Laitila, Thomas; Wang, Lisha

Working Paper

Calibration Estimation under Non-response and Missing Values in Auxiliary Information

Working Paper, No. 2/2015

Provided in Cooperation with:
Örebro University School of Business

Suggested Citation: Laitila, Thomas; Wang, Lisha (2015) : Calibration Estimation under Non-response and Missing Values in Auxiliary Information, Working Paper, No. 2/2015, Örebro University School of Business, Örebro

This Version is available at:
https://hdl.handle.net/10419/244506

WORKING PAPER

2/2015

Calibration Estimation under Non-response and Missing Values in Auxiliary Information

Thomas Laitila and Lisha Wang
Statistics

ISSN 1403-0586

http://www.oru.se/Institutioner/Handelshogskolan-vid-Orebro-universitet/Forskning/Publikationer/Working-papers/
Örebro University School of Business
701 82 Örebro
SWEDEN

Calibration Estimation under Non-response and Missing Values in Auxiliary Information

Thomas Laitila and Lisha Wang
Department of Statistics, Örebro University, S-701 82 Örebro, Sweden

May 6, 2015

Abstract

The calibration approach is suggested in the literature for estimation in sample surveys under non-response given access to suitable auxiliary information. However, missing values in auxiliary information come up as a thorny but realistic problem. This paper considers the consistency of the calibration estimator suggested by Särndal & Lundström (2005) for estimation under nonresponse, connected with how imputation of auxiliary information based on different levels of register information affects the calibration estimator. An illustration is given with results from a small simulation study.

Keywords Sample survey, non-response, imputation, consistency, bias

1 Introduction

Non-response is undesirable but inevitable in surveys and techniques are required to promote the accuracy of estimation. There are several papers addressing the calibration technique for estimation in sample surveys. Deville & Särndal (1992) propose calibration weighting in survey estimation with multivariate auxiliary information. Särndal & Lundström (2005) propose a calibration estimator for nonresponse adjustment. Kott (2006) considered calibration using a known response probability function and suggests calibration for estimation. Montanari & Ranalli (2005) discuss calibration estimator in a neural network mode.

The core idea of the calibration estimation approach is to utilize auxiliary information by replacing the design weights in the Horwitz-Thompson (HT) estimator with weights replicating known population totals when attached to auxiliary variables. Särndal & Lundström (2005) makes a distinction between using population level and sample level information. They also derive several approximate bias expressions for the estimator proposed. These give

guidelines for selection of auxiliary information aiming for the reduction of the bias of the calibration estimator. Attempts to find indicators and algorithms for selection of appropriate sets of auxiliary variables are found in e.g.Särndal & Lundström (2008) and Schouten (2007).

In addition to nonresponse for study variables, missing values of auxiliary variables frequently occur in registers. These missing values can be substituted with imputed values derived from rules defined by information at the response, sample or population levels. After imputation the resulting variable can be treated as any other auxiliary variable. However, the properties of the resulting calibration estimator depend on the way the imputations are derived. Consider for instance the cases of imputations with at constant and conditional regression mean imputation.

This paper is concerned with the potential bias introduced in the Särndal & Lundström (2005) estimator by imputing auxiliary information. Of special interest are the effects of using information at the population, the sample and the response set levels, respectively. Results are obtained by considering the probability limits of the calibration estimator and by using a small simulation illustration. Results indicate that imputation does not add an extra source of bias. Increased bias may be obtained due to the effect of using less powerful information in comparison with the original not fully observed auxiliary variable.

The calibration estimator proposed by Särndal & Lundström (2005) is defined in the next section and its probability limit is considered. Section 3 considers deterministic imputation for missing values of auxiliary and instrument variables and the probability limits of the resulting calibration estimator is derived. Results are illustrated with a small simulation study in Section 4 and a discussion on the results and future research are found in the final section.

2 The Calibration Estimator

Consider a finite population with N elements $U=\{1, 2, \ldots, N\}$, in which y_k is a target variable, x_k is a vector of auxiliary variables with known or estimated population totals, and z_k is a vector of instrument variables with the same length as x_k. The instrument vector will be assumed to satisfy the restriction $\mu' z_k = 1$ for all $k \in U$ and some constant vector μ.

A probability sample s with expected sample size $n(N)$ is selected from U by a probability sampling design $p(s)$. When non-response occurs, only a subset of the sample $r \subset s$ is observable, where the size of the response set is denoted as n_r. The response mechanism is assumed random with $q(r|s)$ denoting the conditional response distribution, such that the probability of a response of element k given its selection to a sample equals $\theta_k = Pr(k \in r|k \in s, s)$. In

addition to the observations y_k ($k \in r$), assume data on x_k and z_k are also available for the units in the response set r.

The calibration estimator of the population total $Y = \sum_U y_k$ suggested by Särndal & Lundström (2005) is

$$\hat{Y}_w = \sum_r w_k y_k \tag{1}$$

where $w_k = d_k v_k$, $v_k = 1 + \lambda_r z_k$, $\lambda_r = (\hat{X} - \sum_r d_k x_k)'(\sum_r d_k z_k x_k')^{-1}$, and $\hat{X}$ denotes the vector with known or estimated population totals of x_k.

The calibration estimator (1) with $z_k = x_k$ can be rewritten in the form of a generalized regression (GREG) estimator (e.g. Deville & Särndal (1992)). Adapting for a general instrument vector, the calibration estimator equals

$$\hat{Y}_w = \hat{X}'\hat{B}_r + \sum_r d_k(y_k - x_k'\hat{B}_r) \tag{2}$$

where

$$\hat{B}_r = (\sum_r d_k z_k x_k')^{-1}(\sum_r d_k z_k y_k) \tag{3}$$

Consider the definitions:

Definition 2.1. *(Sequence of populations) The vector t_k=(y_k, x_k', z_k') is non-random, real-valued and defined for a bounded set such that $\|t_k\| < \kappa$ for some $\kappa < \infty$. Assume the existence of the infinite sequence $\{t_k\}_{k=1}^{\infty}$. The population U_N is then defined as the index set for the first N units in the sequence $\{t_k\}_{k=1}^{\infty}$ with associated set of variable vectors $\{t_1, t_2, \ldots, t_N\}$.*

Definition 2.2. *(Sequence of samples/response sets) For a given population U_N, a probability sample $s_N \subseteq U_N$ of expected size $n(N)$ is drawn using a probability sampling design $p_N(s)$ with positive inclusion probabilities $\pi_k > \varsigma > 0$. Conditionally on the sample, observations of t_k are obtained for a subset of the sample, $r_N \subseteq s_N$ according to the non-response distribution $q(r|s)$ yielding response probabilities $\theta_k = Pr(k \in r_N | k \in s_N)$.*

Let $\hat{B}_{Nr}$ denote the estimator (3) defined on the response set r_N. Also, define

$$B_{N\theta} = (\sum_{U_N} \theta_k z_k x_k')^{-1}(\sum_{U_N} \theta_k z_k y_k) \tag{4}$$

The index N on statistics below is used to indicate its calculation on population U_N or its subsets s_N and r_N .

Lemma 1. *Under the definitions (2.1) and (2.2) and the assumption i)* $\sum\sum_{U_N:k\neq l} \mid \pi_{kl}d_kd_l-1 \mid= O(N)$*, consider the statistic* $\hat{\Psi}_N = N^{-1}\sum_{r_N} d_kc_k$ *where* c_k *(*$k \in U_N$*) are nonrandom real valued scalars bounded by* $\mid c_k \mid< \kappa < \infty$*. Then* $\hat{\Psi}_N = O_p(1)$*,* $(\hat{\Psi}_N - E(\hat{\Psi}_N)) = O_p(N^{-1/2})$ *and* $E(\hat{\Psi}_N) = N^{-1}\sum_{U_N}\theta_kc_k$*.*

A proof of Lemma 1 follows from Corollary 1.7.1.1 in Fuller (2009).

Theorem 2.1. *(Probability limit for* $\hat{B}_{Nr}$*) Under the assumptions in Lemma 1 and the assumption* $N^{-1}\sum_{U_N}\theta_kz_kx_k'$ *is non-singular for all* $N>N_0$*, then*

$$p\lim_{N\to\infty}(\hat{B}_{Nr} - B_{N\theta}) = 0 \tag{5}$$

A proof of Theorem 2.1 follows Slutsky's theorem after applying Lemma (1) to the matrices defining $\hat{B}_r$ in equation (3). From Theorem 2.1, the following corollary is obtained.

Corollary 2.1. *(Consistency of* $\hat{Y}_w$*) If* $plim_{N\to\infty}(\hat{X}_N - X_N)/N = 0$*, then*

$$plim_{N\to\infty}(\hat{Y}_{wN} - Y_{N\theta})/N = 0 \tag{6}$$

where $\hat{Y}_{wN} = \hat{X}_N'\hat{B}_{Nr}$ *and* $Y_{N\theta} = X_N'B_{N\theta}$

Due to the restriction $\mu' z_k = 1$, $Y = X'B_U$ were $B_U = (\sum_U z_kx_k')^{-1}\sum_U z_ky_k$. Corollary 2.1 then yields the approximate bias expression

$$Bias(\hat{Y}_w) \approx X'(B_\theta - B_U) \tag{7}$$

This approximate bias expression equals the nearbias expressions in Corollary 9.1 in Särndal & Lundström (2005).

Corollary 2.1 and the bias expression (7) shows the calibration estimator to be biased and inconsistent in general. However, from the bias expression, Särndal & Lundström (2005) derive conditions for an approximate zero bias. Here, the ability to reduce bias strongly relies on the use of appropriate auxiliary information (Särndal & Lundström (2005), Särndal (2011)).

3 Imputations in Auxiliary Information

3.1 Probability limits

A frequent problem when using register information or sample survey data for construction of sets of auxiliary variables is missing data. It is therefore of interest to understand the effects of using imputed values in the auxiliary information used for calibration. Here, missing values of auxiliary (or instrument) variables are assumed generated by a non-random mechanism.

To keep the derivation general, the problem of imputation for missing values of instrument variables is also treated, since one option for the instrument variable is to use the auxiliary vector, i.e. $z_k = x_k$. Below it is also assumed that the instrument vector, with or without imputed values, satisfies the restriction $\mu' z_k = 1$ for all $k \in U$ and some constant vector μ.

There are several methods available for construction of imputed values. One aspect upon which the methods can be divided is whether the imputations made are deterministic or random. Here, only deterministic imputation is considered.

Let A denote either U, s or r. Also let U_x denote the part of the population with values of the auxiliary variable, and let $\bar{U}_x$ be the part with missing values. Similarly, r_z denotes the part of the response set with values of the instrument variable z_k, and $\bar{r}_z$ denotes the part with missing values. The divisions here of the sets U and r are treated as nonrandom.

For notation of variable vectors which might contain imputed values, the notation of Särndal & Lundström (2005) is adapted and the following notations are introduced

$$\tilde{x}_{\bullet k}(A) = 1(k \in U_x)x_k + 1(k \in \bar{U}_x)\hat{x}_k(A) \tag{8}$$

and

$$\tilde{z}_{\bullet k}(A) = 1(k \in r_z)z_k + 1(k \in \bar{r}_z)\hat{z}_k(A) \tag{9}$$

Here, 1() denotes the indicator function equaling one if the argument is true, and equaling zero otherwise. In (8) and (9) $\hat{x}_k(A)$ and $\hat{z}_k(A)$ denote imputed values derived from calculations on the set A. Different sets A can be used for x_k and z_k.

For a treatment of the asymptotic properties of the calibration estimator based on imputed auxiliary and/or instrument variables, let $x_k(A)$ and $z_k(A)$ denote imputations made when $N \to \infty$. Here the argument A is not defining the actual set used for calculations, it represents the asymptotic counterparts of an imputation method based on the set A. With these "asymptotic" imputations, equations (8) and (9) are rewritten as $x_{\bullet k}(A) = 1(k \in U_x)x_k + 1(k \in \bar{U}_x)x_k(A)$ and $z_{\bullet k}(A) = 1(k \in r_z)z_k + 1(k \in \bar{r}_z)z_k(A)$, respectively. The instrument vectors with imputed information are assumed to satisfy the restriction $\mu' \tilde{z}_{\bullet k}(A) = \mu' z_{\bullet k}(A) = 1$ for all $k \in r$ and some constant μ. The following "uniform convergence" assumptions on the imputations are made.

Assumption 1. *There exists finite constants M_x and M_z such that*

$$\max_{k \in U} \parallel \tilde{x}_{\bullet k}(A) - x_{\bullet k}(A) \parallel < M_x \omega_{xN} \tag{10}$$

and

$$\max_{k\in r} \parallel \tilde{z}_{\bullet k}(A) - z_{\bullet k}(A) \parallel < M_z \omega_{zN} \tag{11}$$

where $\omega_{xN} = o_p(1)$ *and* $\omega_{zN} = o_p(1)$.

For simplicity, the index N for population U_N on the vectors in Assumption 1 has been omitted.

Using the defined auxiliary and instrument variable vectors, the following three parameter vectors are defined.

$$\hat{B}_r(A) = (\sum_r d_k \tilde{z}_{\bullet k}(A)\tilde{x}'_{\bullet k}(A))^{-1}(\sum_r d_k \tilde{z}_{\bullet k}(A) y_k) \tag{12}$$

$$B_\theta(A) = (\sum_U \theta_k z_{\bullet k}(A)x'_{\bullet k}(A))^{-1}(\sum_U \theta_k z_{\bullet k}(A) y_k) \tag{13}$$

and

$$B_U(A) = (\sum_U z_{\bullet k}(A)x'_{\bullet k}(A))^{-1}(\sum_U z_{\bullet k}(A) y_k) \tag{14}$$

The calibration estimator (1) based on imputed values equals $\hat{Y}_w(A) = \hat{X}'(A)\hat{B}_r(A)$, where $\hat{X}'(A)$ is the vector with known or estimated population totals of auxiliary variables with imputations, e.g. $\hat{X}'(A) = \sum_U \tilde{x}_{\bullet k}(A)$.

The following theorem is shown in the appendix.

Theorem 3.1. *(Probability limit for* $\hat{B}_{Nr}(A)$*) Assume Definition 2.1, where* $x_k = x_{\bullet k}(A)$ *and* $z_k = z_{\bullet k}(A)$*, Definition 2.2, and Assumption 1. Also assume i) the sampling design yields second order inclusion probabilities* $\pi_{kl} = Pr(k\&l \in s)$ *such that* $\sum\sum_{U_N, k\neq l} \mid \pi_{kl} d_k d_l - 1 \mid = O(N)$*, and ii)* $N^{-1}\sum_{U_N} \theta_k z_{\bullet k}(A)x'_{\bullet k}(A)$ *is non-singular for all* $N > N_0$*, then*

$$p \lim_{N\to\infty} (\hat{B}_{Nr}(A) - B_{N\theta}(A)) = 0 \tag{15}$$

Proof: Let $\hat{B}^*_r(A) = (\sum_r d_k z_{\bullet k}(A)x'_{\bullet k}(A))^{-1}(\sum_r d_k z_{\bullet k}(A) y_k)$. With Assumption 1 we obtain

$$\begin{aligned}
\parallel \frac{1}{N}\sum_r d_k \tilde{z}_{\bullet k}(A)\tilde{x}'_{\bullet k}(A) - \frac{1}{N}\sum_r d_k z_{\bullet k}(A)x'_{\bullet k}(A) \parallel \leq & \parallel \frac{1}{N}\sum_r d_k (\tilde{z}_{\bullet k}(A) - z_{\bullet k}(A))(\tilde{x}_{\bullet k}(A) - x_{\bullet k}(A))' \parallel \\
& + \parallel \frac{1}{N}\sum_r d_k (\tilde{z}_{\bullet k}(A) - z_{\bullet k}(A))x'_{\bullet k} \parallel + \parallel \frac{1}{N}\sum_r d_k z_{\bullet k}(A)(\tilde{x}_{\bullet k}(A) - x_{\bullet k}(A))' \parallel \\
& \leq O_p(1) M_x \omega_{xN} M_z \omega_{zN} + O_p(1) M_z \omega_{zN} \kappa + O_p(1) M_x \omega_{xN} \kappa = o_p(1)
\end{aligned}$$

and similarly $\parallel \frac{1}{N}\sum_r d_k \tilde{z}_{\bullet k}(A) y_k - \frac{1}{N}\sum_r d_k z_{\bullet k}(A) y_k \parallel \leq o_p(1)$, so that $p\lim_{N\to\infty}(\hat{B}_r(A) - \hat{B}^*_r(A)) = 0$. According to Theorem 2.1, $p\lim_{N\to\infty}(\hat{B}^*_r(A) - B_\theta(A)) = 0$, and the result follow by the triangular inequality. □

The theorem gives the corollary

Corollary 3.1. *Assume the assumptions of Theorem 3.1 and* $plim_{N\to\infty}(\hat{X}_N(A)-X_N(A))/N = 0$*, where* $X_N(A) = \sum_{U_N} x_{\bullet k}(A)$*, then*

$$plim_{N\to\infty}(\hat{Y}_{N\omega}(A) - Y_{N\theta}(A))/N = 0 \tag{16}$$

where $Y_{N\theta}(A) = X'_N(A)B_{N\theta}(A)$

Corollary 3.1 is the major result of this paper. First, it gives the approximative bias expression for the calibration estimator based on auxiliary and instrument variables containing imputations as

$$Bias(\hat{Y}_\omega(A)) \approx X'(A)(B_\theta(A) - B_U(A)) \tag{17}$$

This expression is of the same form as expression (7), i.e. imputation does not add new components to the structure of the bias. Second, the bias expression has the same form irrespective of what set of data (U, s, or r) is used for deriving imputations. Finally, Theorem 3.1 shows that the design weighted least squares solutions (3) and (12) converge in probability to two different population vectors. Also, the two least squares solutions can be considered as inconsistent estimators of two different true population regressions vectors. As the bias expressions in (7) and (17) show, the bias of the calibration estimator is defined by the distance between the probability limits of (3) and (12), and the corresponding true population regression vectors. Thus, without additional assumptions, it is not possible to conclude that the bias of calibration estimators based on imputed values are larger than the bias obtained if all auxiliary and instrument variable values were known.

3.2 Mean value and regression imputation

Mean value imputation for the jth element in x_k is given by

$$\hat{x}_{jk}(A) = \hat{x}_j(A) = n^{-1}_{A_{x_j}} \sum_{A_{x_j}} x_{jl}$$

Then $\hat{x}_{jk}(U) = x_j(U) = n^{-1}_{U_{x_j}} \sum_{U_{x_j}} x_{jl}$ and $\hat{x}_{\bullet jk}(U) - x_{\bullet jk}(U) = 0$. Furthermore, the design weighted sample mean is $\hat{x}_{jk}(s) = x_j(s) = \hat{N}^{-1}_{s_{x_j}} \sum_{s_{x_j}} d_l x_{jl}$ with $\hat{N}_{s_{x_j}} = \sum_{s_{x_j}} d_k$. The sample mean is consistent for the mean of the subpopulation $U_x \subset U$, i.e. $plim_{N\to\infty}(x_j(s) - x_j(U)) = 0$. Thus, using mean imputation based on available observation in the population or in the sample, weighted with the design weight, satisfies Assumption 1.

An interesting result here is that the design weighted sample means converge in probability to the population level means, whereby the two imputation methods yields asymptotically equivalent calibration estimators.

Assumption 1 is also fulfilled by using mean imputation based on the response set. Consider the design weighted response set mean $\hat{x}_{jk}(r) = x_j(r) = \hat{N}_{r_{x_j}}^{-1} \sum_{r_{x_j}} d_l x_{jl}$ with $\hat{N}_{r_{x_j}} = \sum_{r_{x_j}} d_k$. This quantity converges in probability to the θ weighted population mean $x_j(U) = N_{\theta_{x_j}}^{-1} \sum_{U_{x_j}} \theta_l x_{jl}$ with $N_{\theta_{x_j}} = \sum_{U_{x_j}} \theta_k$.

For mean imputation using sample or response set information, respectively, note that the uniform convergence assumed in Assumption 1 is obtained since the same value is imputed for all units with missing values.

For regression imputation, consider the imputations,

$$\hat{x}_{jk}(A) = u_k' \hat{\delta}(A)$$

where u_k is a finite dimensional vector of non-random variables, available for all $k \in U$, and

$$\hat{\delta}(A) = (\sum_{A_{x_j}} v_k u_k u_k')^{-1} \sum_{A_{x_j}} v_k u_k x_{jk}$$

where v_k denotes some positive weight. Suppose this regression coefficient estimator is consistent for $\delta(A)$, i.e. $plim_{N\to\infty}(\hat{\delta}(A) - \delta(A)) = 0$, and let $x_{jk}(A) = u_k' \delta(A)$ where $\| u_k \| < M$. Then Assumption 1 is satisfied since $| \hat{x}_{jk}(A) - x_{jk}(A) | < M \| \hat{\delta}(A) - \delta(A) \|$.

4 Simulation

To illustrate how bias of the calibration estimator is influenced by using imputed values for the auxiliary variable, a simulation experiment is conducted based on a real dataset with 1046 observations, where the amount of fish consumption, birth year, education level and civil status are collected. Fish consumption, education level and civil status are all categorical variables, valuing from 0 to 6, 1 to 3, and 1 to 7 respectively. Variable age is generated from variable birth year for later use.

A population is generated by enlarging the original dataset to 100,000 observations with random sampling with replacement. To achieve a large population without duplicates, a term $\varepsilon/10$ is added to the original values of variable age and education level, where ε is a random number from $N(0, 1)$. Thereafter another variable "personal income" is generated based on linear regression $income$=1.95*age-49*$gender$+53.44*$education$+39.04, where the coefficients is obtained from regression on statistics presented in the report Folk- och bostadsräkningarna 1990. For avoiding duplicates and better controlling the correlation between y_k and x_k, the values of fish consumption is rewritten by the predicted value of the regression $fishconsumption$ =1.1935+0.0008*$income$-0.0167*$civilstatus$+$\varepsilon/10$.

The intention of the simulation study is to provide a numerical example of the results in earlier sections. In the simulation study, the total value of fish consumption y_k is of interest. And it is assumed that the variable income is the only accessible variable and highly correlated with y_k, which is denoted as x_k and will be used as auxiliary information for calibrating the total value of y_k. Also $z_k = x_k$. Both x_k and y_k have 30% missing values at random. The missing values in x_k will be replaced by group-mean in each group categorized by variable civil status which is denoted as u_k. A random sample consisting y_k, x_k and u_k with 1000 observations will be drawn from the population.

The bias of the calibration estimator $Bias(\hat{Y}_w) = E(\hat{Y}_w) - Y$ will be studied in four different cases with different patterns of response probabilities for y_k and probabilities of missing values of x_k.

- Case I: θ_k is constant and ϑ_k is constant.
- Case II: θ_k is varying and ϑ_k is constant.
- Case III: θ_k is constant and ϑ_k is varying.
- Case IV: θ_k is varying and ϑ_k is varying.

Response probabilities and probabilities of missing values are in the four cases given by:

- $\theta_k = 70\%$ in case I/III
- $\theta_k = 50\%$ if income ≤ 215, 70% if income $\in$ (215,265), and 90% if income ≥ 265 in case II/IV.
- $\vartheta_k = 30\%$ in Case I/II
- $\vartheta_k = 50\%$ if $y_k \geq 4.05$, 45% if $y_k \in [3.95,4.05)$, 40% if $y_k \in [3.87,3.95)$, 35% if $y_k \in [3.8,3.87)$, 25% if $y_k \in [3.73,3.87)$, 20% if $y_k \in [3.65,3.73)$, 12% if $y_k \in [3.53,3.65)$, and 5% if $y_k <3.53$ in case III/IV

Here, θ_k is the response probability in y_k and ϑ_k is the probability that x_k is not missing in register system.

The group-mean imputation will be utilized to make up for the missing values in auxiliary variable x_k. In this stage, the following three kinds of collection of objects (i.e., A) are considered.

Imputation 1 $A = U$, i.e., the estimator for imputation is based on the whole population.

Table 1: Group-means in different imputation levels

civil status	Group-Mean		
	Imputation1 $A = U$	Imputation2[a] $A = s$	Imputation3[b] $A = r$
1	244.82	249.46	247.98
2	238.26	235.75	241.24
3	227.56	225.68	225.43
4	225.39	223.37	222.87
5	253.33	274.36	276.67
6	247.84	236.95	239.78
7	242.96	261.59	248.82

[a] Group-mean listed in this column is only one example of a sample in one of the iterations during the simulation

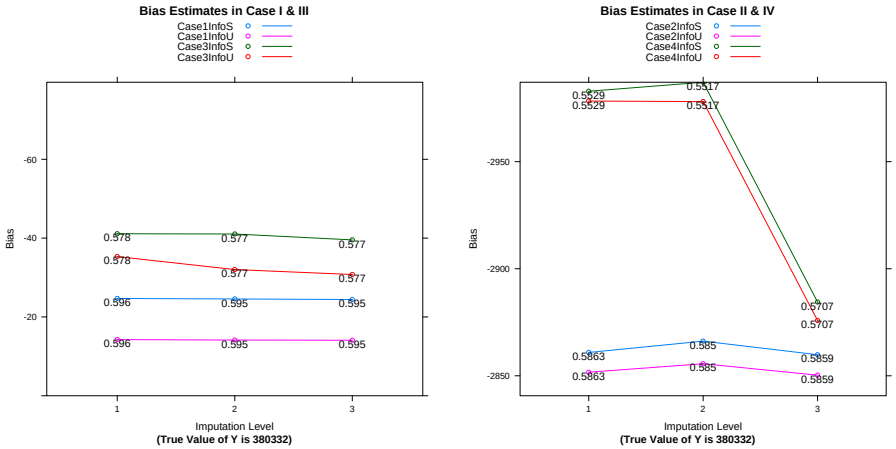
[b] Group-mean listed in this column is only one example of a response set in one of the iterations during the simulation

Imputation 2 $A = s$, i.e., the estimator for imputation is based on the sample level.

Imputation 3 $A = r$, i.e., the estimator for imputation is based on the response level.

In our study, take Case I as an example, the group-means in different imputation levels are displayed in Table 1.

Replicating the simulation for 5000 times, the expectation of the calibration estimator is estimated by $E(\hat{Y}_w) = \sum_{i=1}^{5000} \hat{Y}_{w_i}/5000$ and the bias is estimated with $Bias(\hat{Y}_w) = E(\hat{Y}_w) - Y$. Below shows the profile graph of the bias estimates in each case under different imputation level.



It is told from the graphs above that the biases of the calibration estimator vary very slightly within each case, no matter if the auxiliary variable with missing values is imputed at population level, sample level or response level. And the variance of the calibration estimates are quite close within each case as well.

A steep decrease occurs in Case IV when using response level information for imputation. It could be explained by the increasing correlation between interested variable (fish consumption) and auxiliary variable (income)as labelled on the graphs, whereas the stable correlation corresponds with the stable bias in other cases.

In comparison, the bias estimates of calibration estimator with full-recorded auxiliary are -22 (with InfoU) and -21 (with InfoS) respectively when the response probability of y_k is constant, and the bias estimates increase to -2170 (with InfoU) and -2158 (with InfoS) when the response probability of y_k is varying.

5 Discussion

This paper presents results of importance for applied use of the Särndal & Lundström (2005) estimator when missing values prevail among the instrument and auxiliary variables. The major aim of the estimator is provide estimators with reduced bias due to nonresponse. Results here show that imputation of instrument/auxiliary variable values does not in itself contribute to bias. However, in comparison with fully observed auxiliary information, variables with imputed values can be expected to be less powerful whereby an indirect effect of increased bias is obtained.

This result is valid for imputations made using information from a population register, the sample or the response set, which is a little remarkable. One may expect the response set being less suited for deriving imputations since variable distributions are distorted by the nonresponse. However, one case considered in the simulation indicate the effect might be the reverse. Further explorations on this topic is of interest.

When imputations (deterministic) are made using population level information only, the variance estimator proposed by Särndal & Lundström (2005) can be used. When imputations are based on sample or response set information, imputation adds an additional random component to the estimator whereby this variance estimator may not be valid. This is another topic for further research.

References

Deville, J. & Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376 – 382.

Fuller, W. (2009). *Sampling statistics*. John Wiley & Sons.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* **32**, 133 – 142.

Montanari, G. & Ranalli, M. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* **100**, 1429 – 1442.

Särndal, C. (2011). Three factors to signal non-response bias with applications to categorical auxiliary variables. *International Statistical Review* **79**, 233 – 254.

Särndal, C. & Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Särndal, C. & Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics* **24**, 167 – 191.

Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics* **23**, 51 – 68.