A Simulation Study on Nonresponse-bias for Calibration Estimator with Missing Auxiliary Informaiton

Lisha Wang

Örebro University, e-mail: lisha.wang@oru.se

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 is connected with how imputation of auxiliary information based on different levels of register information affects the calibration estimator. Results show that the level of register information used for deriving imputation models only marginally affects the calibration estimator bias. The results are obtained under different patterns of non-response in the target variable and missing values of the auxiliary variable.
Keywords: Nonresponse, calibration, imputation, bias, auxiliary variable

1 Introduction

The calibration approach is by Särndal & Lundström (2005) suggested for estimation in sample surveys with non-response. Calibration implies computation of weights for sampled elements in the response set such that applied to known auxiliary variables they replicate known population totals. There are several papers addressing the calibration technique for estimation in sample surveys. Deville & Särndal (1992) proposed linear form for the calibration weighting with multivariate auxiliary information. Kott (2006) considered calibration estimation to correct for coverage errors and unit non-response (Quasi-randomization). Montanari (2005) discussed calibration estimator in a neural network mode. Särndal & Lundström (2008) discussed non-response bias for choosing auxiliary information.

In these papers, it is usually assumed that auxiliary variables are fully recorded without missing values. For example, Särndal & Lundström (2005) proposed star vector and moon vector, which are defined as information available at the population level and the sample level, respectively. Unfortunately, in reality, auxiliary information is not that ideal. Missing values occur in all types of data, also in auxiliary variables records.

Missing values can be replaced by imputations and then treated as any other auxiliary variable. However, this depends on the way the imputations are derived. The cases when missing values are replace by a constant, zero say, and the case when imputations are derived from a regression model estimated on response set information are different. This paper addresses the issue of the effects of how imputations of auxiliary variables are derived. Using simulation, the bias of the calibration estimator is studied under regression imputation where the regression imputation model is estimated using register, sample, and response set information respectively.

2 Calibration Estimator

2.1 Definition

Consider a finite population with N elements $U{=}1, 2, \ldots, N$, in which y_k is a target variable and $x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})'$ is a full-recorded auxiliary vector. A probability sample s with sample size n is selected from U by a probability sampling design $p(s)$. When non-response occurs, only a subset of the sample $r \in s$ is answered, where the size of response set is denoted as n_r. To describe the random response

mechanism, $q(r|s)$ is denoted as the conditional response distribution, and the probability of a response of element k given its selection to a sample is denoted $\theta_k = Pr(k \in r | k \in s)$.

To estimate the population total $Y = \sum_U y_k$, the calibration estimator $\hat{Y}_w = \sum_r w_k y_k$ uses calibrated weights w_k subject to the constraint $\sum_r w_k x_k = X$. The weights w_k can be defined in different ways obeying the constraint. Särndal & Lundström (2005) defined the weights using the system $w_k = d_k v_k$, $v_k = 1 + \lambda_r \mathbf{x}_k$, and $\lambda_r = (X - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}$.

2.2 Auxiliary information

Särndal & Lundström (2005) defined three different cases depending on the accessible auxiliary information. In this paper, we will look into two of them. The two cases are

InfoU. Information is available at the level of the population U such that

- the population total $\sum_U \mathbf{x}_k^\star$ is known;
- for every $k \in r$, the value of $\mathbf{x}_k^\star$ is known.

InfoS. Information is available at the level of the sample s such that

- for every $k \in s$, the value of $\mathbf{x}_k^\circ$ is known but $\sum_U \mathbf{x}_k^\circ$ is unknown.

Consider the case that missing values occur in auxiliary variable x_k as well. Imputation is a frequently-used method to allocate artificial values for the missing items. Little & Rubin (1987) regarded imputation as a general and flexible method for handling missing-data problem but with pitfall, such as substantial bias, and summarized different sorts of imputation methods to construct the substitutes.

With imputed values, auxiliary variable will be denoted as

$$x_{\bullet k} = \begin{cases} x_k & \text{for } k \in r_x \\ x_k(\hat{\delta}) & \text{for } k \in U - r_x \end{cases} \tag{1}$$

here r_x is the subset of the population U where x_k is available, and $x_k(\hat{\delta})$ is the imputed value based on the parameter $\hat{\delta}$ which is derived from the register information.

Consider InfoU with $\mathbf{x}_k = \mathbf{x}_k^\star = (1, x_{\bullet k})'$ with the information input $\mathbf{X}_{\bullet k} = (N, \sum_U x_{\bullet k})$, where the calibration estimator for target variable y becomes

$$\hat{Y}_w = N\bar{y}_r + (\sum_U x_{\bullet k} - N\bar{x}_r) * B_r$$

where

$$B_r = \frac{\sum_r d_k (x_{\bullet k} - \bar{x}_r)(y_k - \bar{y}_r)}{\sum_r d_k (x_{\bullet k} - \bar{x}_r)^2} \tag{2}$$

$$\bar{y}_r = \frac{\sum_r d_k y_k}{\sum_r d_k} \tag{3}$$

$$\bar{x}_r = \frac{\sum_r d_k x_{\bullet k}}{\sum_r d_k} \tag{4}$$

Consider InfoS with $\mathbf{x}_k = \mathbf{x}_k^\circ = (1, x_{\bullet k})'$ with the information input $\mathbf{X}_{\bullet k} = (\hat{N}, \sum_s d_k x_{\bullet k})$, where $\hat{N} = \sum_s d_k$. The calibration estimator for target variable y then becomes

$$\hat{Y}_w = \hat{N}\bar{y}_r + (\sum_s d_k x_{\bullet k} - \hat{N}\bar{x}_r) * B_r$$

where B_r, $\bar{y}_r$ and $\bar{x}_r$ are the same as in equation (2), (3) and (4).

2.3 Nearbias

A central issue regarding the effects of nonresponse is estimation bias. Consider an auxiliary vector x_k satisfying $\mu' \mathbf{x}_{\bullet k} = 1$ for all k. Then Särndal & Lundström (2005) shows

$$Nearbias(\hat{Y}_w) = (\sum_U \mathbf{x}_{\bullet k})'(\mathbf{B}_{U;\theta} - \mathbf{B}_U) \tag{5}$$

in which

$$\mathbf{B}_{U;\theta} = (\sum_U \theta_k \mathbf{x}_{\bullet k} \mathbf{x}'_{\bullet k})^{-1} (\sum_U \theta_k \mathbf{x}_{\bullet k} y_k)$$

and

$$\mathbf{B}_U = (\sum_U \mathbf{x}_{\bullet k} \mathbf{x}'_{\bullet k})^{-1} (\sum_U \mathbf{x}_{\bullet k} y_k)$$

Consider the case that missing value occurs in auxiliary variable x_k, x_k becomes $x_{\bullet k}$ as described in equation (1). If all the missing items in x_k is artificially imputed as 0, i.e., $x_{\bullet k} = 0$ when $k \in U - r_x$, the formula of nearbias will be the same as equation (5) with no extra components.

If we consider another case in which regression imputation is utilized for missing values in x_k, i.e., x_k is replaced by $x_{\bullet k}$ and $x_{\bullet k} = x_k(\hat{\delta})$ when $k \in U - r_x$, where $\hat{\delta}$ is the estimate of coefficient derived from the register system. The formula of nearbias in this case will stay the same as equation (5).

According to accessible register information at hand, the estimate of $\hat{\delta}$ could possibly derived from population level, response level or sample level. When $\hat{\delta}$ is obtained from register information only, equation (5) will keep valid through all the cases.

In next section, a simulation study will be conducted on how bias of calibration estimator changes when different imputation for missing values in x_k is used. Missing values in x_k will be imputed by linear regression model $x_k = \mathbf{u}'_k \boldsymbol{\delta} + \varepsilon_k$ and $\boldsymbol{\delta}$ is estimated based on sample level or response level.

3 Simulation Study

The effect of imputation on the calibration estimator bias is here studied by simulation. To simulate a population with 100000 units, the following procedures are performed.

1. x_k is generated from a standard normal distribution $N(0,1)$.
2. error term ξ_1 and ξ_2 are independently generated from $N(0,1)$ distribution.
3. u_k is generated by $u_k = \alpha + \beta * x_k + \rho_1 * \xi_{1k}$.
4. y_k is generated by $y_k = \tau + \eta * x_k + \rho_2 * \xi_{2k}$.

The coefficients β, η, ρ_1 and ρ_2 are used to control the coefficient of determination R^2 between y and x, and x and u respectively.

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 the occurance of non-response in y_k and missing values of x_k.

Case I non-response in y_k occurs with constant probability such that $\theta_k = \theta$ for all $k \in U$, and missing vaule of x_k occurs with constant probability such that $\Pr(x_k \text{ is missing in register system}) = \vartheta_k = \vartheta$ for all $k \in U$.

Case II missing vaule of x_k occurs with constant probability such that $\vartheta_k = \vartheta$ for all $k \in U$, but non-response in y_k occurs with varying probability, i.e., θ_k changes for $k \in U$.

Case III non-response in y_k occurs with constant probability such that $\theta_k = \theta$ for all $k \in U$, but missing vaule of x_k occurs with varying probability, i.e., ϑ_k changes for all $k \in U$.

Case IV non-response in y_k and missing value in x_k both occur with varying probability, i.e., both θ_k and ϑ_k change for all $k \in U$.

In Case II and IV above, y_k is divided into three groups, with response rate

$$\theta_k = \begin{cases} 10\% & \text{when } y > 8 \\ 35\% & \text{when } y < 0 \\ 65\% & \text{when } 0 \leq y \leq 8 \end{cases}$$

Similarly, in Case III and IV, x_k is divided into ten groups, with response rate

$$\vartheta_k = \begin{cases} 45\% & \text{when } x < -1.28 \\ 50\% & \text{when } -1.28 \leq x < -0.84 \\ 55\% & \text{when } -0.84 \leq x < -0.52 \\ 60\% & \text{when } -0.52 \leq x < -0.25 \\ 65\% & \text{when } -0.25 \leq x < 0 \\ 75\% & \text{when } 0 \leq x < 0.25 \\ 80\% & \text{when } 0.25 \leq x < 0.52 \\ 85\% & \text{when } 0.52 \leq x < 0.84 \\ 90\% & \text{when } 0.84 \leq x < 1.28 \\ 95\% & \text{when } x \geq 1.28 \end{cases}$$

The regression imputation will be utilized to make up for the missing values in auxiliary variable x_k. In this stage, imputation will be proceeded based on the linear model

$$x_k = \mathbf{u}_k'\boldsymbol{\delta} + \varepsilon_k = \delta_1 + \delta_2 u_k + \varepsilon_k$$

where u is a full-recorded variable from the register system. The estimator of $\boldsymbol{\delta}$ is $\hat{\boldsymbol{\delta}} = (\sum_A \mathbf{u}_k\mathbf{u}_k')^{-1}(\sum_A \mathbf{u}_k x_k)$, where A is the set of items used for performing the linear regression. The following three kinds of collection of objects (i.e., A) are considered.

Imputation 1 $A = U_x$, where U_x is the whole population of variable x, which means imputation regression will be run based on all the available values of x_k in the population.

Imputation 2 $A = r_x = U_x \cap r$, where imputation regression will be executed based on the available values of x_k in the population where y_k is responsed.

Imputation 3 $A = s_x = U_x \cap s$, where imputation regression will be performed based on the available values of x_k in the sample.

Replicating the simulation for 3000 times, the expectation of the calibration estimator is estimated by $E(\hat{Y}_w) = \sum_{i=1}^{3000} \hat{Y}_{w_i}/3000$ and the bias is estimated with $Bias(\hat{Y}_w) = E(\hat{Y}_w) - Y$. As a benchmark, estimates of the bias of the calibration estimator with full-recorded auxiliary variable x_k is shown in Table 1. The bias in the case II/IV is notably larger than in case I/III.

Bias estimates for the calibration estimator with missing values of the auxiliary variable are reported in tables 2 - 5. In Table 2, the case with $R^2(y, x) = R^2(x, u) = 50\%$ is considered. From the table it is seen that the bias estimates are small for cases I and III, while they are large in cases II and IV. A surprising observation is that bias estimates are in large unaffected by the level of information used for estimation of the regression model used for imputation. Also, bias is in large unaffected by the level of information used in the calibration estimator. However, compared to the beachmark in Table 1, the biases in tables 2 - 5 are larger in general.

In the following tables, bias estimates are reported for different cases of strengths in the relation between y and x, and between x and u. Compared with Table 2, Table 3 reports results where $R^2(y,x)$ is increased to 0.8. It is observed that the results in the two tables are comparable. There are only minor changes in the bias estimates.

In Table 4, bias estimates in the case when $R^2(x,u)$ is increased to 0.8, compared with Table 2, is reported. Again, the results are comparable with those of Table 1 with only small difference in bias estimates. Finally, Table 5 reports results when $R^2(x,u)$ is decreased to 0.26 and it is seen that the reported estimates are comparable with those of tables 2 - 4.

Table 6 reports a case when the auxiliary variable x_k is chi-square distributed instead of normal as in tables 1 - 5. The levels of the bias estimates are different (see Table 6) but the general pattern observed in tables 2 - 5 is also observed here. Biases are negligible in cases I and III, while modest in cases II and IV. The level of the bias is not dependent of the level of information used in the calibration, and finally, the bias is largely unaffected by the level of information used for the estimation of the regression relation used as imputation model.

4 Conclusion and Discussion

It is shown in all the cases of the simulation study that the bias of calibration estimator differs slightly between InfoU case and InofS case, which is also stated in Särndal & Lundström (2005). The bias estimates in Case I/III are always close to 0, implying that the calibration estimator is nearly unbiased when the response rate θ_k is constant and not related to the value of y_k. The bias estimates in Case II/IV, however, are affected by θ_k being related to the value of y_k.

The simulation study also shows that imputation of auxiliary information with coefficient $\hat{\delta}$ derived from different levels of information, i.e., register, sample and response set give negligible differences on bias estimates. This implies that when missing values in auxiliary information need to be imputed, the level of information used for imputation has no essential effect on the bias of the calibration estimator. The results are here derived by simulation and it is desirable to derive more formal and general results. One suggestion for further studies is to consider the asymptotic properties of Y_w, where asymptotic limits $f(\hat{\delta})$ are utilized. It is expected that bias expressions similar to equation (5) can be derived using asymptotics.

References

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

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

Little, R. & Rubin, D. (1987). *Statistical analysis with missing data.* John Wiley & Sons.

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

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.

A Table: simulation result

Table 1: Bias in Normal case when $R^2(y,x)$=$R^2(x,u)$=50%

	Case I	Case II
InfoU	139.28	-11884.92
InfoS	350.78	-11616.12

Note: x_k is full-recorded. $\sum_U y_k$=500915.62

Table 2: Bias in Normal case when $R^2(y,x)$=$R^2(x,u)$=50%

		Case I	Case II	Case III	Case IV
InfoU	Imputation 1	-100.82	-13307.32	167.32	-11509.29
	Imputation 2	-105.22	-13345.59	137.73	-11275.61
	Imputation 3	-102.62	-13343.70	128.98	-11547.56
InfoS	Imputation 1	76.68	-13074.37	455.78	-11223.00
	Imputation 2	84.08	-13091.86	385.32	-10943.10
	Imputation 3	19.18	-13078.22	457.31	-11271.17

Note: $\sum_U y_k$=500915.62

Table 3: Bias in Normal case when $R^2(y,x)$=80% and $R^2(x,u)$=50%

		Case I	Case II	Case III	Case IV
InfoU	Imputation 1	-170.51	-13714.31	97.41	-10606.63
	Imputation 2	-179.55	-13556.80	67.67	-9100.94
	Imputation 3	-158.29	-13693.62	82.72	-10590.23
InfoS	Imputation 1	77.15	-13466.74	378.16	-10386.30
	Imputation 2	37.92	-13290.21	326.84	-8877.21
	Imputation 3	49.62	-13504.88	332.76	-10401.46

Note: $\sum_U y_k$=500967.42

Table 4: Bias in Normal Case when $R^2(y,x)$=50% and $R^2(x,u)$=80%

		Case I	Case II	Case III	Case IV
	Imputation 1	16.13	-12462.12	109.29	-11716.29
InfoU	Imputation 2	18.85	-12490.73	100.82	-11596.08
	Imputation 3	-7.76	-12504.85	65.81	-11755.04
	Imputation 1	275.90	-12216.29	382.02	-11438.96
InfoS	Imputation 2	193.84	-12231.55	296.71	-11291.59
	Imputation 3	246.78	-12257.83	338.56	-11483.31
					Note: $\sum_U y_k$=500915.62

Table 5: Bias in Normal Case when $R^2(y,x)$=50% and $R^2(x,u)$=26%

		Case I	Case II	Case III	Case IV
	Imputation 1	-173.81	-13913.94	231.92	-11403.91
InfoU	Imputation 2	-177.67	-13947.07	196.40	-11148.80
	Imputation 3	-183.65	-13953.12	200.92	-11432.49
	Imputation 1	77.05	-13698.89	522.33	-11111.83
InfoS	Imputation 2	0.48	-13705.81	443.75	-10801.25
	Imputation 3	41.84	-13743.01	478.42	-11165.83
					Note: $\sum_U y_k$=500915.62

Table 6: Bias in chi-square case when$R^2(y,x)$=$R^2(x,u)$=85%

		Case I	Case II	Case III	Case IV
	Imputation 1	-1575	-21848	-1408	-22164
InfoU	Imputation 2	-1649	-21534	-1484	-20986
	Imputation 3	-1668	-21898	-1527	-22184
	Imputation 1	-1171	-21368	-1012	-21743
InfoS	Imputation 2	-1078	-20908	-920	-20406
	Imputation 3	-1027	-21325	-925	-21668
					Note: $\sum_U y_k$=1099883.12