Calibrated Multiple Imputation Under Informative Sampling Using Bayesian Nonparametric Model

Huaiyu Zang

Dept. of Mathematical Sciences
University of Cincinnati

Joint work with Dr. Hang J. Kim

August, 2017

Contents

1. Introduction

2. Non-parametric MI Using Design Information

3. Bayesian Calibrated Imputation

4. Concluding Remarks and Future Studies

Missing data problem

- Survey data suffers from **missing values**
  - Missing data problem occurs frequently in sample surveys.
  - Unit nonresponse: a sampled individual does not respond to the entire survey.

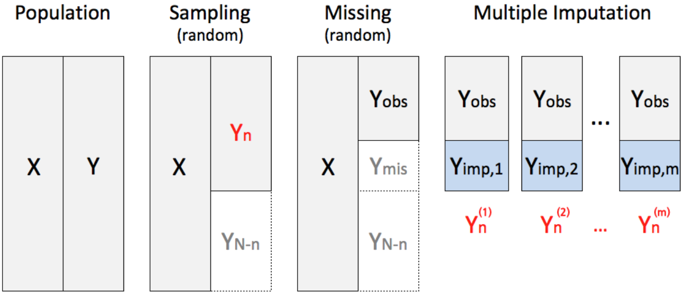
    e.g., attrition: did not respond at all to follow-up survey.
  - Item nonresponse: individuals who do not respond to particular questions.

    e.g., someone answered most of the survey questions, but left a few blank.
- Using only observed values can cause problems for statistical inferences.
  - Biased when the missing data mechanism is not ignorable.
  - Not efficient even when the missing data mechanism is ignorable.

Multiple Imputation (D. Rubin 1978)

- Multiple Imputation (MI) is popular for handling item nonresponse in survey sampling.
- MI is a model-based, Monte Carlo technique in which **missing values** are replaced by multiply imputed datasets
- MI steps:
  1. Find missing values, i.e., $Y_{\text{collect}} = (Y_{\text{obs}}, Y_{\text{mis}})$
  2. Generate multiple versions of imputation $Y_{\text{imp}}^{(1)}$, ..., $Y_{\text{imp}}^{(m)}$ from the predictive distribution of Y_{mis} given Y_{obs}, $f(Y_{\text{mis}}|Y_{\text{obs}})$
     $\rightarrow$ Multiple completed datasets, $Y_{\text{cmp}}^{(l)} = (Y_{\text{obs}}, Y_{\text{imp}}^{(l)})$, $l = 1, \ldots, m$
  3. Using standard statistical software, calculate m completed-data estimates from $Y_{\text{cmp}}^{(l)}$ $l = 1, \ldots, m$,
  4. Make inferences by using **combining rules** with the m completed-data estimates

Multiple Imputation (Combining Rule)



- Estimate $\hat{\theta}^{(l)}$ and its variance $u^{(l)}$ for each $\mathrm{Y}_n^{(l)}$
- $\bar{\theta}_m = \frac{1}{m}\sum_{l=1}^{m}\hat{\theta}^{(l)}$, $\bar{u}_m = \frac{1}{m}\sum_{l=1}^{m}u^{(l)}$, $b_m = \frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{\theta}^{(l)} - \bar{\theta}_m\right)^2$

Multiple Imputation (Combining Rule)

- Under proper imputation,

$$\theta - \bar{\theta}_m \sim t_\nu(0, T_m) \qquad \text{where} \qquad T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

- Consider $m = \infty$,

$$Var(\theta|Y_{obs}) = E(Var(\theta|Y_{obs}, Y_{mis})) + Var(E(\theta|Y_{obs}, Y_{mis})) = \bar{u}_\infty^2 + b_\infty$$

- Benefits of using MI:
  - Accounting for the imputation uncertainty
  - Performing MI can be based on different imputation strategies (eg: sequential modeling, joint modeling).

Survey sampling

- Surveys: gathering data on a usually small subset of population.
  - Measuring the characteristic of a population
  - Aiming to estimate a finite population parameter using a sample with size **n** without measuring the whole population with **N** units
- Sample is selected according to a given sampling design.
  - Population index set $\mathcal{U} = (1, \ldots, N)$ and sample index $\mathcal{S}$ is a subset of $\mathcal{U}$ with size n.
  - Common sampling designs: simple random sampling(SRS), stratified sampling, and probability proportional-to-size (PPS) sampling.

Sampling design

- Sampling design $P(I_k|Z)$:
  - I_k: sample selection indicators. $I_k = 1$ if unit k is selected, $I_k = 0$, otherwise.
  - Z: design variables; e.g, demographic data (age, sex)
    - Known to the sampler before the sample is drawn
    - Determining the sampling probability
  - I_k depends only on Z
- π_k: inclusion probability attached to unit k
  - $E(I_k|Z) = \pi_k$ for $k \in \mathcal{U}$
  - In SRS, every sampling units has equal inclusion probability, $\pi_k = n/N$.
  - Causing the units to be over- or under-represented under unequal probability sampling
- Given $I_k = 1$, collecting survey variables y_k from the sampled units
  - $(y_{1k}, \ldots, y_{pk})$ denote p survey variables (e.g., income, age and social economic status, etc)

Sampling weights

- The base weight
  - Inverse of its inclusion probability $d_k = \frac{1}{\pi_k}$
  - Representing the number of elements of sample in the population
  - Compensating for differential representation due to unequal selection probabilities
- Applying the base weight to the survey variable y leads to a design-unbiased estimator (Horvitz and Thompson, 1952).
  - Population total: $\theta = \sum_{k=1}^{N} y_k$
  - Estimated total: $\hat{\theta} = \sum_{k \in \mathcal{S}} d_k y_k = \sum_{k=1}^{N} d_k y_k I_k$
  - Treating y as fixed and taking expectations with respect to I, we have $E_I(\hat{\theta}) = \theta$

Weighting system

- The first stage: base weight d_k
- The second stage: weights adjusted for nonresponse
- The third stage: final (or calibrated) weights w_k
  - Survey estimates agree with published margin or known (or estimated) totals available from external sources.
  - This process is called calibration (external consistency).
  - The final weight w_k is published with survey variables, i.e., $(y_{k1}, \ldots, y_{kp}, w_k)$
- Additional stage: weight trimming or smoothing

Business establishment microdata

- Business establishment survey/census
  - U.S. Economic Census, Annual Survey of Manufactures, Structural Business Survey (Netherlands)
- **Microdata**: detailed data on units (establishments)
  - accessible only by vetted staff/researchers due to the privacy concerns
  - summarized to produce aggregate-level **tabular data**
- **Microdata** are important for secondary analyses

  e.g., building blocks of GDP; to allocate federal funds; to find new business opportunity, used for high impact economics studies
  - Statistical agencies release microdata for a few survey products

Business establishment microdata (cont.)

- Data characteristics
  - Most of items are continuous variables
  - Marginal distributions are often skewed to the right (monetary values)
  - High correlations among variables
  - To check the accuracy of reported values, the statistical agency uses edit rules.
    - e.g., Total Employees ≥ 0
    - e.g., Total Salary / Total Employees $\leq$ \$1M
  - Suffer from **missing values**
    e.g., 2007 US CM reported 20%-40% missing values [1]

[1] T. K. White, J. P. Reiter, and A. Petrin (2012)

Design-based vs. Model-based inference (Little 2004)

- Design-based inference: y are fixed and inference is based on probability distribution of sample selection I.
  - Automatically taking into account features of the survey design and yielding reliable results for large samples
  - Generally asymptotic and potentially inefficient (Basu's elephant example).
- Model-based inference: y are assumed to come from a statistical model and inference considers the joint distribution for y and I.
  - Yielding robust inference for small sample estimation
  - Equal or more efficient when model assumptions are satisfied

Research question

- How can we handle missing value problems in business establishment mircodata?
- How does sampling design impact the imputation process? How can the imputation model account for design information?
- For external consistency, how can we account for calibration in the imputation process?
- How to construct imputation model whose results are design consistent and model consistent?

Research plan

- Study goals: **1)** Imputation for business microdata **2)** under complex sample design **3)** which ensures that sums are calibrated to the known total.
- Applying our proposed method to 2017 Economic Census to generate the clean version of the microdata.

Contents

1. Introduction

2. Non-parametric MI Using Design Information

3. Bayesian Calibrated Imputation

4. Concluding Remarks and Future Studies

UNIVERSITY OF Cincinnati

Informative sampling

- Informative sampling: the sample distribution not equal to the population distribution

$$P(y_i \in B|x_i, I_i = 1) \neq P(y_i \in B|x_i)$$

  for any measurable set B.
- Leading to inclusion probabilities that are correlated with the survey variable of interest after conditioning on the covariates x_i
  - $P(I_i|y_i, x_i, \phi) \neq P(I_i|x_i, \phi)$, where ϕ denotes parameter of inclusion indicator.
  - Inducing dependence among the selected observations: y_i are not independent and identically distributed for $i \in S$ (Bonnery et al 2012)

Stratified sampling

- Partitioning the population into strata (mutually exclusive and meaningful groups) and then sampling independently in each of the strata.
  - The strata are formed by the auxiliary information (e.g, gender, Industrial types) prior to sampling.
- Benefits
  - Useful if we want to include participants of various minority groups such as race or religion
  - Reducing sampling error if the population within stratum is homogenous and between stratum is heterogonous

PPS sampling

- The probability that a particular sampling unit, which will be chosen in the sample, is proportional to some known size (design) variable Z such as population size or geographic size.
  - E.g, size variable is the number of positions in each section when surveying employees in a company by department
- Offering the possibility of decreased variability of estimators of means and totals if the size variable is correlated with the variable of interest. (Cochran, 1977, p. 255)

Why the design variable is required?

- Design variables provide the information of sampling design.
- A simple motivating example (to weight or not to weight?)

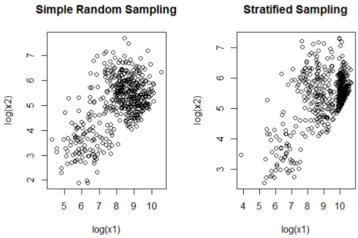


Figure: The figure on the left is the sample selected by SRS. The figure on the right is the sample selected by stratified random sampling when strata variable is formed by x_1.

Motivating example

Population

		x_2=0	x_2=1
Z=0	x_1=0	200	100
Z=1	x_1=0	200	10
	x_1=1	100	100

Sample

		x_2=0	x_2=1
Z=0	x_1=0	2	1
Z=1	x_1=0	200	10
	x_1=1	100	100

Unadjusted by weight: $\hat{p}(x_{1,mis} = 0 \mid x_2 = 1) = \frac{1+10}{1+110} \approx 0.10$

Adjusted by weight: $\hat{p}(x_{1,mis} = 0 \mid x_2 = 1) = \frac{1*100+10*1}{100+110} \approx 0.52$

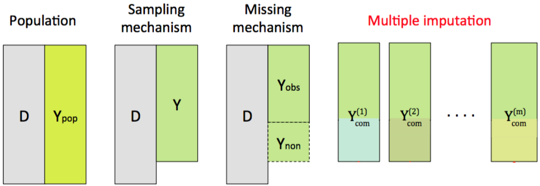
Figure: For Z equal to 0, the units are selected by 1/100 (base weights equal to 100). For Z equal to 1, all units are selected (base weights equal to 1). The red color denote x_1 is missing with true value equal to 0.

Incorporating design information

- Ignoring the sampling design can have severe effects on the inference process if the design is not ignorable. Sampling weights can be used to protect against informative designs and misspecified models (Pfeffermann 1993).
- Noting the challenges in weighting and modeling, Gelman suggested using hierarchical regression, combined with post-stratification (Gelman 2007).
- Si incorporated externally-supplied sampling weights using Bayesian hierarchical model in a fully model-based framework (Si et al 2015).
  - Obtaining better estimates of sparse cells by borrowing strength from the other non-sparse cells

Why the design variable is required?

- Design information is important, however, no consensus exists as to the best way to incorporate design information into the imputation model (Gelman 2007).
- Current practice for MI assumes that the sampling design is non-informative.



- We are interested in investigating the role of informative sampling on the imputation model.

Some existing methods of imputation model

- "Hot Deck" Imputation
  - Replacing missing value with an observed response from a "similar" unit
  - Drawbacks: 1. Underestimation of uncertainty; 2. Constraints are not satisfied.
- Dirichlet Process(DP) Mixture Model Imputation (Kim et al 2014, 2015)
  - The DP Gaussian mixture is a famous form of nonparametric Bayesian models.
  - Benefits:
    1. Preserving the (irregular, skewed) joint density
    2. Maintaining the multivariate relationships
    3. Satisfying the edit constraints
  - However, the performance of DP Mixture Model Imputation has not formally been investigated in the informative sampling setting. In addition, it does not account for calibration constraints.

Hot Deck Imputation

- Widely used by government statistical agencies and survey organizations
- Choosing the "Nearest Neighbor" (NN) Hot Deck imputation in the simulation study.
  - The nearest neighbor are determined using the Mahalanobis distance based on the maximum deviation metric.
  - Similar to what Generalized Edit and Imputation System (GEIS) of Statistics Canada choose for imputation (Andridge and Little 2010)

DP Mixture Model Imputation

- Model notations:
  - Let $\boldsymbol{y}_i$ be a p-dimensional vector $(y_{i1}, \ldots, y_{ip})$ and $\boldsymbol{y}$ be $n \times p$ matrix $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$.
  - Let μ_k is the mean vector of y_i in the kth mixture component and $\mu = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)$.
  - Let $\boldsymbol{\Sigma}_k$ is the the covariance matrix of y_i in the kth mixture component and $\Sigma = (\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$.
  - $\pi = (\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K)$: the mixture component weights, where $\sum_{k=1}^{K} \boldsymbol{\pi}_k = 1$.
  - Let $\theta_y = (\mu, \Sigma, \pi)$.
  - $\mathcal{X}$: the feasible region of $\boldsymbol{y}$ defined only by ratio edit and range restrictions.

DP Mixture Model Imputation

- Model specification:

$$f(\boldsymbol{y}|\theta_y) \propto \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k \, \mathrm{N}\left(\boldsymbol{y}_i; \mu_k, \Sigma_k\right) I\left[\boldsymbol{y}_i \in \mathcal{X}\right] \right)$$

where π_k is estimated by a truncated version of the stick-breaking construction for the **Dirichlet process (DP)**

$$\begin{aligned} \pi_k &= \nu_k \prod_{g<k} (1-\nu_g) \quad \text{for } k = 1, \ldots, K & (1) \\ \nu_k|\alpha &\sim \mathrm{Beta}(1, \alpha) \quad \text{for } k = 1, \ldots, K-1; \;\; \nu_K = 1, & (2) \\ \alpha &\sim \mathrm{Gamma}(a_\alpha, b_\alpha). & (3) \end{aligned}$$

Simulation assuming stratified sampling

1.1 Generating the population

- Mixture of multivariate normal distribution with $K = 3$.
- $N = 10,000$ with 4 variables $y_i = (y_{i1}, \ldots, y_{i4})$
- Introducing edits. e.g., $L_1 \leq y_1 \leq U_1$ and $L_{12} \leq y_1/y_2 \leq U_{12}$
- Similar to Annual Survey of Manufacture

1.2 Stratified into two groups

- Strata variable is formed by size variable. $\text{Size}_i = 500 + 0.1 y_{i1} + N(0, 10^2)$ for each unit i.
- It can be interpreted as historical records of survey variables in the previous year.
- The largest 500 out of 10,000 values of size variable vs. the remaining 9500 values of size variable.
- Stratum 1 with $N_1 = 500$ and Stratum 2 with $N_2 = 9500$

Simulation assuming stratified sampling

2. Sample
   - For the 1st stratum, all the units were selected with sample size $n_1 = 500$; $\pi_{1j} = \frac{n_1}{N_1} = \frac{500}{500} = 1$.
   - For the 2nd stratum, the units were sampled by SRS with sample size $n_2 = 500$; $\pi_{2j} = \frac{n_2}{N_2} = \frac{500}{9500} = \frac{1}{19}$.
   - Sample size $n = n_1 + n_2 = 1,000$
   - e.g., Large companies selected with certainty; Medium companies selected by simple random sampling.
3. Observed sample
   - Introducing item missing: 500 out of 1,000 records were missing.
   - Randomly selecting 200 units with one item missing, 200 units with two item missing, 100 units with three item missing
4. This process of sampling and creating missingness was repeated 500 times.

Simulation assuming stratified sampling

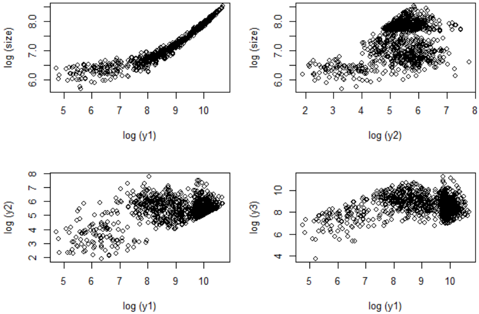


Figure: The scatter-plots show the pairwise relationship between log (size) and log ($\boldsymbol{y}$).The top left figure showed the strong positive relationship between the size variable and y_1. We can clearly see the bounds, which is the ratio edit, for y_2 and y_1 in the bottom left figure.

Simulation assuming stratified sampling

- Compare
  1. DP imputation method w/wo design variable
  2. Hot deck imputation method w/wo design variable
  3. Running DP imputation separately in each stratum
- For DP imputation, a total of 10 multiply imputed data sets were created. The resulting inference is obtained using the combining rules of Rubin (1987)

Results: absolute relative bias

Absolute relative bias ($\times 100$): $\frac{1}{R}\sum_{r=1}^{R} |\hat{\bar{y}}^r - \bar{y}_{\mathcal{U}}|/\bar{y}_{\mathcal{U}}$, where R is the number of repetition and $\bar{y}_{\mathcal{U}}$ is true population mean. (averaging over 500 repeated sim.)

	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Before deletion sample	2.27	3.25	3.26	3.98
Hot deck wo size	21.83	3.66	3.89	4.61
DP wo size	21.36	3.71	3.72	**4.44**
Hot deck w size	2.31	3.81	3.62	4.94
DP w size	**2.29**	3.75	3.60	4.48
Running two DPs	2.57	**3.56**	**3.60**	4.58

The closest estimates to that of before deletion sample are bolded.

Results: relative root MSE

Relative root MSE ($\times 100$): $\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{y}}^r - \bar{y}_{\mathcal{U}}\right)^2}/\bar{y}_{\mathcal{U}}$ (averaging over 500 repeated sim.)

	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Before deletion sample	2.84	4.01	4.02	4.96
Hot deck wo size	22.34	4.65	4.82	5.86
DP wo size	21.73	4.71	4.59	**5.52**
Hot deck w size	2.90	4.73	4.58	6.18
DP w size	**2.86**	4.67	4.46	5.64
Running two DPs	3.17	**4.47**	**4.46**	5.73

Results: 95% nominal CI coverage

95% CI coverage ($\times$100): $\frac{1}{R}\sum_{r=1}^{R} \boldsymbol{I}_r\left[L(\hat{\bar{y}}) < \bar{y}_{\mathcal{U}} < U(\hat{\bar{y}})\right]$ (averaging over 500 repeated sim.)

	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Before deletion sample	95.4	95.6	95.4	94.2
Hot deck wo size	0.2	89.6	89.2	89.8
DP wo size	0.0	**97.2**	93.4	**94.8**
Hot deck w size	95.0	88.4	90.8	87.8
DP w size	**95.6**	98.4	95.0	95.4
Running two DPs	96.6	97.4	**95.4**	95.0

Results: bivariate plot

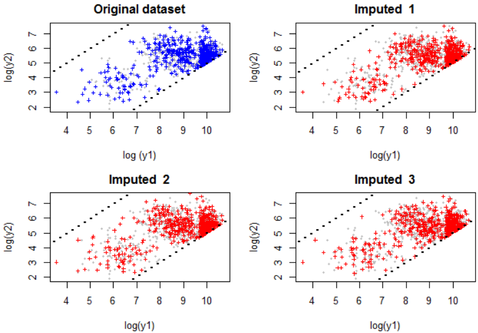


Figure: The original dataset and three imputed (DP w size) datasets. The gray points represent true records. The blue cross on the left top represent the records subject to missing. The red cross on the other three plots represent the corresponding imputed records.

Results: bivariate plot

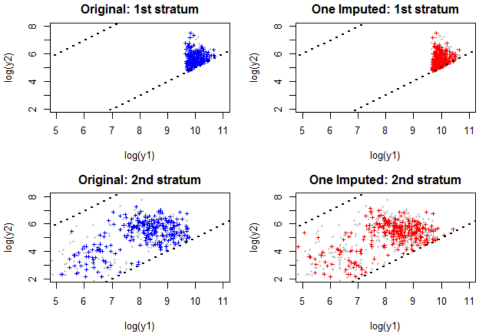


Figure: Running two DPs for each stratum

Discussions: stratified sampling

- For the size variable highly correlated with survey outcome (y_1)
  - Excluding the size variable imputations overestimate the population mean and demonstrated large bias and thus severe under-coverage for $\bar{y}_1$.
  - The imputation method with the size variable outperforms the method without the size variable.
- For the size variable weakly correlated with survey outcomes (y_2, y_3, y_4)
  - There is not significant difference (within simulation error) between imputation with the size variable and without the size variable.

Discussions: stratified sampling

- As compared with Hot Deck imputation, DP imputation does not underestimated the variance, thus producing the confidence interval close to nominal 95%.
- Running two DP for each stratum has similar performance with DP imputation with the size variable.
  - Both methods accurately estimate the population mean
  - Both have small relative root MSE
- Even though imputing separately in each stratum is an ideal method, however, it may be impractical.
  - Requires fitting different models to each stratum, thus it will be computationally more expensive when there are many strata.
  - The extreme case: PPS sampling in which each unit is its own stratum.

Simulation assuming the PPS sampling

1. Generating the population
   - Mixture of multivariate normal distribution with $K = 3$.
   - $N = 100,000$ with 4 variables $y_i = (y_{i1}, \ldots, y_{i4})$
   - Introducing edits. e.g., $L_1 \leq y_1 \leq U_1$ and $L_{12} \leq y_1/y_2 \leq U_{12}$
2. Sample
   - $n = 1000$ units were selected with probability proportional-to-size variable under sampling without replacement.
   - Size variable is defined as $\text{Size}_i = 500 + 0.1 y_{i1} + N(0, 10^2)$. The coefficient of correlation between log (size) and $\log(\boldsymbol{y}_1)$ is 0.9.
3. Observed sample
   - Introducing item missing: 500 out of 1,000 records were missing.
4. This process of sampling and creating missingness was repeated 500 times.

Simulation assuming the PPS sampling

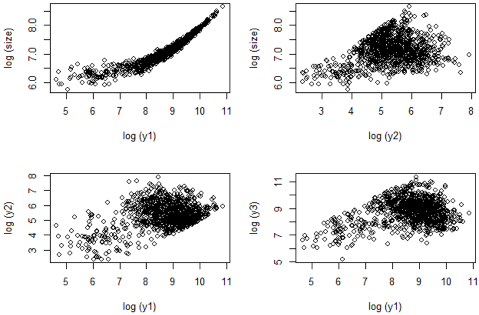


Figure: The scatter-plots show the pairwise relationship between log (size) and log ($\boldsymbol{y}$).

Simulation assuming the PPS sampling

- As we did for stratified sampling, DP multiple imputation was carried out both with size variable and without size variable under PPS sampling design.
- The variances were estimated using design-based Hansen-Hurwitz variance estimator.
- The Hot Deck imputation with and without size variable were employed for comparison.

PPS Results: absolute relative bias

Absolute relative bias (×100) for the mean of x: $\frac{1}{R}\sum_{r=1}^{R}\left|\hat{\bar{y}}^r - \bar{y}_{\mathcal{U}}\right|/\bar{y}_{\mathcal{U}}$ (averaging over 500 repeated sim.)

	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Before deletion sample	1.07	2.87	2.65	3.73
Hot deck wo size	8.14	3.32	3.14	4.36
DP wo size	7.77	3.30	2.94	**4.10**
Hot deck w size	1.13	3.30	3.10	4.28
DP w size	**1.10**	**3.25**	**2.92**	4.19

PPS Results: relative root MSE

Relative root MSE($\times$100) for the mean of x: $\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{y}}^r - \bar{y}_{\mathcal{U}}\right)^2}/\bar{y}_{\mathcal{U}}$ (averaging over 500 repeated sim.)

	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Before deletion sample	1.34	3.64	3.35	4.77
Hot deck wo size	8.61	4.25	3.95	5.46
DP wo size	8.08	4.26	3.68	**5.13**
Hot deck w size	1.42	4.20	3.87	5.34
DP w size	**1.38**	**4.20**	**3.64**	5.31

PPS Results: 95% nominal CI coverage

95% CI coverage ($\times 100$) for the mean of x: $\frac{1}{R}\sum_{r=1}^{R} \boldsymbol{I}_r \left[L(\hat{\bar{y}}) < \bar{y}_{\mathcal{U}} < U(\hat{\bar{y}})\right]$

(averaging over 500 repeated sim.) [1]

	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Before deletion sample	94.6	95.0	95.6	92.8
Hot deck wo size	8.0	90.2	91.0	90.4
DP wo size	21.8	**95.0**	**95.0**	94.8
Hot deck w size	93.2	88.8	90.8	**93.2**
DP w size	**94.8**	95.4	94.6	95.6

[1] The variances were estimated using design-based Hansen-Hurwitz variance estimator.

PPS Results: bivariate plot

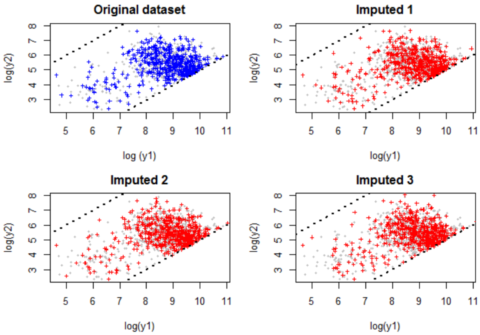


Figure: The original dataset and three imputed (DP w size) datasets

Contents

1. Introduction

2. Non-parametric MI Using Design Information

3. Bayesian Calibrated Imputation

4. Concluding Remarks and Future Studies

Calibration

- Records are often calibrated so that
  - calculated item totals or weighted totals in microdata match to the published margin (or a known total)
- The reasons for using calibration are three-fold (Hazziza and Beaumont 2017):
  - to force consistency of certain survey estimates to known population quantities;
  - to reduce nonsampling errors such as nonresponse errors and coverage errors;
  - to improve the precision of estimates.

Method 1: calibrated weighting

- One way of seeking a calibration weighting system is such that (Deville and Sarndal 1992)
  - final weights w_k are "as close as possible" to the base weights d_k
  - while satisfying the calibration constraints.

$$\sum_{k \in \mathcal{S}} w_k y_k = \mathbf{Y}$$

Method 2: calibrated imputation

- Another calibration approach is to fix the base weights d_k and to adjust imputed missing item y_i so that the calibration constraints are satisfied $\sum_{k \in \mathcal{S}} d_k y_k = \mathbf{Y}$.
- This process is known as calibrated imputation.
  - Has recently been studied by some researchers (e.g, Pannekoek et al. 2013 and Ton De Waal et al. 2017)
  - Benefits:
    1. The differences between estimated total and known total are caused by the systematic errors in the imputed values.
    2. Ensuring that the adjustment does not affect the estimates of all other variables.
  - Drawback:
    1. Based on an optimization-based approach, thus not measuring uncertainty introduced by imputation process.

Suggested method

- Adding the calibration prior to DP mixture model:

$$f(\boldsymbol{y}|\vec{\xi}^2, \vec{\mathbf{Y}}) \propto \exp\left\{-\sum_{j=1}^{p}\left[\frac{1}{2\xi_j^2}\Big(\sum_{i=1}^{n} d_i y_{i,j} - \mathbf{Y}_j\Big)^2\right]\right\}$$

- d_i denotes base weight for unit i.
- $\vec{\mathbf{Y}} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_j)$ are the known totals.
- $\vec{\xi}^2 = (\xi_1^2, \ldots, \xi_j^2)$ control the degree of calibration (tuning parameter).
- p denotes the number of calibration constrains.
- The prior distribution is interpreted as a *penalty function* in frequentist approaches (as in Bayesian Lasso vs. classical Lasso)

Bayesian Calibrated Imputation

Bayesian Calibrated Imputation:

$$f(\boldsymbol{y}|\theta_y, \vec{\xi}^2, \vec{\mathbf{Y}}) \propto \{\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k \mathrm{N}\left(\boldsymbol{y}_i; \mu_k, \Sigma_k\right) I\left[\boldsymbol{y}_i \in \mathcal{X}\right]\} \times f(\boldsymbol{y}|\vec{\xi}^2, \vec{\mathbf{Y}})$$

- Challenge:
  - Independence assumption is violated due to calibration constraints.
  - Small ξ_j^2 may result in low acceptance probability.

Bayesian Calibrated Imputation

Summaries of Bayesian Calibrated Imputation:

1. using a flexible nonparametric model

   : that incorporate the design information (e.g, size variable or survey weights)

   : to capture a complex shape (irregular, skewed) of a joint density

   : to offer protection against model misspecification.

2. adding the "calibration" prior

   : to account for calibration constraints

3. based on multiple imputation (editing) techniques

   : to include nonsampling error in the final variance estimate

Contents

1. Introduction

2. Non-parametric MI Using Design Information

3. Bayesian Calibrated Imputation

4. Concluding Remarks and Future Studies

Concluding remarks

- DP mixture model imputation reflects complex distributional features and accounts for the edit constraints.
- After accounting for size variable, DP mixture model imputation works (impute) well under PPS and stratified sampling design.
- Size variable is only an example of many design variables. Thus, our results are not only for PPS or stratified sampling, but also for any survey design as long as agencies can use the design variables/information for imputing missing values.
- Design variables are very useful and need to be incorporated in the imputation process when sampling design is informative.

Future studies

- Investigating on the performance of DP imputation that adds the calibration prior.
- Proof of the design consistency
- Applying to 2017 Economic Census.

Thank you!

Contact Information

Huaiyu Zang

zanghu@mail.uc.edu

Department of Mathematical Sciences

University of Cincinnati

References

- Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H. and Karr, A. F. (2014), Multiple Imputation of Missing or Faulty Values Under Linear Constraints, *Journal of Business & Economic Statistics*, 32, 375–386.
- Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015a), Simultaneous Edit-Imputation for Continuous Microdata, *Journal of the American Statistical Association*, 110, 987–999.
- Andridge, R. R. and Little, R. J. (2010), A review of hot deck imputation for survey nonresponse, *International statistical review*, 78(1), 40-64.
- Deville, Jean-Claude, and Carl-Erik Sarndal (1992). "Calibration estimators in survey sampling." *Journal of the American statistical Association* 87, no. 418 : 376-382.

References (cont.)

- Cochran, W. G. (2007). Sampling techniques, *John Wiley & Sons.*
- Pannekoek, J., Shlomo, N., & De Waal, T. (2013). Calibrated imputation of numerical data under linear edit restrictions. *The Annals of Applied Statistics*, 7(4), 1983-2006.
- Ton De Waal, Wieger Coutinho & Natalie Shlomo (2017). Calibrated Hot Deck Imputation for Numerical Data Under Edit Restrictions. *Journal of Survey Statistics and Methodology*, Volume 5, Issue 3, 372-397.
- Danny Pfeffermann (1993), The Role of Sampling Weights When Modeling Survey Data, *International Statistical Review*, Vol. 61, No. 2, pp. 317-337

References (cont.)

- Berger, Y. G., and Tille, Y. (2009). Sampling with unequal probabilities, *Handbook of statistics*, 29, 39-54.
- Gelman, A. (2007), Struggles with survey weighting and regression modeling, *Statistical Science*, 153-164.
- Si, Y., Pillai, N. S., & Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10(3), 605-625.
- Little, R. J. (2004), To model or not to model? Competing modes of inference for finite population sampling., *Journal of the American Statistical Association*, 299(466), 546-556.

References (cont.)

- HAZIZA, D. and BEAUMONT, J.F. (2017). Construction of weights in surveys: a review. *Statistical Science*, 32, 206-226.
- Elizabeth A. Stuart (2012), Recent Advances in Missing Data Methods: Imputation and Weighting. *Conference in MCHB/HRSA*, 153-164.
- Bonnery, D., Breidt, J. and Coquet, F (2012). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population *Bernoulli*, 18, (4), 1361-1385.
- S Yang and JK Kim (2016). Fractional imputation in survey sampling: a comparative review *Statistical Science*,Volume 31, Number 3 (2016), 415-432.
- H Sang and JK Kim (2017). An approximate Bayesian inference on propensity score estimation under unit nonresponse *arxiv.org*

Appendix: Finite population parameters

- A finite population parameter is any function of survey variable $Y = (y_1, \ldots, y_N)$
  - Population total: $t = \sum_{k \in \mathcal{U}} y_k = \sum_{k=1}^{N} y_k$
  - Population mean: $\bar{y}_u = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k = \frac{1}{N} \sum_{k=1}^{N} y_k$
  - Population variance: $S^2 = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (y_k - \bar{y}_u)^2 = \frac{1}{N-1} \sum_{k=1}^{N} (y_k - \bar{y}_u)^2$
- To estimate finite population parameter, it is common to select a sample $\mathcal{S}$ with size n from the finite population.

Appendix: Model and Design consistency

- Model consistency
  1. Preserve multivariate relation among variables
- Design consistency
  1. The sample estimate becomes exactly equal to the population value when $n = N$.
  2. Calibration (external consistency): the values of variables across units have to sum up to known totals.

Appendix: Notation

- Y: $N \times p$ matrix of partially observed outcome
- X: $N \times q$ matrix of fully observed covariates
- I: **Indicator for inclusion** in the survey
  - $I_{ij} = 1$: sampled unit; $I_{ij} = 0$: non-sampled unit
- R: **Indicator for response** in the survey
  - $R_{ij} = 1$: responded unit; $R_{ij} = 0$: non-responded unit
- M: **Missingness indicator** in the survey
  - $M_{ij} = 1$: missing unit; $M_{ij} = 0$: observed unit
  - $M_{ij} = 1 - R_{ij}$

Appendix: Notation

- ϕ models the inclusion indicator I
- ψ describes the missing-data mechanism M
- θ is a parameter indexing a model for Y
- $\text{obs} = \{(i,j)|I_{ij}R_{ij} = 1\}$ and $\text{mis} = \{(i,j)|I_{ij}M_{ij} = 1\}$
- $\text{inc} = \{(i,j)|I_{ij} = 1\}$
  - $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$
- $\text{nob} = \{(i,j)|I_{ij} = 0 \quad \text{or} \quad R_{ij} = 0\}$
  - $Y = (Y_{\text{obs}}, Y_{\text{nob}})$

Appendix: Missing data mechanism

(Source: Elizabeth A. Stuart 2012)

- **Missing Completely at Random (MCAR)**: Missingness does not depend on any data.
  - $P(M|Y,\psi) = P(M|Y_{obs}, Y_{mis}, \psi) = P(M|\psi)$
  - Cases with missing values a random sample of the original sample
  - No systematic differences between those with missing and observed values
- **Missing at Random (MAR)**: Missingness depends on observed data
  - $P(M|Y_{obs}, Y_{mis}, \psi) = P(M|Y_{obs}, \psi)$
  - MCAR and MAR are ignorable missing data mechanism.
- **Nonignorable Missing Data**: Missingness depends on unobserved values
  - $P(M|Y_{obs}, Y_{mis}, \psi)$ can not be ignored.
  - e.g., probability of someone reporting their income depends on what their income is

UNIVERSITY OF Cincinnati

Appendix: Ignorable sampling mechanism

- Ignorable sampling mechanism
  - Ignorability conditions
    - a $P(I|Y_{obs}, Y_{nob}, X, \theta, \phi) = P(I|Y_{obs}, X, \phi)$
    - b Prior independent: $P(\theta, \phi) = P(\theta)P(\phi)$
  - $P(Y_{nob}|Y_{obs}, X, I)$ reduces to $P(Y_{nob}|Y_{obs}, X)$
- Non-informative sampling design
  - $P(I|Y_{obs}, Y_{nob}, X, \theta, \phi) = P(I|X, \phi)$
  - Special case of ignorable sampling design
  - Population missing at random implies sample missing at random

Appendix: Edit rules for continuous data

1. Logical conditions, called **edit rules**, are used to determine if a reported value has some errors or not
2. Imputing missing values also often considers the constraints

Examples of edit rules

- Range restriction $L_j \leq y_j \leq U_j$,
- Ratio edit $L_{jl} \leq y_j/y_l \leq U_{jl}$,
- Balance edit: $y_j = y_{j1} + y_{j2}$

Appendix: Basu's elephant example

(Source: Jae-Kwang Kim's STAT 521: Survey Sampling Chap. 2)

- Circus with N=50 elephants. Want to estimate the total weights of the elephants using a sample of size n = 1
- About three years ago, every elephant is weighted and "Sambo" was in the middle in terms of the weight. (and "Jumbo" was the largest one.)
- Circus owner's idea: measure Sambo's weight and multiply it by 50.
- Statistician: No ! It's not a probability sampling.
- Circus owner: Well, what is your sampling scheme ?
- Statistician: Let's select Sambo with high probability. Say, select Sambo with probability 99/100, and select the other 49 elephants with probability 1/4900.

Appendix: Basu's elephant example

(Source: Jae-Kwang Kim's STAT 521: Survey Sampling Chap. 2)

- Circus owner: OK. Let's select one with this scheme. (Sambo is selected.) OK. Let's multiply 50 to sambo's weight.
- Statistician: No ! You should multiply the inverse of the inclusion probability. So, you should multiply by 100/99, not by 50.
- Circus owner: ????? What if Jumbo was selected? What number should we multiply?
- Statistician: Well, it is $4,900$.
- Circus owner: What??? You are fired!
- That is how the statistician lost his job (and perhaps became teacher of statistics!)

Appendix: Hansen-Hurwitz estimator

The Hansen-Hurwitz estimator for the population mean μ is

$$\hat{\mu}_{HH} = \frac{1}{N}\left(\frac{1}{n}\cdot\sum_{i=1}^{n}\frac{y_i}{p_i}\right), \text{where} \quad p_i = \frac{z_i}{\sum_{i=1}^{N} z_i}.$$

The unbiased estimator for $Var(\hat{\mu}_{HH})$:

$$\widehat{Var}(\hat{\mu}_{HH}) = \frac{1}{N^2}\cdot\frac{1}{n}\cdot\frac{\sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\sum_{i=1}^{N} y_i\right)^2}{n-1}$$

An approximate (1 - α) 100% confidence interval for μ_{HH} is

$$\hat{\mu}_{HH} \pm t_{\frac{\alpha}{2},n-1}\cdot\sqrt{\widehat{Var}(\hat{\mu}_{HH})}$$

UNIVERSITY OF Cincinnati

Appendix: Hansen-Hurwitz estimator

The Hansen-Hurwitz estimator is used for PPS **sampling with replacement**. It is easy to be implemented in practice without calculating joint inclusion probabilities.

If n/N is negligible, the Hansen-Hurwitz estimator can be used to approximate the variance of Horvitz-Thompson estimator (e.g, Sen-Yates-Grundy variance estimator) under sampling without replacement.

However, the Hansen-Hurwitz estimator can lead to overestimation of the variance for large sampling fractions.

Appendix: Stratified sampling estimator

- The population is partitioned into H stata $U_1, \ldots, U_H$ of size $N_1, \ldots, N_H$. That is $N = \sum_{h=1}^{H} N_h$.
- From stratum h, we select a sample S_h of size n_h, according to a simple random sampling without replacement.
- Let y_{hi} be the value of ith unit in stratum h. Then sample mean $\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ in stratum h.

Appendix: Stratified sampling estimator

The stratified sampling estimator for the population mean μ is

$$\hat{\mu}_{st} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h$$

.

The (unbiased) variance estimator is

$$\widehat{Var}(\hat{\mu}_{st}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

When all of the stratum sizes are small, an approximate 100(1-α)% CI for μ_{st} is

$$\hat{\mu}_{st} \pm t_{\frac{\alpha}{2},n-H}\sqrt{\widehat{Var}(\hat{\mu}_{st})}$$

However, when the stratum sample sizes are at least 30, use z to approximate t.