Can Post-Imputation Calibration Mitigate Possible Bias in 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 estimated parameters. Variance estimates incorporate contributions to variance from both within and between completed data set analyses. Advances have been made on computational issues for highly multivariate data sets and for variables exhibiting complex patterns and relationships. Despite the advantages of multiple imputation for missing data in public use release of data from sample surveys and other studies, it has been noted that multiple imputation variance estimators can be biased to some degree. Bias has been found to be possible when key subpopulations are ignored in the imputation model, survey weights are not used in the imputation model, and in the case of linear estimators prior distributions for model parameters other than single observation unbiased priors (SOUPs) are used in the imputation model. 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. We explore possibilities for using calibration weighting in combination with multiple imputation when there are missing data in a survey from a finite population. In particular, we examine the potential for calibration to remove or reduce bias in multiple imputation variance estimation when a SOUP is not used. Methods could apply to both sample survey and more general study design contexts.

Key words and phrases: Calibration weighting; Single observation unbiased priors; 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 estimated parameters. Variance estimates incorporate contributions to variance from both within and between completed data set analyses. Advances have been made on computational issues for highly multivariate data sets and for variables exhibiting complex patterns and relationship (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 public use release of data from sample surveys and other studies, it has been noted that MI variance estimators can be biased to some degree. Bias has been found to be possible when key subpopulations are ignored in the imputation model, survey weights are not used in the imputation model, and in the case of linear estimators prior distributions for model parameters other than single observation unbiased priors (SOUPs) are 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.

An instance of biased variance estimation that appears in the literature concerns estimating a domain (or subgroup) mean and total when the domain is ignored by the imputation model (Fay 1992; see also Fay 1993 and 1994). Ignoring key aspects of the data generally causes bias in any analysis or approach to handling missing data. When a domain is ignored, imputations can have too much variability, because the variability in imputations can reflect both observed variability within the group plus variability due to group differences. If the imputer (the person doing the imputations) and the analyst use models that both incorporate domain effects, then this criticism has much less strength. In a large-scale complex survey, however, imputation might be implemented by a statistical agency before data are released. Subsequently, analysts conduct many analyses without intimate knowledge of the imputation model, including which domains have been modeled. Response to this issue can be found in Meng (1994), Rubin (1996), and, briefly, in Kott (1995).

The second example occurs 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 a model is posited for the missing data that gives equal weight to each respondent in the domain, then there is a disconnection between sampling/response rates and the use of individual units when estimating the model. 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 *et al.* (2006) in this context.

These two conditions (domains and survey weights) were explored later by Kim *et al*. (2006). The study in Kim *et al*. (2006) extends the work of Kott (1995) on weighting and Fay (1992, 1993) on domains. The paper examines estimators that are linear combinations of the sampled data. The authors consider a joint distribution for superpopulation, sampling, response, and imputation models. The MI estimator ($\hat{\theta}_M$) of a parameter θ is decomposed 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 ($V(\hat{\theta}_M)$) 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.

The third situation in which the MI variance estimator can have bias is when certain prior distribution assumptions are made. Kim (2004) demonstrated bias in a finite population context for Rubin's MI variance estimator for an estimator that is a linear combination of sample values. In order to implement Bayesian multiple imputation, one specifies a prior distribution on unknown parameters. Under the model, which includes a data distribution and a prior distribution, one generates random values as imputations for missing data. Under Kim's (2004) scenario, when a prior distribution other than a single observation unbiased prior (SOUP; Meng and Zaslavsky 2002) is used in the imputation model, Rubin's MI variance estimator is biased. The bias is larger when there is a small sample size in the study.

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 preformed 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 the proposed research we set out to explore this question for each of the three conditions in which biased variance estimation has been noted to occur. In this paper we will focuses on the selection of priors.

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 multiple imputation (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, $B = \left({}^{1}/_{(M-1)}\right)\sum_{m=1}^{M}(\hat{\theta}_m - \left({}^{1}/_{M}\right)\sum_{m=1}^{M}\hat{\theta}_m)^2$.

If the 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 multiple imputation scheme (see, e.g., Kim 2004, Schenker and Welsh 1988, Rubin and Schenker 1986, Rubin 1987) is as follows. We assume 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* is a degrees of freedom. The degrees of freedom is 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 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-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-p+2*.

2.2 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 estimators, 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 control totals, such as known population totals, and thus produce estimates that have the same value 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 change in the weights while meeting the control totals under a selected distance function. 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.

2.3 Multiple Imputation with Calibration and Choice of Prior Distribution

Kim (2004) and Schenker and Welsh (1988) considered MI for missing data in the outcome variable y in a homoscedastic regression model with independent errors. Schenker and Welsh (1988) showed that, as the sample size increases, the MI point and variance estimator are consistent for their desired quantities. Kim (2004) showed that in the finite population setting that the MI variance estimator can have bias. This occurs in the case of simple regression when $df=r-2$, which corresponds to a prior distribution uniform on the (β, log σ) scale (Gelman *et al.*, 2004). If the sample size is large, then the bias gets small. The bias goes away in the simple regression scenario when $df = r$, corresponding to a prior distribution on σ^2 proportional to $(\sigma^2)^{-2}$. Meng and Zaslavsky (2002) called this a single observation unbiased prior (SOUP) distribution. It should be noted that even if a sample survey has a large total sample size, one often desires to make inferences within strata or groups of strata and for small domain groups. Thus, one cannot always rely on large sample sizes to address bias in Rubin's multiple imputation variance estimator.

The bias in the Rubin's multiple imputation variance estimator, when using the SW prior, is a function of both the within (W) and the between (B) variance (Kim 2004). Calibration, if the axillary variable is significantly correlated with the variable of interest, reduces W by incorporating axillary information and reduces B by pulling the estimates in each imputed data set closer together.

3. Simulation

Three separate simulations are being run to evaluate each of the sources of potential bias in the multiple imputation variance estimators described earlier in this paper. Here results are reported for the issue of the choice of the prior distribution. These simulations were all 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 2 x 3 x 4 factorial design with the following three factors:

1. factor A, method of imputation - SW or SOUP
2. factor B, response rate – 0.8, 0.6, 0.4
3. factor C, sample size – 20, 40, 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 σ^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, two variables were generated to be used for raking 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 correlation(Y, U_i^{ρ}) = ρ for ρ = (0.6, 0.9). 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 Missing Data Mechanisms

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

3.3 Choice of Prior Distribution and Imputation Algorithm

For this simulation two imputation models are compared: the standard linear-model framework studied by Schenker and Welsh (1988) and the single observation unbiased priors (SOUPs) modification to this model which appears in Kim (2004) and is based on Meng (1994) and Meng and Zalavsky (2004).

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

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

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

The standard Bayesian approach to a classical linear 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).

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

1. Draw

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

2. Draw

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

3. Then for each missing y_j draw

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

4. Finally impute for y_j for the kth implicate as

$$y_{j(k)} = \beta_{0(k)} + \beta_{1(k)} x_j + e_{j(k)}.$$

The difference between the method proposed by Schenker and Welsh (1988) and Kim (2004), from now on called the SW method and the SOUP method, respectively, lies in the choice the degrees of freedom 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 $df = r\text{-}p\text{-}2$.

3.4 Raking

The weights where all adjusted by raking to evaluate the effect of post-imputation weighting adjustments on the bias and coverage properties of multiple imputation variance estimators. Raking was performed using the calib function from the sampling package in R (Tillé & Matei 2008). Raking was performed by using the U_i^{ρ} variable and the marginal total $\sum U_i^{\rho}$.

3.5 Variances and Degrees of Freedom

Within variance for each imputed data set was calculated using a traditional jackknife method. The between variance estimate and the combined variance estimate were calculated using the method laid out in Section 2.1. For calculation of confidence intervals, a *t*-distribution was assumed with degrees of freedom, calculated using the method proposed by Barnard and Rubin (1999), which is a method of calculating degrees of freedom for multiple imputation analyses that is more robust for small sample sizes.

3.5 Evaluation Criterion

Two evaluation criterion where used to evaluate the bias of the variance estimator and coverage properties of each variance estimator. The first is the relative bias

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

where $E_L(\hat{V})$ is the 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. Values of $E_L(\hat{V})$ and $Var_L(\hat{\theta})$ are found in the appendix. The second evaluation criterion is the coverage of the finite population value of θ for the 95% *t*-distribution based confidence interval over the L simulations.

4. Results

There is not much variation in coverage rates or relative biases of variance estimators between the multiple imputation methods with the SOUP prior distributions and the multiple imputation methods with the three versions using the SW prior distribution. Since 50,000 simulations were performed, the Monte Carlo simulation variance of these estimates are on the order of $O(n^{-1})$. Each draw was taken from the same finite population and the four methods applied to each sample data set. There should be a positive correlation between results that is induced by using the same sample to produce four estimates. As a result, when considering the difference between coverage rate or bias estimates for two 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 and relative bias are likely not due to simulation error alone.

Table 1: Actual Coverage Rates of Nominal 95% Confidence Intervals and Relative Bias of Variance Estimators based on *L=50,000* replicates. *n* is the sample size, *r/n* is the response rate, SOUP is single observation unbiased prior and SW is the Schenker-Welsh prior distribution. In the SW scenario, the three options are no calibration (No-Cal) or calibration (raking) with an auxiliary variable with correlation 0.45 or 0.90.

n	r/n	95 % Coverage				Relative Bias			
		SOUP	SW			SOUP	SW		
			No-Cal	0.45	0.90		No-Cal	0.45	0.90
20	0.8	95.1	95.2	95.2	95.4	-0.0013	0.0056	0.0341	0.0158
	0.6	95.0	95.1	95.2	95.7	-0.0056	0.0020	0.0310	0.0299
	0.4	95.1	95.0	95.3	96.2	-0.0005	0.0059	0.0291	0.0563
40	0.8	95.1	95.0	94.8	95.2	-0.0069	-0.0070	-0.0005	0.0088
	0.6	95.1	94.9	95.1	95.2	0.0018	-0.0055	0.0085	0.0093
	0.4	95.1	95.1	95.0	95.6	-0.0090	-0.0009	0.0094	0.0284
200	0.8	95.0	95.0	95.1	95.1	0.0053	0.0082	0.0073	0.0101
	0.6	95.0	94.9	95.1	95.3	-0.0046	-0.0123	0.0053	0.0163
	0.4	94.8	94.8	95.1	95.4	-0.0179	-0.0137	0.0051	0.0370
400	0.8	95.1	94.9	95.0	95.3	0.0046	-0.0013	-0.0016	0.0241
	0.6	95.0	94.9	95.0	95.2	0.0064	0.0025	0.0065	0.0154
	0.4	95.1	94.9	95.1	95.5	-0.0063	-0.0069	-0.0021	0.0362

For larger values of *n* the SW with no calibration tends to cover well. There is only slight over coverage in some cases. This does not seem to be a problem for the SOUP model or either of the calibrated SW coverage rates. In coverage, the SW calibrated estimator seems in general to far better than the no calibration SW method but this comes at a steep cost. The calibration to a variable with low correlation to the variable of interest considerably increases the variance, which can be seen in Table 2. The calibrated SW to a variable with 0.90 correlation always covers more than the nominal rate and always has a relative bias greater than zero, which is seen as preferable, although none of the methods has very large relative biases for the variance. As expected, the benefit of calibrating to such a highly correlated variable is that the variance shrinks considerably which also can be found in Table 2. Overall the simulation shows that the choice of prior does not seem to effect bias or coverage in Rubin's variance estimator and that the addition of calibration does not change these results appreciably.

5. Discussion

The framework used in Kim's (2004) simulation is both the optimal and unrealistic. That scenario assumes that *X* and *Y* are correlated almost perfectly and the missingness mechanism is MCAR, both of which are not seen in practice. Additionally, having a variable that has 0.90 correlation with the outcome variable to rake to is not very likely unless the variable is a size measurement taken repeated over time in a longitudinal survey or from some high quality administrative source. The other two sources of bias are most likely more interesting and need similar simulations to be done to truly understand bias and coverage issues in Rubin's multiple imputation variance estimator. It would also be useful to explore missing at random mechanisms, covariates that are not as correlated as *X* and *Y* are in this simulation and multiple covariates and control totals.

Table 2: Empirical Variances and Mean of Estimated Variances based on *L=50,000* replicates. *n* is the sample size, *r/n* is the response rate, SOUP is single observation unbiased prior and SW is the Schenker-Welsh prior distribution. In the SW scenario, the three options are no calibration (No-Cal) or calibration (raking) with an auxiliary variable with correlation 0.45 or 0.90.

n	r/n	Mean Estimated Variance				Empirical Variance			
		SOUP	SW			SOUP	SW		
			No-Cal	0.45	0.90		No-Cal	0.45	0.90
20	0.8	6.71	6.74	17.54	1.58	6.72	6.71	16.96	1.56
	0.6	6.76	6.76	17.58	1.64	6.80	6.75	17.06	1.60
	0.4	6.82	6.85	17.71	1.75	6.83	6.81	17.21	1.66
40	0.8	3.36	3.36	8.45	0.78	3.38	3.38	8.46	0.78
	0.6	3.37	3.38	8.49	0.80	3.37	3.39	8.42	0.80
	0.4	3.40	3.39	8.50	0.84	3.43	3.40	8.43	0.82
200	0.8	0.67	0.67	1.65	0.16	0.67	0.67	1.64	0.15
	0.6	0.67	0.67	1.66	0.16	0.68	0.68	1.65	0.16
	0.4	0.68	0.68	1.66	0.16	0.69	0.69	1.66	0.16
400	0.8	0.34	0.34	0.83	0.08	0.33	0.34	0.83	0.08
	0.6	0.34	0.34	0.83	0.08	0.33	0.34	0.82	0.08
	0.4	0.34	0.34	0.83	0.08	0.34	0.34	0.83	0.08

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.

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., & OLIVIER, S. (1993) Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013-1020.

FAY, R. E. (1992) When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research*, American Statistical Association, Alexandria, VA, 227-232.

FAY, R. E. (1993) Valid inference from imputed survey data. *Proceedings of the Section on Survey Research*, American Statistical Association, Alexandria, VA, 41-48.

FAY, R. E. (1994) Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490-498.

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

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.

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.

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.

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**.