Supplementary Material of "Matrix completion under complex survey sampling"

Xiaojun Mao · Zhonglei Wang · Shu Yang

The Supplementary Material contains an additional simulation study under single-stage sampling and analysis for the longitudinal trend of the NHANES Questionnaire Data.

S1 Single-stage sampling

Based on the generated finite population U_N, the following sampling designs are considered:

I Poisson sampling (POI) with inclusion probability $\pi_i = ns_i(\sum_{i=1}^N s_i)^{-1}$, where $s_i > 0$ is a size measure of the ith subject, and the generation of s_i is discussed later. Specifically, for $i = 1, \ldots, N$, a sampling indicator I_i is generated by a Bernoulli distribution with success probability π_i.

II Simple random sampling without replacement (SRS) with sample size n.

III Probability-proportional-to-size sampling (PPS) with size measure s_i. That is, a sample of size n is selected independently from the finite population U_N with replacement, and the selection probability of the ith subject is proportional to its size measure s_i.

The first two sampling designs are without replacement, but the third one is not. We consider the third one to test the robustness of the proposed method. Instead of the Horvitz-Thompson estimator, we use the Hansen-Hurwitz estimator (Hansen and Hurwitz; 1943) for the third sampling design. We consider two scenarios for the sampling procedure. One is non-informative sampling with $s_i = d^{-1}\sum_{j=1}^d x_{ij} + e_i + 1$,

Xiaojun Mao
School of Mathematical Sciences, Ministry of Education Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai, 200240, P.R.C.

Zhonglei Wang
Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University, Xiamen, Fujian, 361005, P.R.C.
E-mail: wangzl@xmu.edu.cn

Shu Yang
Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695, U.S.A.

Table 1 Summary of MSE for different estimation methods under single-stage sampling under Scenario I. Values out side of the parenthesis are the average of MSE among 500 items, and the ones inside correspond to the standard errors.

		Non informative		Informative	
		$n = 200$	$n = 500$	$n = 200$	$n = 500$
PPS	HDI	24.66 (10.35)	24.41 (10.31)	24.74 (10.33)	24.50 (10.42)
	MI	10.98 (3.17)	3.92 (1.48)	11.73 (3.95)	4.00 (1.58)
	IPW	11.39 (4.53)	4.55 (1.84)	11.58 (4.62)	4.58 (1.83)
	AIPWLR	12.37 (4.99)	4.63 (1.88)	12.38 (5.22)	4.59 (1.86)
	AIPWMC	7.53 (2.82)	2.96 (1.07)	7.62 (2.76)	2.88 (1.05)
	Full	5.74 (2.07)	2.33 (0.87)	5.81 (2.06)	2.24 (0.83)
POI	HDI	24.78 (11.10)	22.57 (9.93)	24.66 (10.74)	22.67 (10.70)
	MI	10.96 (4.09)	3.93 (1.25)	12.92 (4.80)	4.65 (1.76)
	IPW	11.99 (4.67)	4.83 (1.86)	12.22 (4.84)	4.80 (1.87)
	AIPWLR	12.42 (5.08)	4.64 (1.82)	12.87 (5.11)	4.71 (1.88)
	AIPWMC	7.61 (2.76)	2.88 (1.02)	7.60 (2.96)	2.93 (1.12)
	Full	6.21 (2.31)	2.51 (0.93)	6.41 (2.34)	2.48 (0.90)
SRS	HDI	25.04 (10.63)	24.56 (10.20)	24.82 (10.91)	24.38 (10.46)
	MI	10.44 (3.03)	3.47 (1.06)	10.52 (3.48)	3.53 (1.11)
	IPW	10.52 (4.11)	4.01 (1.63)	10.40 (4.17)	4.10 (1.63)
	AIPWLR	11.42 (4.70)	4.14 (1.69)	11.38 (4.63)	4.20 (1.69)
	AIPWMC	7.01 (2.51)	2.61 (0.93)	6.88 (2.60)	2.60 (0.95)
	Full	5.19 (1.99)	2.00 (0.76)	5.17 (1.93)	2.00 (0.76)

where $e_i \sim \text{Exp}(1)$. The other is informative sampling with $s_i = 7^{-1}\sum_{j=1}^{7} y_{ij} - m_s + 1$, where $m_s = \min\{7^{-1}\sum_{j=1}^{7} y_{ij} : i = 1, \ldots, N\}$. Two different sample sizes are considered, $n = 200$ and $n = 500$.

We conduct 1 000 Monte Carlo simulations for each estimation method in Section **??**. Tables 1–2 summarizes MSE_j for $j = 1, \ldots, L$ under the two scenarios. The MSE of the hot deck imputation is much larger than other estimators regardless of the sampling design and sample size. The MSE of the inverse probability weighting method and the AIPW estimator using linear regression are similar, and they are larger than the multiple imputation and the AIPW estimator using the proposed estimator. Under the three sampling designs, the AIPW estimator using linear regression performs worse than the inverse probability weighting method, since it does not correctly specify the regression model in (**??**). Although the multiple imputation performs slightly better than the inverse probability weighting method and AIPW estimator using linear regression, it is not preferable due to the computation complexity, especially when the number of items is large. The AIPW estimator using the proposed method has smaller MSE than other alternatives, and it is almost as efficient as the Horvitz-Thompson estimator using full data when the sample size is large. Although we do not investigate theoretical properties the proposed method under Scenario II, the AIPW estimator using the propose method is robust in this case. Besides, the AIPW estimator using the propose method still outperforms its competitors. However, since the random errors are correlated under Scenario II, the performance of the AIPW estimator using the propose method is slightly undermined.

Table 2 Summary of MSE for different estimation methods under single-stage sampling under Scenario II. Values out side of the parenthesis are the average of MSE among 500 items, and the ones inside correspond to standard errors.

		Non informative		Informative	
		$n = 200$	$n = 500$	$n = 200$	$n = 500$
PPS	HDI	23.25 (10.32)	22.81 (10.06)	23.37 (10.69)	22.98 (10.23)
	MI	9.41 (3.25)	3.45 (1.41)	10.85 (3.80)	3.72 (1.60)
	IPW	10.81 (4.49)	4.37 (1.82)	10.90 (4.61)	4.28 (1.87)
	AIPWLR	11.63 (4.90)	4.40 (1.86)	11.77 (5.08)	4.37 (1.84)
	AIPWMC	8.19 (3.32)	3.05 (1.22)	7.77 (2.95)	3.01 (1.28)
	Full	5.43 (2.11)	2.21 (0.87)	5.44 (2.17)	2.13 (0.88)
POI	HDI	22.87 (10.48)	20.84 (10.41)	22.86 (10.75)	21.17 (10.21)
	MI	10.53 (3.80)	3.90 (1.28)	12.12 (4.27)	4.30 (1.68)
	IPW	12.26 (4.69)	4.93 (1.90)	12.37 (4.81)	4.81 (1.84)
	AIPWLR	11.93 (5.03)	4.35 (1.79)	12.04 (5.00)	4.39 (1.81)
	AIPWMC	8.12 (3.22)	3.11 (1.25)	8.16 (3.28)	3.01 (1.32)
	Full	6.63 (2.43)	2.75 (0.95)	6.82 (2.42)	2.63 (0.95)
SRS	HDI	23.46 (10.66)	22.92 (10.14)	23.74 (10.58)	22.97 (10.43)
	MI	9.35 (3.11)	3.05 (1.09)	9.47 (3.01)	3.07 (1.00)
	IPW	9.84 (4.04)	3.81 (1.62)	9.87 (4.05)	3.82 (1.67)
	AIPWLR	10.60 (4.60)	3.91 (1.65)	10.71 (4.69)	3.91 (1.70)
	AIPWMC	8.14 (3.65)	2.69 (1.08)	8.37 (3.82)	2.75 (1.11)
	Full	4.80 (2.02)	1.89 (0.76)	4.94 (1.98)	1.89 (0.77)

We also test the performance of the plug-in variance estimator in terms of relative bias,

$$RB_j = \frac{\widehat{V}_{p,j} - \widehat{V}_{s,j}}{\widehat{V}_{s,j}} \quad (j = 1, \ldots, L),$$

where $\widehat{V}_{p,j} = 1000^{-1} \sum_{m=1}^{1000} \widehat{V}_{p,j}^{(m)}$, $\widehat{V}_{p,j}^{(m)}$ is the plug-in variance estimator based on the mth Monte Carlo sample, and $\widehat{V}_{s,j}$ is the corresponding Monte Carlo variance. Table 3 summarizes the simulation results. From the simulation results, we can conclude that the relative bias of the plug-in variance estimator is less than 10% generally. Besides, the plug-in variance estimator is conservative for the proposed AIPW estimator. This observation makes sense since the imputed value for each missing cell may be slightly biased. We have also checked the relative bias under other sampling designs, and similar conclusion can be drawn. The plug-in variance estimator performs similarly for both scenarios.

S2 Longitudinal trend of the NHANES Questionnaire Data

In this section, we analyze the longitudinal trend for the three questions in Table **??** using the NHANES Questionnaire Data since 1999. The question about self-reported greatest weight has remained in the Questionnaire since the 1999–2000 cycle, but the other two were added in the Questionnaire in the 2007–2008 cycle.

Figure S1 shows the longitudinal trend for the self-reported greatest weight in pounds. The self-reported greatest weight increases from 1999 to 2014 gradually, but it remains stable in the 2015–2016 cycle. Compared with the 1999–2000 cycle, people

Table 3 Summary of the relative bias of the plug-in variance estimator under single-stage sampling and different scenarios. Values out side of the parenthesis are the average of the relative bias among 500 items, and the ones inside are the corresponding standard error.

		Non informative		Informative	
		$n = 200$	$n = 500$	$n = 200$	$n = 500$
PPS	I	0.10 (0.06)	0.08 (0.06)	0.06 (0.06)	0.11 (0.07)
	II	0.09 (0.06)	0.08 (0.05)	0.06 (0.06)	0.09 (0.06)
POI	I	0.05 (0.05)	0.04 (0.04)	0.07 (0.06)	0.08 (0.06)
	II	0.07 (0.06)	0.06 (0.05)	0.06 (0.05)	0.07 (0.05)
SRS	I	0.06 (0.06)	0.06 (0.05)	0.06 (0.06)	0.06 (0.06)
	II	0.08 (0.07)	0.06 (0.05)	0.07 (0.07)	0.05 (0.05)

gained about 15 pounds in terms of greatest weights up to the 2015–2016 cycle. Thus, more attention should be paid to the obesity problems for the people in the age group from 20 to 59.

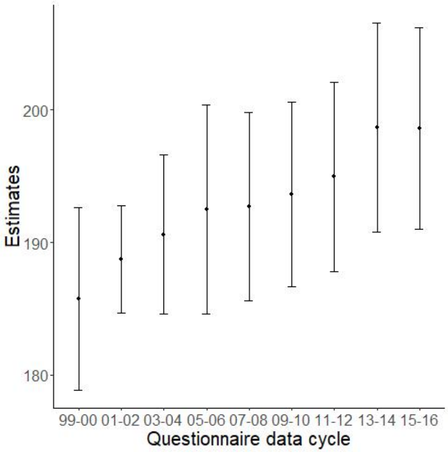

Fig. S1 Longitudinal trend of the self-reported greatest weight in pounds since 1999. The dot corresponds to the estimated population mean of the age group 20–59, and the vertical segment to the two standard error area around the estimated mean. “99-00” stands for the “1999–2000 cycle”, ..., and “15-16” for the most recent “2015–2016 cycle”.

Figure S2 shows the estimation results for the second and the third questions since the 2007–2008 cycle. For the money spent on eating out, we can conclude that people tended to spend less on eating out during 2009 and 2010, and this phenomenon may due to the economic crisis from 2007 to 2009. After that, the money on eating out increased dramatically in the 2011–2012 cycle, and it was even larger than that in the 2007–2008 cycle. On the other hand, the monthly family income has been stable since 2007. The economic crisis did not have significant influence on the estimation of monthly family income, and one possible reason is that, instead of using the actual amount, the survey uses the raw 12 levels to characterize the monthly family income.

However, we can still see a slight decrease in the monthly family income during the 2009–2010 cycle.

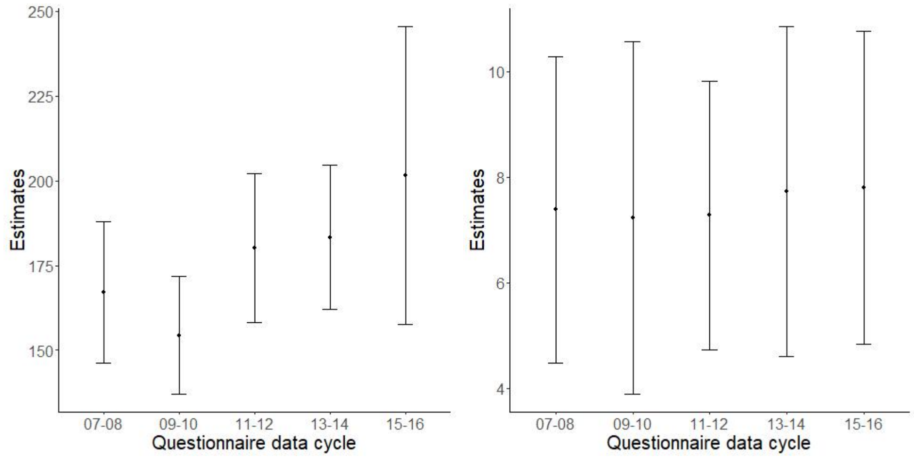

Fig. S2 Longitudinal trend of the money spent on eating out (left) and the monthly family income (right) since the 2007–2008 cycle. The dot corresponds to the estimated population mean, and the vertical segment to the two standard error area around the estimated mean. "07-08" stands for the "2007–2008 cycle", ..., and "15-16" for the most recent "2015–2016 cycle".

References

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations, *The Annals of Mathematical Statistics* **14**(4): 333–362.