DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 963–978, http://dx.doi.org/10.1515/JOS-2017-0045

How to Obtain Valid Inference under Unit Nonresponse?

Laura Boeschoten[1], *Gerko Vink*[2], *and Joop J.C.M. Hox*[2]

Weighting methods are commonly used in situations of unit nonresponse with linked register data. However, several arguments in terms of valid inference and practical usability can be made against the use of weighting methods in these situations. Imputation methods such as sample and mass imputation may be suitable alternatives, as they lead to valid inference in situations of item nonresponse and have some practical advantages. In a simulation study, sample and mass imputation were compared to traditional weighting when dealing with unit nonresponse in linked register data. Methods were compared on their bias and coverage in different scenarios. Both, sample and mass imputation, had better coverage than traditional weighting in all scenarios.

Imputation methods can therefore be recommended over weighting as they also have practical advantages, such as that estimates outside the observed data distribution can be created and that many auxiliary variables can be taken into account. The use of sample or mass imputation depends on the specific data structure.

Key words: Weighting; mass imputation; sample imputation; coverage.

1. Introduction

Missing data form a ubiquitous source of problems in survey research. A common research scenario occurs when respondents that are sampled from the population cannot be contacted, or when they are reluctant to conform to the survey. If no analysable information about the respondent is collected, we deem it unit nonresponse. In such a scenario, we can distinguish between two missing data problems. The first problem is that, when sampling from the population, not all units from the population are recorded (which is the usual process of sampling producing missing data by design). The second problem is that the sample is found to be incomplete. The severity of these problems is related to the probability each data point has of being missing.

The mechanism that governs these probabilities is called the missing data mechanism (Rubin 1976). To describe these mechanisms, we assume to have a data set consisting of an incomplete target variable Y and a fully observed covariate X. The incomplete target variable Y has two parts: an observed part Y_{obs} and a missing part Y_{mis}. An indicator variable R can be defined that scores a 0 when Y is missing and a 1 when Y is observed.

[1] Tilburg School of Social and Behavioral Sciences, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands. Email: L.Boeschoten@tilburguniversity.edu
[2] Department of Methodology & Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. Email: G.Vink@uu.nl and J.Hox@uu.nl

Acknowledgments: The authors would like to thank the anonymous associate editor and three referees for their constructive feedback on an earlier version of this manuscript. Furthermore, the authors would like to thank Barry Schouten and Sander Scholtus for their valuable input.

If the data are Missing Completely At Random (MCAR, Rubin 1976), the response probability for the respondents and nonrespondents is equal. This can be formally defined as:

$$P(R = 0|Y_{obs}, Y_{mis}, X) = P(R = 0) \tag{1}$$

"An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck" (Van Buuren 2012, 7). If the data are Missing At Random (MAR, Rubin 1976), the distribution of the missing values is related to other observed values, formally defined:

$$P(R = 0|Y_{obs}, Y_{mis}, X) = P(R = 0|Y_{obs}, X) \tag{2}$$

"For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and if we can assume MCAR within the type of surface, then the data are MAR" (Van Buuren 2012, 7). If the distribution of the missing values relates to unobserved values, it is called Missing Not At Random (MNAR, Rubin 1976), formally defined:

$$P(R = 0|Y_{obs}, Y_{mis}, X) = P(R = 0|Y_{obs}, Y_{mis}, X) \tag{3}$$

"For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we fail to note this. If the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted" (Van Buuren 2012, 7).

Sometimes register data is available with information about the characteristics of the respondents and the nonrespondents that can be linked to the survey data (Bethlehem et al. 2011, 211). If there is a relationship between the selection mechanism and the survey variables, the estimators will systematically over- or under-represent the population characteristics. Such deviations can be corrected by weighting the observed data to conform to the known population parameters. If done properly, both distinct missing data problems can in theory be solved. However, there are several arguments against the use of weighting techniques to handle nonresponse. We list them in no particular order:

1. Weighting ignores the uncertainty about the missing data. This may result in too little variation about the estimates (Bethlehem et al. 2011, 184).
2. Weighting methods cannot create estimates that lie outside the observed data distribution. Although some researchers might view this as an advantage of weighting and would worry when a method could yield estimates outside the observed data distribution, an example given by Rubin illustrates when this could be problematic: "Consider dealing with censored data by weighting – data beyond or approaching the censoring point have zero or very small probabilities of being observed, and so either cannot be dealt with by weighting or imply a few observations with dominant weights. Weighting by inverse probabilities cannot create estimates outside the convex hull of the observed data, and estimates involving weights near the boundary have extremely large variance" (Rubin 1996, 486).

3. Uncertainty about the weights is ignored when weights are estimated from the data and thereby treated as fixed, given that the data conform to sampling variance. When taking additional measures, such as combining jackknife procedures with calibration, or by using design based analysis, weights can be treated as random.
4. Weighting has difficulties with handling large numbers of auxiliary variables, which are potentially needed to make the nonresponse ignorable (Rubin 1987, 155). Additional measures should then be taken, such as dimension reduction or propensity score estimation.
5. Weighting can have difficulties with creating sensible weights when more auxiliary information is incorporated. As a result, it is possible that the score on a target variable of an individual is used to represent a large group in the population. An illustrative example from the United States of America 2016 presidential elections show how one man heavily influenced the outcome of a poll due to extreme weights being given to his demographic category (Cohn 2016).
6. Some weighting methods cannot handle continuous variables.
7. Weighting cannot handle partial response. It is an all or nothing approach and may thereby discard valuable information (Van Buuren 2012, 22).

Because of arguments 1 and 3, we expect weighting to create too little variance and therefore to yield invalid inference (with confidence validity as defined by Rubin (1996)). We expect multiple imputation (MI) to be a good alternative method to correct for unit nonresponse, since it takes sampling variability as well as uncertainty due to missing values into account (Rubin 1987, 76). Furthermore, with MI there is no limit to the use of auxiliary information: continuous variables or the number of variables are less likely to pose problems, as the likelihood of the observed data given the unobserved data is taken into account. In cases of large numbers of variables or nonlinear associations, principal component analysis can be used (Howard 2012). In addition, item and unit non-response can be handled simultaneously with MI.

The goal of this article is to investigate whether MI is a suitable alternative for weighting when correcting for unit nonresponse. In this article, we distinguish between sample and mass imputation. With sample imputation, both item and unit nonresponse (occuring both in the sample) can be imputed. If the sample is a simple random sample without replacement ($\mathrm{SRS}_{\mathrm{WOR}}$) auxiliary information is only needed for the sample. However, sometimes registers with information about the whole population can be linked on a unit level to sample data sets. This is for example the case at Statistics Netherlands where complete population registers were used in the 2011 Dutch census (Schulte Nordholt et al. 2014). If this is the case, the nonsampled units can be imputed as well (besides the item and unit nonresponse within the sample). Mass imputation can then be applied with $\mathrm{SRS}_{\mathrm{WOR}}$ or complex samples.

Our definition of mass imputation should not be confused with the approach of Zhou et al. (2016), who generate a synthetic data set based on known population totals. A benefit of mass imputation is that every source of (linked) auxiliary information can be used for imputation. This means that a MNAR missing mechanism can become MAR, leading to more efficient estimation of (population) parameters.

We investigate the performance of weighting and both sample and mass imputation. As a reference, we also investigate complete case analysis (CCA), where no correction for unit nonresponse is made. With performance, summarized as 'valid inference' in the title, we mean obtaining unbiased parameter estimates and unbiased variance estimates.

2. Methodology

In this article, we distinguish between multiple auxiliary variables **X** and a single target variable y, which we assume to be normally distributed with mean μ and variance σ^2. If we would take a $\mathrm{SRS}_{\mathrm{WOR}}$, the estimate of the sample mean of a target variable y is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{4}$$

where y_i is the observation on the i^{th} sampled unit with $i = 1, \ldots, n$, where n is the sample size. The estimate of the variance of the mean is:

$$\mathrm{VAR}(\hat{\mu}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 \frac{1}{n}\left(1 - \frac{n}{N}\right), \tag{5}$$

where N is the size of the (finite) population. This is how μ and $\mathrm{VAR}(\mu)$ are estimated when the sample is completely observed. We will now discuss different methods to estimate these parameters in case of unit nonresponse.

2.1. Complete Case Analysis

When CCA is applied, nonrespondents are completely removed from the sample. μ and $\mathrm{VAR}(\mu)$ are estimated with the same equations used for a completely observed sample, as in Equations 4 and 5. However, with unit nonresponse, not all values in y are observed, and only the observed values in y are used to estimate μ and $\mathrm{VAR}(\mu)$ of the target variables:

$$\hat{\mu} = \frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}} y_{obs_i}, \tag{6}$$

$$\mathrm{VAR}(\hat{\mu}) = \frac{1}{n_{obs}-1}\sum_{i=1}^{n_{obs}}(y_{obs_i} - \hat{\mu}_{obs})^2 \frac{1}{n_{obs}}\left(1 - \frac{n_{obs}}{N}\right). \tag{7}$$

2.2. Weighting

The weighted mean of a target variable is defined as

$$\hat{\mu} = \frac{\sum_{i=1}^{n_{obs}} w_i y_{obs_i}}{\sum_{i=1}^{n_{obs}} w_i} \tag{8}$$

where w_i is the weight corresponding to the i^{th} observation (Biemer and Christ 2008, 318) and μ is a vector quantity (and is so throughout the remainder of the article). The weights, w_i, can be estimated with different methods, such as poststratification, linear weighting, multiplicative weighting and propensity weighting. A full description of how to apply the different methods can be found in Chapter 8 of Bethlehem et al. (2011). De Waal et al. (2011, 237–244) show that under certain conditions, linear weighting and mass imputation yield the same estimate. Therefore, it would be interesting to use this method to estimate the weights, and investigate whether these methods also yield the same inference. For this reason, we use linear weighting to estimate w_i.

Linear weighting is a calibration method, and is thoroughly discussed by, among others, Deville and Särndal (1992) and Särndal et al. (1992). When estimating weights, it is important to note first that these weights (w_i) consist of two parts:

$$w_i = d_i \delta_i, \tag{9}$$

where d_i are the sampling design weights. For a SRS_{WOR}, N and n are fixed numbers, d_i is constant and does not need to be estimated:

$$d_i = N/n. \tag{10}$$

δ_i is the adjustment factor. Our goal is to find a δ_i which makes w_i as close as possible to d_i, while respecting the calibration equation

$$\sum_{i=1}^{n_{obs}} w_i \mathbf{X}_i = \mathbf{t}_{\mathbf{X}}, \tag{11}$$

where $\mathbf{X}$ represents the auxiliary variables and $\mathbf{t}_{\mathbf{X}}$ are the population totals of $\mathbf{X}$. Minimizing the function

$$\sum_{i=1}^{n_{obs}} (w_i - d_i)^2 / d_i \tag{12}$$

leads to what is also known as linear weighting, which is a special case of calibration. We derive new weights here that modify as little as possible to the original sampling design weights d_i by minimizing the conditional value of the distance, given the realized observed sample n_{obs}. This leads to the calibrated weight

$$w_i = d_i(1 + \mathbf{X}_i' \lambda) \tag{13}$$

where λ is a vector of Lagrange multipliers determined from Equation 12:

$$\lambda = \mathbf{T}_{n_{obs}}^{-1}(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}\pi}). \tag{14}$$

The inverse of $\mathbf{T}_{n_{obs}}$ is

$$\mathbf{T}_{n_{obs}}^{-1} = \left(\sum d_i \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \tag{15}$$

and $\hat{\mathbf{t}}_{\mathbf{X}\pi}$ is the Horvitz-Thompson (Horvitz and Thompson 1952) estimator for $\mathbf{X}$:

$$\hat{\mathbf{t}}_{\mathbf{X}\pi} = \sum_{i=1}^{n_{obs}} d_i \mathbf{X}_{\mathbf{i}} \tag{16}$$

(Deville and Särndal 1992). The variance of a weighted mean can be approximated with methods such as Taylor linearization or Jacknife resampling (Stapleton 2008, 355). We use Taylor linearization and we assume for convenience that there is a vector of constants γ, such that $\gamma'\mathbf{X}_i = 1$ for all i. In that case, $\sum_{i=1}^{n_{obs}} w_i = N$. Then, the variance of a weighted mean can be approximated by:

$$\mathrm{VAR}(\hat{\mu}) = \frac{1}{N^2}\sum_{i=1}^{n_{obs}}\sum_{h=1,h\neq i}^{n_{obs}} \frac{\pi_{ih} - \pi_i\pi_h}{\pi_{ih}}\left(\delta_i\frac{e_i}{\pi_i}\right)\left(\delta_h\frac{e_h}{\pi_h}\right) \tag{17}$$

where π_i and π_h are the first order and π_{ih} the corresponding second order inclusion probabilities of observations i and h, and e_i (and e_h) are defined as:

$$e_i = y_i - \mathbf{X}_i'\mathbf{T}_{n_{obs}}^{-1}\sum_{l=1}^{n_{obs}}\mathbf{X}_l y_l d_l \tag{18}$$

(Särndal et al. 1992, 225–236).

2.3. Sample Imputation

With MI, each missing datapoint is imputed $m \geq 2$ times, resulting in m completed data sets. At least two imputations are needed to reflect the uncertainty about the imputations, although performing more imputations is often advisable. The m data sets can then be analyzed by standard procedures and the analyses combined into a single inference. A clear introduction to multiple imputation and different methods to impute the missing datapoints is given in Van Buuren (2012, Chapter 2).

With sample imputation, we only impute the nonrespondents in the sample. Because the imputation theory aims at inference about the population, sampling uncertainty is taken into account and we can use the standard rules for pooling.

The pooled estimate of μ is obtained by

$$\bar{\mu} = \frac{1}{m}\sum_{j=1}^{m}\hat{\mu}_j\,, \tag{19}$$

where m is the number of imputations with $j = 1, \ldots, m$ and $\hat{\mu}_j$ is the $\hat{\mu}$ of the j th imputed sample. $\overline{\mathrm{VAR}(\hat{\mu})}$ consists of multiple components (we therefore name it $\overline{\mathrm{VAR}(\hat{\mu})}_{\mathrm{total}}$) and is estimated

$$\overline{\mathrm{VAR}(\hat{\mu})}_{\mathrm{total}} = \overline{\mathrm{VAR}(\hat{\mu})}_{\mathrm{within}} + \mathrm{VAR}(\hat{\mu})_{\mathrm{between}} + \frac{\mathrm{VAR}(\hat{\mu})_{\mathrm{between}}}{m}, \tag{20}$$

where $\overline{\mathrm{VAR}(\hat{\mu})}_{\mathrm{within}}$ is the within imputation variance and $\mathrm{VAR}(\hat{\mu})_{\mathrm{between}}$ is the between imputation variance. $\overline{\mathrm{VAR}(\hat{\mu})}_{\mathrm{within}}$ is calculated by

$$\overline{\mathrm{VAR}(\hat{\mu})}_{\mathrm{within}} = \frac{1}{m}\sum_{j=1}^{m}\mathrm{VAR}(\hat{\mu})_{\mathrm{within}_j} \tag{21}$$

and $\mathrm{VAR}(\hat{\mu})_{\mathrm{between}}$ is calculated by

$$\mathrm{VAR}(\hat{\mu})_{\mathrm{between}} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\mu}_j - \bar{\mu})(\hat{\mu}_j - \bar{\mu})'. \tag{22}$$

2.4. *Mass Imputation*

With mass imputation, the estimate of μ is also obtained by Equation 19, although $\hat{\mu}_j$ now corresponds to the j th imputed version of the population instead of the the j th imputed sample.

Because we impute the population, there is no variance due to sampling. Therefore, $\overline{\text{VAR}(\hat{\mu})}_{\text{within}} = 0$ and we can adjust Equation 20 to

$$\overline{\text{VAR}(\hat{\mu})}_{\text{total}} = \text{VAR}(\hat{\mu})_{\text{between}} + \frac{\text{VAR}(\hat{\mu})_{\text{between}}}{m}. \tag{23}$$

For a thorough description of making multiply imputed inference when sampling variance is not of interest see Vink and Van Buuren (2014).

3. Simulation Approach

To empirically evaluate the performance of the different analysis methods, we conducted a simulation study using R (R Core Team 2015, version 3.2.2). The properties we manipulate in the simulation design can be summarized as follows:

- The correlation between the auxiliary variables and the target variables: 0.30; 0.50.
- The amount of missingness: 25%; 50%.
- The missingness mechanism: MCAR; left-tailed MAR.
- The analysis method: CCA; lineair weighting (calibration); Bayesian normal linear imputation of the sample; Bayesian normal linear imputation of the population.

We now discuss the properties of the simulation design in more detail.

3.1. *The Correlation Structure*

We start by creating a large but finite population of 100,000 units with two auxiliary (X_1 and X_2) and two target variables (Y_1 and Y_2). The population data is multivariate normally distributed with $\boldsymbol{\mu}$ and Σ:

$$\begin{pmatrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{pmatrix} = MVN(\boldsymbol{\mu}, \Sigma),$$

where $\boldsymbol{\mu}$ is:

$$\boldsymbol{\mu} = \begin{array}{c} X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{array} \begin{pmatrix} 3 \\ 2 \\ 0 \\ 170 \end{pmatrix}$$

and Σ is either:

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \end{array} \\ \begin{pmatrix} 1.00 & 0.08 & 1.34 & 1.90 \\ 0.08 & 0.25 & 0.67 & 0.95 \\ 1.34 & 0.67 & 20.00 & 4.24 \\ 1.90 & 0.95 & 4.24 & 40.00 \end{pmatrix} \end{array}$$

when the correlations between the target variables and the auxiliary variables are 0.30, and

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \end{array} \\ \begin{pmatrix} 1.00 & 0.08 & 2.24 & 3.16 \\ 0.08 & 0.25 & 1.12 & 1.58 \\ 2.24 & 1.12 & 20.00 & 4.24 \\ 3.16 & 1.58 & 4.24 & 40.00 \end{pmatrix} \end{array}$$

when the correlations between the target variables and the auxiliary variables are 0.50. The target variables X_1 and X_2 are transformed into categorical variables with respectively six and four categories, because auxiliary register information is in practice often categorical.

3.2. The Amount of Missingness and the Missingness Mechanism

From the population of size 100,000, a random sample of size 5,000 is drawn. In each sample, either 25% or 50% missingness is induced in the Y_1 and Y_2 variables.

The missingness in the target variables follow MCAR or left-tailed MAR mechanisms conform the procedure described by Van Buuren (2012, 63). With a left-tailed MAR mechanism, the probability of having missing values in the target variables is larger for smaller values on the auxiliary variables. For example, consider the number of employees of a company to be the auxiliary variable on which the missingness depends and working conditions of the company as target variables. In this situation, it is likely that more missing values are found at the companies with fewer employees. The first reason for this is that smaller companies are often less well organized. However, researchers are also probably more interested in larger companies, and are more likely to re-contact these in cases of nonresponse. If you sort companies on an axis with number of employees, you find more missing values on the left side of this axis, where the smaller companies are found.

3.3. The Analysis Method

We estimate $\hat{\mu}$ and VAR($\hat{\mu}$) of the target variables by making use of CCA, weighting, sample imputation and mass imputation. There are slight differences between the simulation setup within the different methods. For CCA, 96.25% or 97.50% of the 100,000 population values could be deleted directly from the target variables using MAR or MCAR to come to a sample of 5,000 with 25% or 50% missing values. The estimates of the incomplete sample can be compared directly to the population values.

Table 1. Smallest and largest adjustment factor per simulated condition.

		MCAR		MARleft	
cor.	% mis	min	max	min	max
	25	1.2305	1.4511	0.9682	4.2467
0.3	50	1.7472	2.3080	0.9419	13.8479
	25	1.2298	1.4513	0.9646	4.2426
0.5	50	1.7454	2.3128	0.9273	13.4502

For weighting, we first select randomly 5,000 cases from the population. Next, we create unit missingness following one of the missingness mechanisms. We weight the respondents to the total sample using the population totals. Weights are calculated using the survey package (Lumley 2014, version 3.30-3) in R (R Core Team 2015, version 3.2.2) with the `calibrate()` function. We evaluate the performance of weighting by comparing the results of the weighted sample to the population values. The design weights are $d_i = N/n = 100{,}000/5{,}000 = 20$. The adjustment factors δ_i can be found in Table 1, which can be used to compute the weights $w_i = d_i\delta_i$.

We are aware that some of the correction weights are considered large and that weighted estimates may be inefficient in such scenarios. An option would be to trim the weights to predefined boundaries. However, by not trimming the weights, we are able to investigate the performance of the method itself and its default options to other methods and their default options.

For sample imputation, we also 5,000 cases from the population and create unit missingness in the sample. Next, we multiply impute the sample and compare the results of the imputed sample to the population results.

For mass imputation, we can directly delete 96.25% or 97.50% of the values of the target variables and multiply impute the population. The results of the imputed population are compared to the original population results. Both sample and mass imputations are executed with `mice` (Van Buuren and Groothuis-Oudshoorn 2011) in `R` (R Core Team 2015) using Bayesian normal linear imputation (`mice.impute.norm()`) as the imputation method with five imputations and five iterations for the algorithm to converge.

3.4. Performance Measures

We estimate $\hat{\mu}$ and VAR($\hat{\mu}$) by using the previously discussed methods and replicate this procedure 1,000 times. In each replication, we investigate these estimates by looking at two performance measures. First, we look at the bias of $\hat{\mu}$ of the two target variables. This bias is equal to the difference between the average estimate over all replications and the population value. Next, we look at the coverage of the 95% confidence interval. This is equal to the proportion of times that the population value falls within the 95% confidence interval constructed around the $\hat{\mu}$'s of the two target variables over all replications.

3.5. Expectations

When CCA is applied and the missingness is MCAR, the probability of being missing is equal for every unit in the sample. Therefore, we do not expect biased estimates of $\hat{\mu}$.

However, with MAR, the probability of being missing is not equal for every unit, and we do expect bias. Since parameter uncertainty and uncertainty about the missing values is not taken into account when estimating the variance of the mean, we also expect undercoverage with MAR.

When weighting is applied, we expect unbiased estimates of $\hat{\mu}$ under both MCAR and MAR. The variance estimate takes the weights and parameter uncertainty into account, but not the uncertainty about the missing values. Therefore, we expect an estimate of the variance of the mean that is a bit too small, resulting in undercoverage under MAR.

For sample imputation we expect unbiased estimates and adequate coverage under both MCAR and MAR.

For mass imputation, we also expect unbiased estimates and adequate coverage under both MCAR and MAR.

4. Results

The simulation results are depicted in Table 2. Note that the results for CCA in terms of coverage and confidence interval width with correlation 0.30 and 0.50 look identical under MCAR. Small differences in the results were found, but these occur after the fourth decimal.

4.1. The Missingness Mechanism

The methods that aim to correct for the nonresponse show equivalent bias and coverage patterns under MCAR and left-tailed MAR missingness mechanisms. Naturally, the loss of observed information results in larger confidence interval widths under left tailed MAR missingness than under MCAR missingness mechanisms. CCA is unable to handle the estimation under left-tailed MAR missingness and yields large bias, zero coverage and confidence intervals that are, as expected, equally wide to those under MCAR.

4.2. The Correlation Structure

Larger correlations are often beneficial when solving incomplete data problems because the correlations give strong direction to the estimation procedure. This is clearly visible in all methods that aim to solve the missingness problem as confidence intervals tend to become smaller when the correlation between the target variables and the linked register data increases. Interestingly, the coverage rates for weighting are negatively impacted under large correlations. In this specific situation the bias remains roughly the same as under low-correlation simulations, while the confidence interval widths decrease. As a result, the simulations for weighting demonstrate lower coverage of the population mean.

4.3. The Amount of Missingness

In general, it can be said that when amounts of missingness become larger, incomplete data problems become more difficult. More specifically, the probability that you deal with a MNAR mechanism increases. None of the methods seem negatively impacted by the increased amount of missingness, when compared to the results under less missingness. However, the confidence intervals naturally tend to become wider as there is less information about the observed data.

Table 2. Simulation results. Depicted are the bias of the mean of Y_1 and Y_2, coverage of the 95% confidence interval and width of the 95% confidence interval for the four methods under varying simulation conditions.

				MCAR			MARleft		
Method	Correlation	% mis	*Y*	bias	coverage	CI width	bias	coverage	CI width
CCA	0.3	25	1	−0.0003	0.9620	0.2872	−1.1480	0.0000	0.2866
			2	0.0020	0.9580	0.4049	−1.6535	0.0000	0.4060
		50	1	−0.0003	0.9590	0.3516	−1.1728	0.0000	0.3479
			2	−0.0000	0.9620	0.4957	−1.6889	0.0000	0.4930
	0.5	25	1	−0.0002	0.9620	0.2872	−1.9291	0.0000	0.2848
			2	0.0020	0.9570	0.4049	−2.7578	0.0000	0.4024
		50	1	−0.0003	0.9590	0.3516	−1.9691	0.0000	0.3460
			2	−0.0000	0.9620	0.4957	−2.8181	0.0000	0.4888
Weighting	0.3	25	1	−0.0021	0.9310	0.2626	−0.0037	0.9370	0.2736
			2	0.0033	0.9400	0.3693	0.0007	0.9360	0.3835
		50	1	−0.0015	0.9340	0.3237	−0.0021	0.9390	0.3710
			2	0.0020	0.9370	0.4551	0.0014	0.9390	0.5184
	0.5	25	1	−0.0031	0.8820	0.2236	−0.0029	0.8950	0.2331
			2	0.0017	0.9060	0.3140	0.0004	0.9030	0.3266
		50	1	−0.0032	0.9070	0.2756	−0.0015	0.9180	0.3160
			2	−0.0004	0.9250	0.3868	0.0035	0.9080	0.4417

Table 2. Continued.

				MCAR			MARleft		
Method	Correlation	% mis	Y	bias	coverage	CI width	bias	coverage	CI width
Sample imputation	0.3	25	1	−0.0021	0.9650	0.2959	−0.0040	0.9520	0.3037
			2	0.0010	0.9460	0.4125	0.0076	0.9480	0.4293
		50	1	0.0004	0.9540	0.3422	−0.0062	0.9430	0.4281
			2	0.0005	0.9550	0.4879	0.0106	0.9540	0.6103
	0.5	25	1	−0.0016	0.9650	0.2824	−0.0014	0.9460	0.2905
			2	0.0013	0.9440	0.3954	0.0009	0.9450	0.4077
		50	1	0.0005	0.9540	0.3422	−0.0044	0.9540	0.3842
			2	0.0007	0.9550	0.4879	0.0098	0.9390	0.5402
Mass imputation	0.3	25	1	0.0003	0.9450	0.3857	−0.0200	0.9510	0.5419
			2	0.0005	0.9590	0.5480	0.0268	0.9460	0.7713
		50	1	−0.0008	0.9390	0.4772	−0.0237	0.9570	0.6636
			2	0.0030	0.9560	0.6752	0.0229	0.9440	0.9117
	0.5	25	1	−0.0001	0.9570	0.3289	−0.0051	0.9490	0.4663
			2	0.0007	0.9630	0.4603	0.0423	0.9400	0.6507
		50	1	−0.0033	0.9390	0.4033	−0.0051	0.9530	0.5743
			2	−0.0010	0.9470	0.5665	0.0438	0.9620	0.8202

Note that results of two target variables Y_1 and Y_2 are shown, which both have their own mean and variance, as illustrated in Subsection 3.1.

4.4. Overall Efficiency

We investigate efficiency of the methods in the sense that we investigate which methods have the smallest confidence interval widths under which conditions. When investigating the results, we see that CCA is an efficient method yielding valid inference under MCAR. There is no need for handling the nonresponse as the nonresponse is perfectly ignorable: the set of observed values can simply be analyzed to obtain unbiased estimates about the population. Even though the missingness is MCAR, treating the missingness can increase the statistical power of the analyses at hand. This is demonstrated by weighting and imputing the sample as the confidence intervals under these approaches are generally more narrow than under CCA. Mass imputation, on the other hand, does not show this result. This can simply be explained by the severity of the problem that is considered with mass imputation in our simulation setup. After all, under mass imputation we aim to solve at least a 96.25% missingness problem.

Even though mass imputation may yield less sharp inference than sample imputation and weighting, the inference is valid and exhibits correct variance properties under all simulation conditions. The same can be said of sample imputation, but with much sharper inference. The estimates obtained under weighting are unbiased, the intervals are among the smallest, but the coverage rates are somewhat low. Especially when larger correlations occur in the data, one could question the validity of inference obtained by weighting. Furthermore, it is surprising that these low coverage rates occur under both MCAR and MAR, indicating that the variance of a weighted mean estimated using Taylor linearization indeed ignores uncertainty about the missing data and possibly about the weights as well.

5. Discussion

We have demonstrated that weighting and imputation are practically equivalent when unbiased estimation is of interest. However, the inference obtained under weighting may be questionable in situations where multiple imputation approaches exhibit correct variance properties and well-covered population estimates. In general it holds that inferring about the population by imputing the sample yields efficient, unbiased estimates in all simulated conditions, which is in line with conclusions drawn by Peytchev (2012).

A main characteristic of our simulation approach is that it deals with a $\mathrm{SRS}_{\mathrm{WOR}}$. With more complex sampling approaches, it would not be sufficient to only impute the sample, since the complex sampling structure is then ignored. Although we did not investigate this, we do expect that mass imputation will lead to unbiased and efficient estimates when a more complex sample is drawn because the design of the complex sample is always based on observed information, so the missingness mechanism describing the sample to the population is always MAR. However, this is not included in this simulation study, and additional research should be done.

Furthermore, in this simulation we assume quite an ideal situation, where the sample is perfectly linked to a completely observed population register. Of course, this is not often the case in practice. In addition to the traditional Total Survey Error framework introduced by Groves et al. (2009), Zhang (2012) introduced a two-phase life cycle of integrated statistical micro data, which also discusses the errors that might be encountered when multiple data sets are combined, such as identification or comparability error. Furthermore, we also assume that our population register is perfectly observed. This is in practice also not often the case, although this is commonly assumed by many researchers. Recently, imputation methods have been developed to take misclassification in combined data sets into account, for example by assuming that a certain proportion of the data is misclassified (Manrique-Vallier and Reiter 2016) or by estimating the number of misclassified units by using information from multiple sources (Boeschoten et al. 2016).

It is clear that weighting does not include all sources of uncertainty. This limits the validity of the inference obtained under weighting. Theoretically, these sources of uncertainty could be added to the estimations that are obtained from weighted data sets. However, we have demonstrated that the imputation approaches take the sources of variations about the observed and missing data properly into account. Adjusting the weighted estimation to allow for valid inference under unit nonresponse would therefore be redundant as it is a complicated step to solve a problem that can be straightforwardly solved by another approach.

In addition, weighting cannot handle partial response (Van Buuren 2012, 22). Analyzing multivariate response data with partial responses will be particularly problematic when weighting is applied, and multiple imputation is a very suitable alternative in this setting.

It is known that complete case analysis yields valid inference under MCAR mechanisms and that its performance may be severely impaired under MAR missingness. The results of complete case analysis in simulations can be very informative, as it can act as a point of reference for the performance of other methods. At the same time, the validity of the simulation scheme can be assessed, because we know the theoretical properties under which complete case analysis can be applied. Failure to meet these expectations indicates a faulty simulation scheme. This is not the case.

The simulation study conducted in this article illustrated that multiple imputation methods lead to valid inference in situations of unit nonresponse and have practical advantages over weighting. Whether sample or mass imputation methods should be used depends on the specific data structure.

6. References

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys,* volume 568 of *Wiley Handbooks in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Boeschoten, L., D. Oberski, and T. de Waal. 2016. “Estimating Classification Error under Edit Restrictions in Combining Survey-Register Data.” *Journal of Official Statistics* 33:921–962. Doi: http://dx.doi.org/10.1515/JOS-2017-0044.

Cohn, N. 2016. “How One 19-Year-Old Illinois Man is Distorting National Polling Averages.” *The New York Times*. Available at: https://nyti.ms/2k5sB5z (accessed September 26, 2017).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation,* volume 563 of *Wiley Handbooks in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Deville, J.-C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American statistical Association* 87: 376–382.

Groves, R.M., F.J. Fowler, Jr, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology,* volume 561 of *Wiley Series in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Howard, W.J. 2012. *Using Principal Component Analysis (pca) to Obtain Auxiliary Variables for Missing Data in Large Data Sets*. University of Kansas. PhD Dissertation.

Lumley, T. 2014. *Analysis of Complex Survey Samples*. Available at: http://cran.r-project.org/web/packages/survey/survey.pdf (accessed September 26, 2017).

Manrique-Vallier, D. and J.P. Reiter. 2016. "Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data." *Journal of the American Statistical Association*. Doi: http://dx.doi.org/10.1080/01621459.2016.1231612.

Peytchev, A. 2012. "Multiple Imputation for Unit Nonresponse and Measurement Error." *Public Opinion Quarterly* 76: 214–237. Doi: https://doi.org/10.1093/poq/nfr065.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, USA.

Rubin, D.B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91: 473–489.

Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag.

Schulte Nordholt, E., J. van Zeijl, and L. Hoeksma. 2014. *Dutch Census 2011, Analysis and Methodology*. The Hague/Heerlen. Available at: https://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/0/2014b57pub.pdf (accessed 26 September 2017).

Stapleton, L.M. 2008. "Analysis of Data from Complex Surveys." In *International Handbook of Survey Methodology*, edited by E.D. De Leeuw, J.J. Hox, and D. Dillman, 342–369. Psychology Press, Taylor & Francis Group, New York.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data*. CRC press, Boca Raton, Florida.

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45: 1–67. Doi: http://dx.doi.org/10.18637/jss.v045.i03.

Vink, G. and S. van Buuren. 2014. "Pooling Multiple Imputations when the Sample Happens to be the Population." *arXiv preprint arXiv:1409.8542*. Available at: https://arxiv.org/pdf/1409.8542.pdf (accessed 26 September 2017).

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016. "A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation." *Biometrics* 72: 242–252.

Received November 2015
Revised January 2017
Accepted March 2017