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

Lisha Wang

Swedish Business School Örebro University

Aug 28, 2012

▲日▼ ▲□▼ ▲ □▼ ▲ □▼ ■ ● ● ●

◆□▶ ◆□▶ ◆三▶ ◆三▶ - 三 - のへぐ



- Calibration Estimator
- Auxiliary Information
- Nearbias
- Simulation

Interest and Target

- The calibration approach is suggested in the literature for estimation in sample survey under non-response given access to suitable auxiliary information.
- Missing values occur in auxiliary variables records.
- To investigate how imputation of auxiliary information based on different levels of register information affect the calibration estimator.

Calibration Estimator

- Population total: $Y = \sum_U y_k$
- Calibration estimator: $\hat{Y}_w = \sum_r w_k y_k$
- 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. For example, Särndal & Lundström (2005) defined the weights using the system $w_k = d_k v_k$, $v_k = 1 + \lambda_r \mathbf{x}_k$, and $\lambda_r = (X \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}'_k)^{-1}$.

・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・

Auxiliary Information

Two cases will be considered in this paper.

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^* 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° is known but $\sum_U \mathbf{x}_k^{\circ}$ is unknown.

With imputed values, auxiliary variable will be denoted as

$$x_{\bullet k} = \begin{cases} x_k & \text{for } k \in r_x \\ \delta_0 + \delta_1 * u_k & \text{for } k \in U - r_x \end{cases}$$

here r_x is the subset of the population U where x_k is available, and u_k is available for all $k \in U$. Three different cases will be discussed, where estimation is based on poplation set U_x , sample set $s_x = U_x \cap s$ and response set $r_x = U_x \cap r$.

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. Särndal & Lundström (2005) shows

$$Nearbias(\hat{Y}_w) = (\sum_U \mathbf{x}_{\bullet k})'(\mathbf{B}_{U;\theta} - \mathbf{B}_U)$$
(1)

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

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

Simulation Study

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 $y_k = \tau + \eta * x_k + \rho_2 * \xi_{2k}$.

Patterns of occurance of non-response in y_k and missing in x_k .

	θ_k	ϑ_{k}
Case I	Constant	Constant
Case II	Varying	Constant
Case III	Constant	Varying
Case IV	Varying	Varying

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.

<ロト < 団 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < 国 > < B > < B > < B > < 国 > < 国 > < 国 > < B > < B > < B > < 国 > < 国 > < 国 > < B > < B > < B > < B > < B > < B > < B > < B > < B > < B > < B > < B > < B

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	$x_k \text{ is fu} \\ \sum_U y_k^{=}$	III-recorded. =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
	Imputation 1	-100.82	-13307.32	167.32	-11509.29
InfoU	Imputation 2	-105.22	-13345.59	137.73	-11275.61
	Imputation 3	-102.62	-13343.70	128.98	-11547.56
	Imputation 1	76.68	-13074.37	455.78	-11223.00
InfoS	Imputation 2	84.08	-13091.86	385.32	-10943.10
	Imputation 3	19.18	-13078.22	457.31	-11271.17
			Not	e: $\sum_U y_k =$	=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
	Imputation 1	-170.51	-13714.31	97.41	-10606.63
InfoU	Imputation 2	-179.55	-13556.80	67.67	-9100.94
	Imputation 3	-158.29	-13693.62	82.72	-10590.23
	Imputation 1	77.15	-13466.74	378.16	-10386.30
InfoS	Imputation 2	37.92	-13290.21	326.84	-8877.21
	Imputation 3	49.62	-13504.88	332.76	-10401.46
			Not	e: $\sum_U y_k$ =	=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
			Not	e: $\sum_U y_k =$	=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
			Not	e: $\sum_U y_k =$	=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
InfoU	Imputation 1	-1575	-21848	-1408	-22164
	Imputation 2	-1649	-21534	-1484	-20986
	Imputation 3	-1668	-21898	-1527	-22184
InfoS	Imputation 1	-1171	-21368	-1012	-21743
	Imputation 2	-1078	-20908	-920	-20406
	Imputation 3	-1027	-21325	-925	-21668
			Note: 2	$\sum_{U} y_k = 10$	99883.12

Expected results:

- Increased bias under Case II/IV.
- Imputation only slightly increase bias in some cases.

The important empirical conclusion: The effects of using different levels of auxiliary information (population, sample, response set) for estimation of imputation model are negligible.

▲日▼ ▲□▼ ▲ □▼ ▲ □▼ ■ ● ● ●

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

Thank you for your attention!

◆□▶ ◆□▶ ◆三▶ ◆三▶ - 三 - のへぐ