Post-Imputation Calibration Under Rubin's Multiple Imputation Variance Estimator[1]

Benjamin M. Reist[2] and Michael D. Larsen[3]

[2]U.S. Census Bureau and [3] The George Washington University
[2] U.S. Department of Commerce, Washington, DC 20233; Benjamin.M.Reist@census.gov
[3]6110 Executive Boulevard, Suite 750, Rockville, MD 20852; mlarsen@bsc.gwu.edu

Abstract: Multiple imputation has become one of the most popular and successful methods for dealing with missing data in statistical analyses. Multiple imputation allows one to use observed data to model relationships among variables, represent uncertainty in missing values through multiple draws from conditional distributions, and produce both point estimates and variance estimates for parameters. Variance estimates incorporate contributions to variance from both within and between completed data set analyses. Despite the advantages of multiple imputation, it has been noted that multiple imputation variance estimators can be biased. Bias is possible when, in the imputation model, survey weights are not used. Calibration weighting and its familiar forms, including raking and post-stratification, are often used in sample surveys to adjust sample estimates to match control total values and reduce variance. We explore the possibility of using calibration weighting in combination with multiple imputation to remove or reduce bias in multiple imputation variance estimation when survey weights are not used in the imputation model. Methods could apply to both sample survey and more general study design contexts.

*Key words and phrases***:** Calibration weighting; Missing data; Sample survey; Variance estimation.

1. Introduction

Multiple imputation (MI; Rubin 1978, 1987, 1996) has become one of the most popular and successful methods for dealing with missing data in statistical analyses (e.g., Barnard and Meng 1999, Klebanoff and Cole 2008, and Reiter and Raghunathan 2007). MI allows one to use observed data to model relationships among variables, represent uncertainty in missing values through multiple draws from conditional distributions, and produce both point estimates and variance estimates for parameters. Variance estimates incorporate contributions to variance from both within and between completed data set analyses. Advances have been made in computational issues of multivariate data sets and for variables exhibiting complex patterns and relationships (e.g., Raghunathan *et al.* 2001, Burgette and Reiter 2010, Azur *et al.* 2011, and Ofer and Zhou 2007)

Despite the advantages of MI for missing data in sample surveys and other studies, it has been noted that MI variance estimators can be biased to some degree. Bias has been found when survey weights are not used in the imputation model.

[1] Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or of The George Washington University.

An example is when survey weights are not used in the imputation model under informative sampling (Kott 1995). Specifically, Kott (1995) considers a situation in which there is interest on a domain that crosses strata and sampling and/or response rates vary by strata. If an unweighted model is posited for the missing data, then the fitted model may produce biased parameter estimates and predictions. If the weights, which reflect the sampling/response rates, had been used to develop the imputation model, then Kott's scenario would not be a concern. Ignoring key aspects of the sampling design and response mechanisms for an analysis and approach to missing data can lead to bias in general (Rubin 1983). Kott and Folsom (2010) comment further on the interaction of MI models and survey weights for a multi-item survey. See also Reiter, Raghunathan, and Kinney (2006) in this context.

This condition was later explored by Kim *et al*. (2006). They decompose the MI estimator ($\hat{\theta}_M$) of a parameter θ into three pieces: the complete data point estimator ($\hat{\theta}_n$), the difference between the infinite replicate MI estimator and the complete data point estimator ($\hat{\theta}_n - \hat{\theta}_\infty$), and the difference between the finite replicate and the infinite replicate MI estimators ($\hat{\theta}_\infty - \hat{\theta}_M$), where *n* is the planned sample size and *M* is the number of imputations per missing value. The bias of Rubin's MI variance estimator ($\hat{V}(\hat{\theta}_M)$) is shown to occur due to covariance between the complete data point estimator and the finite replicate MI estimator. Special cases of their presentation apply to domain estimation and linear regression models with fully observed independent variables. One troubling result from Kim *et al*. (2006) is that if survey weights are not used in the imputation model under informative sampling it is possible that the MI variance estimator can have negative bias. Prior papers had pointed to the MI variance estimator having positive bias. Unlike a point estimator, the direction of the bias for variance estimator is almost as important as the magnitude of the bias since negative bias can lead to less than the nominal confidence interval (CI) coverage. The general rule of thumb is that one unit of negative bias is equivalent to three units of positive bias for a variance estimator (Johnson and King 1987).

One of the major limitations of the prior work on bias in Rubin's MI variance estimator is that the correlation between imputation and weighting as it often occurs in practice is not considered. In many large-scale government surveys some form of calibration weighting, such as raking and post-stratification, is performed after imputation has been completed. The calibration then is dependent on the imputed values. This raises the question, if these weighting steps are performed after imputation and can take into account information external to the survey, not available for imputation, can these weighting steps reduce the bias in Rubin's MI variance estimator? In this paper we will focus on the case in which survey weights are not used in the imputation model under informative sampling.

Section 2 reviews the theoretical background used in this paper. Section 3 describes the design of the simulations used to study the question under consideration. Section 4 presents the results of the simulations for *one* of the questions under study. Finally, Section 5 provides some concluding remarks and some topics that could be considered in future research.

2. Background

Let θ be a population parameter. Let $\hat{\theta}$ be the estimator of θ. Let $V(\hat{\theta})$ be the variance of $\hat{\theta}$, and $\hat{V}(\hat{\theta})$ be the estimator of this variance. When some data that one intended to observe are missing, then one must decide what to do about the missing data when estimating θ and the variance of the estimator.

2.1 Multiple Imputation Variance Estimation

In MI (Rubin 1978, 1987, 1996), one fills in the missing data from random imputations under a (Bayesian) model. The data are completed multiple times, yielding multiple completed data sets. For filled-in data set *m*, let the estimate of θ be $\hat{\theta}_m$. Suppose there are *M* imputed data sets and estimates. The MI estimator of θ is the average of the *M* estimates: $\left(1/M\right)\sum_{m=1}^{M}\hat{\theta}_m$. The variance of this estimator can be estimated by $\hat{V} = U + \left(1 + 1/M\right)B$, where *U* is the average within analysis variance and *B* is the variance between estimates (Rubin 1987). In formulas, $U = \left(1/M\right)\sum_{m=1}^{M}\hat{V}(\hat{\theta}_m)$, where $\hat{V}(\hat{\theta}_m)$ is the estimated variance for the analysis using the *m*-th data set, and, for a scalar parameter, $B = \left(1/(M-1)\right)\sum_{m=1}^{M}(\hat{\theta}_m - \left(1/M\right)\sum_{m=1}^{M}\hat{\theta}_m)^2$.

2.2 Imputation Models

If the unweighted model for the data is $Y_i = x_i'\beta + e_i, e_i \sim^{iid} N(0,\sigma^2)$ and an *r* subscript indicates respondents and an *m* subscript indicates missing values, then a typical MI scheme (see, e.g., Kim 2004, Schenker and Welsh 1988, Rubin and Schenker 1986, Rubin 1987) is as follows. Assuming a non-informative, flat prior distribution on β ($p(\beta) \propto 1$) and that only some values of *y* are missing:

1. For each replication, *m*=1, …, *M*, draw the error variance as $\sigma_m^2|y_r \sim^{iid} SSE/\chi_{df}^2$ where SSE is the sum of squared errors from the regression of *y* on *x* for the cases with *y* observed and *df* as degrees of freedom. The degrees of freedom are influenced by the choice of the prior distribution on σ^2.

2. For each replication, *m*=1, …, *M*, draw regression coefficients as $\beta_m|y_r,\sigma_m^2 \sim^{iid} N(\hat{\beta}_r,(X_r'X_r)^{-1}\sigma_m^2)$ where $\hat{\beta}_r = (X_r'X_r)^{-1}X_r'y_r$ is the least-squares estimate of the regression coefficients using the cases with *y* observed and where X_r and y_r are the design matrix and response vector, respectively, for respondents.

3. For each replication, *m*=1, …, *M*, and unit with missing y_i, draw imputations independently of one another: $y_i^m \sim N(x_i\beta_m,\sigma_m^2)$.

Alternatively, for a weighted model, let $D_\pi = diag(\pi_1,\pi_2,\cdots,\pi_r)$ and let π_i be the inclusion probability of the i^{th} respondent. After nonresponse adjustments, π_i might be replaced by the inverse of the survey final weights. Then

1. For each replication, *m*=1, …, *M*, draw the error variance as $\sigma_m^2|y_r \sim^{iid} SSE/\chi_{df}^2$ where SSE is the sum of squared errors from the weighted regression of *y* on *x* for the cases with *y* observed and *df* as degrees of freedom. The degrees of freedom are influenced by the choice of the prior distribution on σ^2.

2. For each replication, *m*=1, …, *M*, draw regression coefficients as $\beta_m|y_r,\sigma_m^2 \sim^{iid} N(\hat{\beta}_r,(X_r'D_\pi^{-1}X_r)^{-1}\sigma_m^2)$ where $\hat{\beta}_r = (X_r'D_\pi^{-1}X_r)^{-1}X_r'D_\pi^{-1}y_r$ is the weighted least-squares estimate of the regression coefficients using the cases with *y* observed and where X_r and y_r are the design matrix and response vector, respectively, for respondents.

3. For each replication, m=1, …, M, and unit with missing y_i, draw imputations independently of one another: $y_i^m \sim N(x_i\beta_m, \sigma_m^2)$.

If the prior distribution is uniform on the (β, log σ) scale (Gelman *et al.* 2004, chapter 14; $p(\beta, \sigma^2) \propto \sigma^{-2}$) , then $df = r\text{-}p$, where p is the dimension of X. If the prior distribution on σ^2 is proportional to $(\sigma^2)^{-2}$ (Kim 2004, Meng and Zaslavsky 2002), then $df=r\text{-}p+2$.

2.3 Calibration

Calibration estimation, or calibration weighting, is a method used to incorporate axillary information based on known marginal totals into analysis of survey data from a finite population. It is used with the aim of achieving two goals. The first goal is to increase the efficiency of estimates, which can be done when the axillary information is highly predictive of the variable of interest. The second is to ensure that the estimates meet the marginal totals and thus produce consistent estimates across surveys and for known population quantities. Calibration encompasses many familiar weighting adjustments and estimators such as raking, poststratification, and generalized regression estimators. The basic theory of calibration can be found in Deville & Särndal (1992) and Deville *et al.* (1993). The basic framework for calibration estimation is to minimize the distance between base weights and new calibration weights while meeting the marginal totals. The choice of distance function and specification of control totals (i.e., the margins of which variables and interactions among variables) are what make the methods different versions of calibration estimation. Developments in the literature on survey weight calibration are not reviewed here; see Kim and Park (2010), Särndal (2007), Gelman (2007), and Zhang (2000).

2.4 Multiple Imputation with Calibration and the Use of Survey Weights

Using a superpopulation argument Kim *et. al.*(2006) showed that the Rubin's MI variance estimator is biased if the weights under informative sampling or weighting adjustments (such as cell based nonresponse adjustments where there is differential nonresponse across cells) are not used in the imputation model. They show that this bias in the variance estimator is a function of both B, the variance between estimates, and the survey weights. In the bias formula (either of formulas 5.5 or 5.8 in Kim *et al.* 2006), if the between imputation variance B goes to zero, then the bias of the MI variance estimator goes to zero as well. That is, if the between imputation variance B is zero, then the amount of bias in the MI variance estimator also should be zero.

The interpretation of the bias formula from Kim *et al.* (2006) in the previous paragraph has an implication for the potential of calibration estimation to reduce bias in the MI formula. If the outcome variable that is being analyzed is highly correlated with auxiliary variables used in calibration, then all estimates of the parameter of interest should be nearly the same across sets of imputations. In that situation, the value of B should be small and the variance of the MI variance estimator should be small as well. If the variables that are used in the calibration are reasonably well correlated with the desired outcome variable, then B will be reduced in comparison to what it would have been without calibration. The more similar calibration to the same control totals for each set of imputation can make the estimates, the smaller the value of B should be, thus reducing the bias in the MI variance estimator. The simulation in the next section examines this idea.

3. Simulation

The following simulation was used to evaluate the potential bias in the MI variance estimators when survey weights are not used in the imputation model, under informative sampling. This simulation evaluated the following three imputation models specifications:

1. Without sample weights (NW)
2. With sampling weights as predictors (WP)
3. Weighted model (WM)

This simulation was run in R (R Development Core Team, 2008) and are similar in nature to but different in detail from those presented in Kim (2004). All of the simulations are based on a common finite population with N = 100,000 members. The simulation design used was a 3 x 2 x 2 factorial design with the following three factors:

1. Factor A, method of imputation – NW, WP, WM
2. Factor B, response rate – 0.8, 0.6
3. Factor C, sample size – 200, 400

For each factor combination, 50,000 simulations (L) were performed.

3.1 Finite Population

A common finite population was used for all three simulations. The finite population was generated by taking 100,000 independent draws out of a normal distribution with μ = 10 and variance of $\sigma^2 = 25/3$ (X).

$$Y_i = 2 + 4X_i + e_i$$

where e_i are drawn independently from a standard normal distribution. This model specification causes X and Y to be highly correlated (i.e $\rho \approx 0.99$). This follows the basic setup of (Kim 2004).

Additionally, three variables were generated to be used for creating strata, raking and the sample design as follows:

$$U_i^{\rho} = Y_i + \epsilon_i$$

where ϵ_i was drawn independently from a normal distribution with μ = 10 and σ^2 in a manner such that the correlation(Y, U_i^{ρ}) = ρ for ρ = 0.8. Specifically the variance of Y is v= 16(25/3)+1, the variance of U is v+ σ^2, and the covariance of Y and U is also v. The correlation of Y and U then is $v/\sqrt{v(v+\sigma^2)}$, and one can solve for ρ.

3.2 Sampling

Stratified simple random sampling was used to create an informative sampling scheme. Four strata were created based on the quartiles of $U_i^{0.8}$, each stratum with 25,000 units. Sample was allocated between strata with the upper quartile receiving 50% of the sample, 30% of the sample in the next quartile, 15% in the next, and finally 5% in the bottom quartile. As Kott (1995) notes, this can be viewed as either unequal sampling rates or unequal response rates across four nonresponse adjustment cells where the inverse response rate is used to adjust the weights. Under

the latter setting, the nonresponse adjustment cells would be optimal since the cells would be good predictors of *Y* and the propensity to respond thus reducing both nonresponse bias and variance (e.g., Little and Vartivarian 2005 and West 2009).

3.3 Missing Data Mechanisms

Missing values of *Y* for each sample were generated by taking a simple random sample of size one minus response rate times the sample size. This is a uniform response mechanism that insures constant sample size for each simulate. This is the missing completely at random (MCAR) response mechanism, which also was used by Kim (2004).

3.4 Imputation Algorithms

For this simulation, two imputation models were compared: the standard linear-model framework studied by Schenker and Welsh (1988) and the SOUP modification to this model, which appears in Kim (2004) and is based on Meng (1994) and Meng and Zaslavsky (2004).

For this simulation, simple random without replacement sampling was used to ensure that the sampling was non-informative. Imputation is performed assuming a classic linear model framework:

$$y_i = \beta \vec{x}_i + e_i$$

$$e_i \sim N(0, \sigma^2).$$

The standard Bayesian approach to a classical linear unweighted model based only on the respondents is to assume that

$$\beta_r | (y, \sigma^2) \sim N(\hat{\beta}_r, \hat{V}_{\beta_r}\sigma^2),$$

where

$$\begin{aligned} \hat{\beta}_r &= (X_r' X_r)^{-1} X_r' y_r \\ \hat{V}_{\beta_r} &= (X_r' X_r)^{-1}. \end{aligned}$$

Additionally,

$$\sigma^2 | y_r \sim Inv\chi^2(df, s_r^2)$$

$$s_r^2 = (r-2)^{-1} y_r' [I - X_r (X_r' X_r)^{-1} X_r'] y_r.$$

For more details on the Bayesian approach to classical linear models see chapter 14 of Gelman *et al*. (2004).

For the classical weighted model, $\hat{\beta}_r$, $\hat{V}_{\beta_r}$, and s_r^2 are replaced by the following formulas:

$$\begin{aligned} \hat{\beta}_{r,\pi} &= (X_r' D_\pi^{-1} X_r)^{-1} X_r' D_\pi^{-1} y_r \\ \hat{V}_{\beta_{r,\pi}} &= (X_r' D_\pi^{-1} X_r)^{-1} \\ s_{r,\pi}^2 &= (r-2)^{-1} y_r' [D_\pi^{-1} - D_\pi^{-1} X_r (X_r' D_\pi^{-1} X_r)^{-1} X_r' D_\pi^{-1}] y_r \end{aligned}$$

These estimates were used in the WM.

To make this model and procedure operational in a MI context the following algorithm is used independently for each replicate $k = 1,\ldots, M$:

1. Draw

$$\sigma^2_{(k)}|y_r \sim Inv\chi^2(df, s_r^2)$$

2. Draw

$$\beta_{(k)}|\left(y, \sigma^2_{(k)}\right) \sim N(\hat{\beta}_r, \hat{V}_{\beta_r}\sigma^2_{(k)})$$

3. Then for each missing y_j, draw

$$e_{j(k)}|\left(\beta_{(k)}, \sigma^2_{(k)}\right) \sim N(0, \sigma^2_{(k)})$$

4. Finally, impute for y_j for the k^{th} implicate as

$$y_{j(k)} = \beta_{(k)}\vec{x}_j + e_{j(k)}.$$

The difference between the methods proposed by Schenker and Welsh (1988) and Kim (2004), from now on called the SW method and the SOUP method, respectively, lie in the choice the df used in the prior distribution on σ^2. The SW method uses $df = r\text{-}p$, where r is the number of respondents and p is the number of parameters that are being estimated in the model. In this case, $p = 2$. The SOUP method uses $df = r\text{-}p\text{-}2$.

3.5 Imputation Models

Three models were evaluated in this simulation—one that does not incorporate any design information and two that incorporate the design weights differently.

1. Without sample weights (NW)

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$e_i \sim N(0, \sigma^2).$$

2. With sampling weights as predictors (WP)

$$y_i = \beta_0 + \beta_1 x_i \ + \beta_2 w_i + e_i$$

$$e_i \sim N(0, \sigma^2),$$

where w_i is the weight for the i^{th} unit. In this case, $w_i = \frac{1}{\pi_i}$, the inverse of the inclusion probability.

3. Weighted model (WM)

$$y_i = \beta_{\pi,0} + \beta_{\pi,1} x_i + e_i$$

$$e_i \sim N(0, \sigma^2).$$

3.6 Calibration Method

For the NW model, three estimators of the mean were calculated.to evaluate the effect of post-imputation raking schemes on the bias and coverage properties of MI variance estimators:

1. NW model without raking (NW),
2. NW model raking to the $U_i^{0.8}$ variable and the marginal total $\sum U_i^{0.8}$ (NW-RT)
3. NW model raking to stratum totals (NW-RST).

Estimators 2 through 4 are typical weighting adjustments used in practice to incorporate design information into the weighting process to reduce variance. Raking was performed using the *calib* function from the sampling package in R (Tillé & Matei 2008).

3.7 Variances and Degrees of Freedom

The within variance for each imputed data set was calculated using the traditional jackknife method. The between variance and the combined variance were calculated using the method described in Section 2.1. For CI calculation, a *t*-distribution was assumed, with df calculated using the method proposed by Barnard and Rubin (1999)—a method of calculating degrees of freedom for MI analyses that is robust for small sample sizes.

3.8 Evaluation Criterion

Three evaluation criterion were used to evaluate the bias of the variance estimate and coverage properties of each variance estimator. One was the relative bias

$$\frac{E_L(\hat{V}) - Var_L(\hat{\theta})}{Var_L(\hat{\theta})},$$

where $E_L(\hat{V})$ is the Monte Carlo mean of the variance estimator for $\hat{\theta}$ over L simulations and $Var_L(\hat{\theta})$ is the estimated Monte Carlo variance over L simulations. The second evaluation criterion was the CI coverage of the finite population value of θ for the 95% *t*-distribution based CI over the L simulations using the estimated variance. The final evaluation criterion was mean lengths of the 95% *t*-distribution based CI over the L simulations.

4. Results

Results are presented in Table 1. Methods are summarized by coverage rate of nominal 95% CIs, relative bias of variance estimates, and mean length of CIs. Coverage should be 95% or above for a CI procedure to be judged valid, and coverage rates closer to 95% and still above 95% are desirable. Relative bias should be small for the point estimate of variance to be considered high quality. Other things being equal, the mean length of CIs should be small. If a method has short CIs but fails to achieve 95% or greater coverage, then it is inferior to a method with wider CIs but adequate coverage.

The Monte Carlo simulation variance of these estimates was on the order of $O(n^{-1})$. Each draw was taken from the same finite population and the three methods applied to each sample data set. The estimators should be positively correlated between simulations. As a result, when considering the difference between coverage rate, bias, and mean length of CI estimates for different methods, the standard error of the difference should be smaller than the Monte Carlo standard error for the individual estimation methods. Thus, the variance of the difference between Monte Carlo estimates might be slightly lower than the variance of individual Monte Carlo simulation estimates. Overall, this means that small differences like those found in coverage, relative bias, and mean length of CI are likely not due to simulation error alone.

Using the approximate simulation variance, we see that the WM produces consistently large coverage rates (over 99%). When compared to the other three models the WM model, produces larger coverage rates, which are statistically significant at the 95% confidence level. Since the variance estimates are positively biased and mean length of intervals is large, the WM is acceptable from a CI coverage perspective, but it is not very desirable due to the imprecise inferential statements it generates. The other three methods that have been implemented produce coverage rates close to the desired 95%. Although, none of these methods consistently outperforms the others in terms of coverage rates and relative biases. Additionally, there does not seem to be any appreciable effect caused by raking back to design variable totals except for statistically significant reductions in the mean length of the CIs.

Table 1: Actual Coverage Rates (CR) of Nominal 95% Confidence Intervals, Relative Bias (RB) of Variance Estimators, and Mean Length (ML) of 95% Confidence Intervals based on *L=50,000* replicates. *n* is the sample size, *r/n* is the response rate.

N	r/n	Measure	Model				
			Weighted model (WM)	Weights as predictors (WP)	Not weighted w/o raking (NW)	Not weighted, rake to totals (NW-RT)	Not weighted, rake to stratum (NW-RST)
200	80%	CR	99.9%	95.0%	94.9%	95.1%	95.1%
		RB	2.041	0.033	0.028	0.027	0.034
		ML	5.521	2.544	2.525	2.318	2.452
	60%	CR	100.0%	94.8%	95.1%	95.2%	95.3%
		RB	3.223	0.015	0.039	0.042	0.044
		ML	7.67	2.550	2.557	2.333	2.466
400	80%	CR	99.5%	95.0%	95.1%	95.1%	95.1%
		RB	1.165	0.014	0.029	0.017	0.019
		ML	2.947	1.782	1.784	1.626	1.715
	60%	CR	99.9%	95.1%	95.0%	95.0%	95.1%
		RB	1.967	0.019	0.019	0.017	0.016
		ML	3.870	1.788	1.790	1.634	1.724

5. Discussion

MI is a popular method for dealing with missing data in statistical analyses that produces both point estimates and variance estimates for parameters. It has been noted that MI variance estimators can be biased when survey weights are not used. In this paper we explore the possibility of using post stratification, a form of calibration weighting, in combination with MI to reduce bias in MI variance estimation when survey weights are not used in the imputation model. Three methods were evaluated in a simulation.

Using a WM produced undesirable results (i.e. coverage rates greater than 99%). This is because design consistent covariance estimates were not used in the model thus not reflecting the variance reduction induced by the stratification. Both using the NW model and using the WP model produce comparable results across sample sizes and response rates. This might be caused by the fact that X and Y are so highly correlated. Thus, design information may not be needed at all in the imputation model. Finally, raking only had the effect of lowering the mean lengths of CIs. This implies that calibration could have minimal effects on coverage rates and bias.

This simulation, however, is both optimal and unrealistic. The scenario used in simulations here assumes that X and Y are correlated almost perfectly and the missingness mechanism is MCAR, both of which are not seen in practice. The simulation was chosen here to be comparable to methods employed in Kim (2004).

Future work can explore a number of extensions. First, it would be useful to explore missing at random (MAR) mechanisms, such as when response rates vary by strata or by domain. Second, it would be interesting to implement simulations with covariates that are not as correlated as the X and Y in this simulation. Third, it would be important to examine nonlinear relationships among variables. Fourth, it would be desirable to study properties in higher dimension examples. Finally, as noted above the weighted model does not use design consistent estimates of the variance-covariance matrix as seen in Binder (1983), chapter 6 Fuller (2009), and chapter 5 section 10 of Särndal *et al* (1992). This would be of interest.

6. References

AZUR, M. J., STUART, E. A., FRANGAGKIS, C., & LEAF, P.J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* **20**, 40-49.

BARNARD, J. & MENG, X.-L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research* **8** 17-36.

BARNARD, J. & RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948-955.

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279-292.

BURGETTE, L. F., & REITER, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology* **172**, 1070-1076.

DEVILLE, J.-C., & SÄRNDAL, C.-E. (1992). Calibration estimation in survey sampling. *Journal of the American Statistical Association* **87**, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., & SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013-1020.

FULLER, W. A. (2009). *Sampling Statistics*, Hoboken: John Wiley & Sons, Inc.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 153-164.

GELMAN, A., CARLIN, J. B., STERN, H. S., & RUBIN, D. B. (2004). *Bayesian Data Analysis* 2nd ed., Boca Raton: Chapman & Hall/CRC.

JOHNSON, E. G., & KING, B. F. (1987). Generalized variance functions for a complex sample survey, *Journal of Official Statistics* **38**, 235-250.

KIM, J. K., (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics* **32**, 766-783.

KIM, J. K., BRICK, J. M., FULLER, W. A., & KALTON, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B* **68**, 509-521.

KIM, J. K. & PARK, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, **78**, 21-39.

KLEBANOFF, M. A., & COLE, S. R. (2008). Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology* **168**, 355-357.

KOTT, P. S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research*, American Statistical Association, Alexandria, VA, 380-383.

KOTT, P. S., & FOLSOM, R. E. (2010). Weights, double protection, and multiple imputation. *Proceedings of the Section on Survey Research*, American Statistical Association, Alexandria, VA, 431-440.

LITTLE, R. J., & VARTIVARIAN, S. (2005). Dose weighting for nonresponse increase the variance of survey means? *Survey Methodology* **31**, 161-168.

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538-558.

MENG, X.-L., & ZASLAVSKY, A. M. (2002). Single observation unbiased priors. *The Annals of Statistics* **30**, 1345-1375.

OFER, H. & ZHOU, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* **26**, 3057-3077.

R DEVELOPMENT CORE TEAM. (2008). R: a language and environment for statistical computing **V2.7.2.** Vienna: R Foundation for Statistical Computing.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J., & SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85-95.

REITER, J. P., & RAGHUNATHAN, T. E. (2007). Multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462-1471.

REITER, J. P., RAGHUNATHAN, T. E., & KINNEY, S. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology* **32**,143 - 150.

RUBIN, D. B. (1978). Multiple imputations in sample surveys – a phenomenological response to missing data. *Proceedings of the Survey Research Methods Section*, 20-28.

RUBIN, D. B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys – probabilities of selection and their role for Bayesian modeling in sample-surveys – comment. *Journal of the American Statistical Association*, **78**, 803-805.

RUBIN, D. B. (1987). *Multiple Imputation for Survey Nonresponse*, New York: John Wiley & Sons Inc.

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473-489.

RUBIN, D. B. & SCHENKER, N. (1986). Multiple imputation for interval estimation from sample random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366-374.

SÄRNDAL, C.–E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**, 99-119.

SÄRNDAL, C.–E, SWENSSON, B., & WRETMAN J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SCHENKER, N. & WELSH, A. H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics,* **16**, 1550-1566.

TILLÉ, Y. & MATEI, A. (2008). Sampling: survey sampling. *R package* V**2.0**.

WEST , B.T. (2009). A simulation of alternative weighting class adjustments for nonresponse when estimating a population mean from complex sample survey data. *Proceedings of the Section on Survey Research*, American Statistical Association, Alexandria, VA, 4920-4933.

ZHANG, L. C. (2000). Post-stratification and calibration – a synthesis. *American Statistician,* **54,** 153-164.