Combining Non-probability and Probability Survey Samples Through Mass Imputation

Jae-Kwang Kim [1]

Iowa State University & KAIST

[1] Joint work with Seho Park, Yilin Chen, and Changbao Wu

Outline

1. Introduction
2. Proposed method
3. Variance estimation
4. Replication variance estimation
5. A real data application
6. Conclusion

1. Introduction

- We are interested in combining information from two samples, one with probability sampling and the other with non-probability sampling (such as voluntary sample).
- We observe X from the probability sample and observe (X, Y) from the non-probability sample.
- The sampling mechanism for sample B is unknown.

Table: Data Structure

Data	X	Y	Representativeness
A	✓		Yes
B	✓	✓	No

Mass Imputation

- Rivers (2007) idea
  1. Use X to find the nearest neighbor for each unit $i \in A$.
  2. Compute
  $$\hat{\theta} = \sum_{i \in A} w_i y_i^*$$
  where w_i is the sampling weight of unit $i \in A$ and y_i^* is the imputed value of y_i using nearest neighbor imputation taken from sample B.
- Based on MAR (missing at random) assumption
$$f(y \mid x, \delta = 1) = f(y \mid x)$$
where
$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{if } i \notin B. \end{cases}$$

Mass Imputation for data integration

- Basic Steps
  1. Use sample B to estimate the conditional distribution $f(y \mid x)$.
  2. Predict y-values for sample A using the estimated conditional distribution.

- If $f(y \mid x)$ is correctly specified and MAR holds, then the mass imputation estimator is unbiased.

- Question: How to estimate the variance of mass imputation estimator?

2. Proposed method

- Assume
$$Y_i = m(\mathbf{x}_i; \boldsymbol{\beta}_0) + e_i \tag{1}$$
for some $\boldsymbol{\beta}_0$ with a known function $m(\cdot)$, with $E(e_i \mid \mathbf{x}_i) = 0$.
- We assume that $\hat{\boldsymbol{\beta}}$ is the unique solution to
$$\hat{U}(\boldsymbol{\beta}) \equiv \sum_{i \in B} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\} h(\mathbf{x}_i; \boldsymbol{\beta}) = 0 \tag{2}$$
for some p-dimensional vector $h(\mathbf{x}_i; \boldsymbol{\beta})$.
- Under MAR, the solution to (2) is consistent for $\boldsymbol{\beta}_0$.
- Thus, we use the observations in sample B to obtain $\hat{\boldsymbol{\beta}}$ and then use it to construct $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ for all $i \in A$.

Theorem

Suppose that model (1) and MAR condition hold. Under some regularity conditions, the mass imputation estimator

$$\bar{y}_I = \frac{1}{N} \sum_{i \in A} w_i m(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \tag{3}$$

satisfies

$$n_B^{1/2} \left(\bar{y}_I - \tilde{y}_I(\boldsymbol{\beta}_0) \right) = o_p(1) \tag{4}$$

where $o_p(1)$ denotes convergence in probability, n_B is the size of sample B,

$$\tilde{y}_I(\boldsymbol{\beta}) = N^{-1} \sum_{i \in A} w_i m(\mathbf{x}_i; \boldsymbol{\beta}) + n_B^{-1} \sum_{i \in B} \left\{ y_i - m(\mathbf{x}_i; \boldsymbol{\beta}) \right\} h(\mathbf{x}_i; \boldsymbol{\beta})' \mathbf{c}^*,$$

$$\mathbf{c}^* = \left[n_B^{-1} \sum_{i \in B} \dot{m}(\mathbf{x}_i; \boldsymbol{\beta}_0) h'(\mathbf{x}_i; \boldsymbol{\beta}_0) \right]^{-1} N^{-1} \sum_{i=1}^{N} \dot{m}(\mathbf{x}_i; \boldsymbol{\beta}_0),$$

$\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ in (1), and $\dot{m}(\mathbf{x}; \boldsymbol{\beta}) = \partial m(\mathbf{x}; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$.

Also,

$$E\{\tilde{y}_I(\boldsymbol{\beta}_0) - \bar{y}_N\} = 0, \tag{5}$$

and

$$\begin{aligned} V\left\{\tilde{y}_I(\boldsymbol{\beta}_0) - \bar{y}_N\right\} &= V\left\{N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - N^{-1}\sum_{i\in U} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\} \\ &+ E\left[n_B^{-2}\sum_{i\in B} E\left(e_i^2 \mid \mathbf{x}_i\right)\left\{h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^*\right\}^2\right], \end{aligned} \tag{6}$$

where $e_i = y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)$.

Example

- Under the special case of linear model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i$$

with $e_i \sim (0, \sigma_e^2)$, we can use $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}} = \left(\sum_{i \in B} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in B} \mathbf{x}_i y_i$ to construct regression mass imputation.

- If sample A is obtained from the simple random sampling, the asymptotic variance in (6) reduces to

$$V\left(\hat{\theta}_{I,reg}\right) = \frac{1}{n_A} \beta_1^2 \sigma_x^2 + \frac{1}{n_B} \sigma_e^2 + E\left\{ \frac{(\bar{x}_N - \bar{x}_B)^2}{\sum_{i \in B} (x_i - \bar{x}_B)^2} \right\} \sigma_e^2.$$

- If sample B were an independent random sample of size n_B, then the third term would of order $O(n_B^{-2})$ and is negligible. However, as sample B is a non-probability sample, the third term is not negligible.

3. Variance estimation

- For variance estimation of the mass imputation estimator (3), we have only to estimate the variance of the linearized estimator $\tilde{y}_I(\boldsymbol{\beta}_0)$ in (4). Since the variance formula can be written as

$$V\left\{\tilde{y}_I(\boldsymbol{\beta}_0) - \bar{y}_N\right\} = V_A + V_B$$

where

$$\begin{aligned} V_A &= V\left\{N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - N^{-1}\sum_{i\in U} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\} \\ V_B &= E\left[n_B^{-2}\sum_{i\in B} E\left(e_i^2 \mid x_i\right)\left\{h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^*\right\}^2\right], \end{aligned}$$

we can estimate V_A and V_B separately.

- To estimate $\hat{V}_A$, we can use

$$\hat{V}_A = \frac{1}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} w_i m(\mathbf{x}_i; \hat{\beta}) w_j m(\mathbf{x}_j; \hat{\beta}).$$

  where π_{ij} is the joint inclusion probability for unit i and j, which is assumed to be positive.

- To estimate V_B, we can use

$$\hat{V}_B = n_B^{-2} \sum_{i \in B} \hat{e}_i^2 \left\{ h(\mathbf{x}_i; \hat{\beta})' \hat{\mathbf{c}}^* \right\}^2, \tag{7}$$

  where $\hat{e}_i = y_i - m(\mathbf{x}_i; \hat{\beta})$ and

$$\hat{\mathbf{c}}^* = \left[n_B^{-1} \sum_{i \in B} \dot{m}(\mathbf{x}_i; \beta_0) h'(\mathbf{x}_i; \beta_0) \right]^{-1} N^{-1} \sum_{i \in A} w_i \dot{m}(\mathbf{x}_i; \beta_0)$$

- Hence, the variance of $\bar{y}_{I,reg}$ can be estimated by $\hat{V}(\bar{y}_{I,reg}) = \hat{V}_A + \hat{V}_B$.

Remark

- If $n_A/n_B = o(1)$, then V_B is smaller order than V_A and total variance is dominated by V_A. Otherwise, the two variances both contribute to the total variance. If sample B is a big data, n_B is huge and V_B can be safely ignored.
- However, to compute $\hat{V}_B$ in (7), we use individual observations of (x_i, y_i) in sample B, which is not necessarily available when only sample A with mass imputation is released to the public.
- Note that the goal of mass imputation is to produce a representative sample with synthetic observations using sample B as a training data. Once the mass imputation is performed, the training data is no longer necessary in computing the point estimation.
- So, it is desirable to develop a variance estimation method that does not require access to the sample B observations.

4. Replication variance estimation

- We consider a bootstrap method for variance estimation that creates replicated synthetic data $\{\hat{y}_i^{(k)}, i \in A\}$ corresponding to each set of bootstrap weights $\{w_i^{(k)}, i \in A\}$ associated with sample A only.
- This method enables the user to correctly estimate the variance of the mass imputation estimator $\bar{y}_{I,reg}$ without access to the training data $\{(y_i, x_i) : i \in B\}$ from sample B. The data file will contain additional columns of $\{y_i^{(k)} : i \in A\}$ associated with the columns of weights $\{w_i^{(k)}; i \in A\}$ $(k = 1, \cdots, L)$, where L is the number of replicates created from sample A.

- Kim and Rao (2012) developed a replication method for survey integration when sample B is also a probability sample.
  1. Obtain $\hat{\boldsymbol{\beta}}^{(k)}$, the k-th replicate of $\hat{\boldsymbol{\beta}}$, by solving the same estimating equation for $\boldsymbol{\beta}$ using the replication weights for sample B.
  2. The k-th replicate of the mass imputation estimator $\bar{y}_{I,reg} = \sum_{i \in A} w_i \hat{y}_i$ is

$$\bar{y}_{I,reg}^{(k)} = \sum_{i \in A} w_i^{(k)} \hat{y}_i^{(k)}$$

  where $\hat{y}_i^{(k)} = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(k)})$.
- How to modify the method of Kim and Rao (2012) when sample B is a non-probability sample?

- In order to develop a valid bootstrap method for mass imputation estimator $\bar{y}_{I,reg}$ in (3), it is critical to develop a valid bootstrap method for estimating $V(\hat{\beta})$ when $\hat{\beta}$ is computed from (2). Note that, under assumption (1) and MAR, we can obtain

$$V(\hat{\boldsymbol{\beta}}) \doteq J^{-1}\Omega J^{-1'} \tag{8}$$

where $J = E\left\{n_B^{-1}\sum_{i \in B} \dot{\mathbf{m}}_i \mathbf{h}_i'\right\}$ and $\Omega = E\left\{n_B^{-2}\sum_{i \in B} E(e_i^2 \mid \mathbf{x})\mathbf{h}_i\mathbf{h}_i'\right\}$ with $\dot{\mathbf{m}}_i = \dot{m}(\mathbf{x}_i; \boldsymbol{\beta}_0)$ and $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\beta}_0)$.

- In (8), the reference distribution is the joint distribution of the superpopulation model and the unknown sampling mechanism for sample B.

- Interestingly, the variance formula in (8) is exactly equal to the variance of $\hat{\beta}$ under simple random sampling (SRS) for sample B.
- That is, even though the sample design for sample B is not SRS, its effect on the variance of $\hat{\beta}$ is essentially equal to that under SRS.
- Therefore, we can safely apply the bootstrap method under SRS for variance estimation of $\hat{\beta}$, even though the true sampling mechanism for sample B is not SRS.

Thus, the proposed bootstrap method can be described as in the following steps:

1. Treating sample B as a simple random sample, generate the k-th bootstrap sample from sample B to compute $\hat{\beta}^{(k)}$, the k-th bootstrap replicate of $\hat{\beta}$, using the same estimation formula (2) applied to the bootstrap sample.
2. Using $\hat{\beta}^{(k)}$ from [Step 1], compute $\hat{y}_i = m(\mathbf{x}_i; \hat{\beta}^{(k)})$ for each $i \in A$. Using the $\hat{y}_i^{(k)}$, we obtain the replicated mass imputation estimator
$$\bar{y}_{I,reg}^{(k)} = \sum_{i \in A} w_i^{(k)} \hat{y}_i^{(k)}.$$
3. The resulting bootstrap variance estimator of $\bar{y}_{I,reg}$ is then
$$\hat{V}_b(\bar{y}_{I,reg}) = L^{-1} \sum_{k=1}^{L} \left(\bar{y}_{I,reg}^{(k)} - \bar{y}_{I,reg} \right)^2 . \tag{9}$$

5. A real data application

- Pew Research Center (PRC) data in 2015: a non-probability sample data of size $n = 9,301$ with 56 variables, provided by eight different vendors with unknown sampling and data collection strategies.
- The PRC dataset aims to study the relation between people and community. We choose 9 variables, among them 8 are binary and 1 is continuous, as response variables in our analysis.
- We consider two probability samples with common auxiliary variables. The first is the Behavioral Risk Factor Surveillance System (BRFSS) survey data and the second is the Volunteer Supplement survey data from the Current Population Survey (CPS), both collected in 2015.

Comparison of covariates from three datasets

Table: Estimated Population Mean of Covariates from the Three Samples

		$\hat{X}_{PRC}$	$\hat{X}_{BRFSS}$	$\hat{X}_{CPS}$
Age category	<30	0.183	0.209	0.212
	>=30,<50	0.326	0.333	0.336
	>=50,<70	0.387	0.327	0.326
	>=70	0.104	0.131	0.126
Gender	Female	0.544	0.513	0.518
Race	White only	0.823	0.750	0.786
Race	Black only	0.088	0.126	0.125
Origin	Hispanic/Latino	0.093	0.165	0.156
Region	Northeast	0.200	0.177	0.180
Region	South	0.275	0.383	0.373
Region	West	0.299	0.232	0.235
Marital status	Married	0.503	0.508	0.528
Employment	Working	0.521	0.566	0.589

Comparison of covariates from three datasets (Cont'd)

		$\hat{X}_{PRC}$	$\hat{X}_{BRFSS}$	$\hat{X}_{CPS}$
Education	High school or less	0.216	0.427	0.407
Education	Bachelor's degree and above	0.416	0.263	0.309
Education	Bachelor's degree	0.221	NA	0.198
Education	Postgraduate	0.195	NA	0.111
Household	Presence of child in household	0.289	0.368	NA
Household	Home ownership	0.654	0.672	NA
Health	Smoke everyday	0.157	0.115	NA
Health	Smoke never	0.798	0.833	NA
Financial status	No money to see doctors	0.207	0.133	NA
Financial status	Having medical insurance	0.891	0.878	NA
Financial status	Household income $<$ 20K	0.161	NA	0.153
Financial status	Household income $>$100K	0.199	NA	0.233
Volunteer works	Volunteered	0.510	NA	0.248

Remark

- There are noticeable differences between the naive estimates from the PRC sample and the estimates from the two probability samples for covariates such as Origin (Hispanic/Latino), Education (High school or less), Household (with children), Health (Smoking) and Volunteer works.
- It is a strong evidence that the PRC dataset is not a representative sample for the population.

Mass Imputation using a single set of common covariates

Table: Estimated Population Mean Using A Single Set of Common Covariates

Binary Response y		$\hat{\theta}$	$v_l(\times 10^{-5})$	$v_b(\times 10^{-5})$
Talked with neighbours frequently	PRC	0.461		
	BRFSS	0.457	4.323	4.187
	CPS	0.458	4.195	4.055
Tended to trust neighbours	PRC	0.590		
	BRFSS	0.553	4.200	4.221
	CPS	0.557	4.070	4.044
Expressed opinions at a government level	PRC	0.265		
	BRFSS	0.240	2.858	2.881
	CPS	0.243	2.878	2.925
Voted local elections	PRC	0.750		
	BRFSS	0.707	3.687	3.498
	CPS	0.716	3.447	3.258

Mass Imputation using a single set of common covariates

Table: Estimated Population Mean Using A Single Set of Common Covariates

Binary Response y		$\hat{\theta}$	$v_l(\times 10^{-5})$	$v_b(\times 10^{-5})$
Participated in school groups	PRC	0.210		
	BRFSS	0.200	2.599	2.615
	CPS	0.206	2.602	2.607
Participated in service organizations	PRC	0.141		
	BRFSS	0.133	1.910	1.886
	CPS	0.135	1.922	1.930
Participated in sports organizations	PRC	0.168		
	BRFSS	0.165	2.278	2.221
	CPS	0.170	2.262	2.257
No money to buy food	PRC	0.251		
	BRFSS	0.289	3.681	3.562
	CPS	0.286	3.516	3.457

Mass Imputation using a single set of common covariates

Table: Estimated Population Mean Using A Single Set of Common Covariates

Continuous Response y		$\hat{\theta}_I$	$v_I(\times 10^{-2})$	$v_b(\times 10^{-2})$
Days had at least	PRC	5.301		
one drink last month	BRFSS	4.931	1.010	0.996
	CPS	4.986	0.978	0.952

Discussion

- There are substantial discrepancies between the mass imputation estimator and the naive estimator in most cases.
- The mass imputation estimates obtained with two different probability samples are comparable for all cases.
- The two variance estimators obtained by using the linearization and the bootstrap methods generally agree with each other.

6. Conclusion

- Using a non-probability sample data as a training set for prediction, we can implement mass imputation for survey sample data and obtain unbiased estimates under some reasonably weak assumptions.
- Statistical inference including variance estimation under the parametric mass imputation is developed.
- Nonparametric model can be used for the imputation model, which will be presented elsewhere.
- Machine learning algorithm can also be used for mass imputation, but its variance estimation is more challenging.
- A promising area of research.

REFERENCES

Kim, J. K. and J. N. K. Rao (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika 99*, 85–100.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*.