imputation and ABB need to be conducted in the future. Also, more theoretical development related to WSHD particularly when used in multiple imputation should be undertaken.

References

Chromy, J.R. (1979). “Sequential Sample Selection Methods.” *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Pp. 401-406.

Cox, Brenda (1980). “The Weighted Sequential Hot Deck Imputation Procedure.” *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Pp. 721-726.

Demirtas, Hakan, Lester Arguelles, Hwan Chung, and Donald Hedeker (2007). “On the Performance of Bias-reduction Techniques for Variance Estimation in Approximate Bayesian Bootstrap Imputation.” *Computational Statistics and Data Analysis*. Vol. 51, Pp. 4064-4068.

Iannacchione, Vincent (1982). “Weighted Sequential Hot Deck Imputation Macros.” *Seventh Annual SAS User’s Group International Conference.*

Kim, J. (2002). “A Note on Approximate Bayesian Bootstrap Imputation.” *Biometrika*. Vol. 89, No. 2, Pp. 470-477.

Parzen, Michael, Stuart Lipsitz, and Garrett Fitzmaurice (2005). “A Note on Reducing the Bias of the Approximate Bayesian Bootstrap Imputation Variance Estimator.” *Biometrika*. Vol. 92, No. 4, Pp. 971-974.

RTI International (2008). *SUDAAN Language Manual, Release 10*. Research Triangle Park: RTI International.

Rubin, Donald (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, Donald and Nathaniel Schenker (1986). “Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse.” *Journal of the American Statistical Association*. Vol. 81, No. 394, Pp. 366-374.