Imputation for the Treatment of Item Nonresponse in Surveys: An Overview

David Haziza

Department of mathematics and statistics
Université de Montréal

ITACOS 2019
Florence, Italy

June 5, 2019

Outline

Outline

Levels of nonresponse

- It is customary to distinguish unit nonresponse from item nonresponse:
  1. Unit nonresponse:
     - No usable information is collected on a sample unit;
     - Reasons: refusal, inability to contact the sample unit, etc.
  2. Item nonresponse:
     - Absence of information limited to some survey variables only;
     - Reasons: sensitive items (e.g., income), invalid or inconsistent responses.

Methods of treatment

- Treatment of unit nonresponse: typically, handled through weight adjustment procedures methods. They consist of
  - eliminating the nonrespondents from the data file;
  - increasing the sampling weight of the respondents to compensate for the elimination of the nonrespondents;
- Treatment of item nonresponse: `Imputation!`
  - It consists of consists of constructing one or more replacement values to "fill in" for a missing value.
  - Single imputation: replace a missing value by a single imputed value $\longrightarrow$ a single completed data file.
  - Multiple imputation: replace a missing value by $M \geq 2$ replacement values $\longrightarrow$ M completed data files.

Effects of nonresponse

- Why is nonresponse an issue?
  - Nonresponse bias: due to the fact that respondents and nonrespondents do not have the same characteristics with respect to the survey variables.
  - Sample size is smaller than expected $\longrightarrow$ Variance of estimators is greater than that of estimators that would have been used if there were no nonresponse. $\longrightarrow$ nonresponse variance.
- Main objective of all the treatment methods: reduce the nonresponse bias and possibly control the variance due to nonresponse.
- Some review papers on imputation: Durrant (2005), Haziza (2009), Andridge and Little (2010) and Chen and Haziza (2019).

Questions of interest in the last two decades

(1) How to obtain asymptotically unbiased and efficient point estimators?

(2) How to obtain some protection against the misspecification of the underlying model(s)?

(3) How to consistently estimate the variance of imputed estimators?

(4) How to preserve relationships between survey variables?

Finite population parameters

- Let $\mathcal{P}$ be a finite population of size N.
- y: a survey variable
- y_i: y-value attached to unit i, $i = 1, \cdots, N$.
- Goal: estimate a finite population parameter θ_N defined as the solution of the census estimating equation:

$$U_N(\theta_N) = \frac{1}{N} \sum_{i \in \mathcal{P}} u(y_i; \theta_N) = 0.$$

Parameter	$u(y_i; \theta_N)$	Explicit form of θ_N
Total	$y_i - n^{-1}\pi_i\theta_N$	$t_y = \sum_{i \in \mathcal{P}} y_i$
Mean	$y_i - \theta_N$	$\overline{Y} = t_y/N$
Distribution function	$1(y_i \leq t) - \theta_N$	$F_N(t) = \frac{1}{N}\sum_{i \in \mathcal{P}} 1(y_i \leq t)$

Outline

Estimation in ideal conditions

- Ideal conditions: no nonsampling errors
- S: sample of size n selected according to a given sampling design $P(S)$.
- $\pi_i = P(i \in S) > 0$: first-order inclusion probability for unit i (known prior to sampling)
- $w_i = 1/\pi_i$: design weight attached to unit i
- A full sample estimator of θ_N, denoted by $\widehat{\theta}_F$, is the solution of the sample estimating equation

$$\widehat{U}_F(\theta_N) = \frac{1}{\widehat{N}} \sum_{i \in S} w_i u(y_i; \theta_N) = 0.$$

Estimation in ideal conditions

Parameter	$u(y_i;\theta_N)$	Full sample estimator $\widehat{\theta}_F$
Total	$y_i - n^{-1}\pi_i\theta_N$	$\widehat{t}_{HT} = \sum_{i\in S} w_i y_i$
Mean	$y_i - \theta_N$	$\widehat{\overline{Y}}_{HA} = \sum_{i\in S} w_i y_i / \sum_{i\in S} w_i$
Distribution function	$1(y_i \leq t) - \theta_N$	$\widehat{F}_n(t) = \frac{\sum_{i\in S} w_i 1(y_i \leq t)}{\sum_{i\in S} w_i}$

- Under ideal conditions, $\widehat{\theta}_F$ is design-consistent for θ_N; that is,

$$\widehat{\theta}_F - \theta_N = O_p(1/\sqrt{n}) \quad \text{or } O_p(N/\sqrt{n});$$

e.g., Wang and Opsomer (2011) and Breidt and Opsomer (2017).

Estimation after imputation

- In practice the y-variable is subject to missingness.
- Let r_i be a response indicator for unit i such that $r_i = 1$ if y is observed and $r_i = 0$, otherwise.
- Let $\widetilde{y}$ denote the y-variable after imputation. We have

$$\widetilde{y}_i = r_i y_i + (1 - r_i) y_i^*,$$

where y_i^* denotes the imputed value used to replace the missing y_i.

- An estimator of θ_N after imputation, denoted by $\widehat{\theta}_I$, is obtained by solving the estimating equation

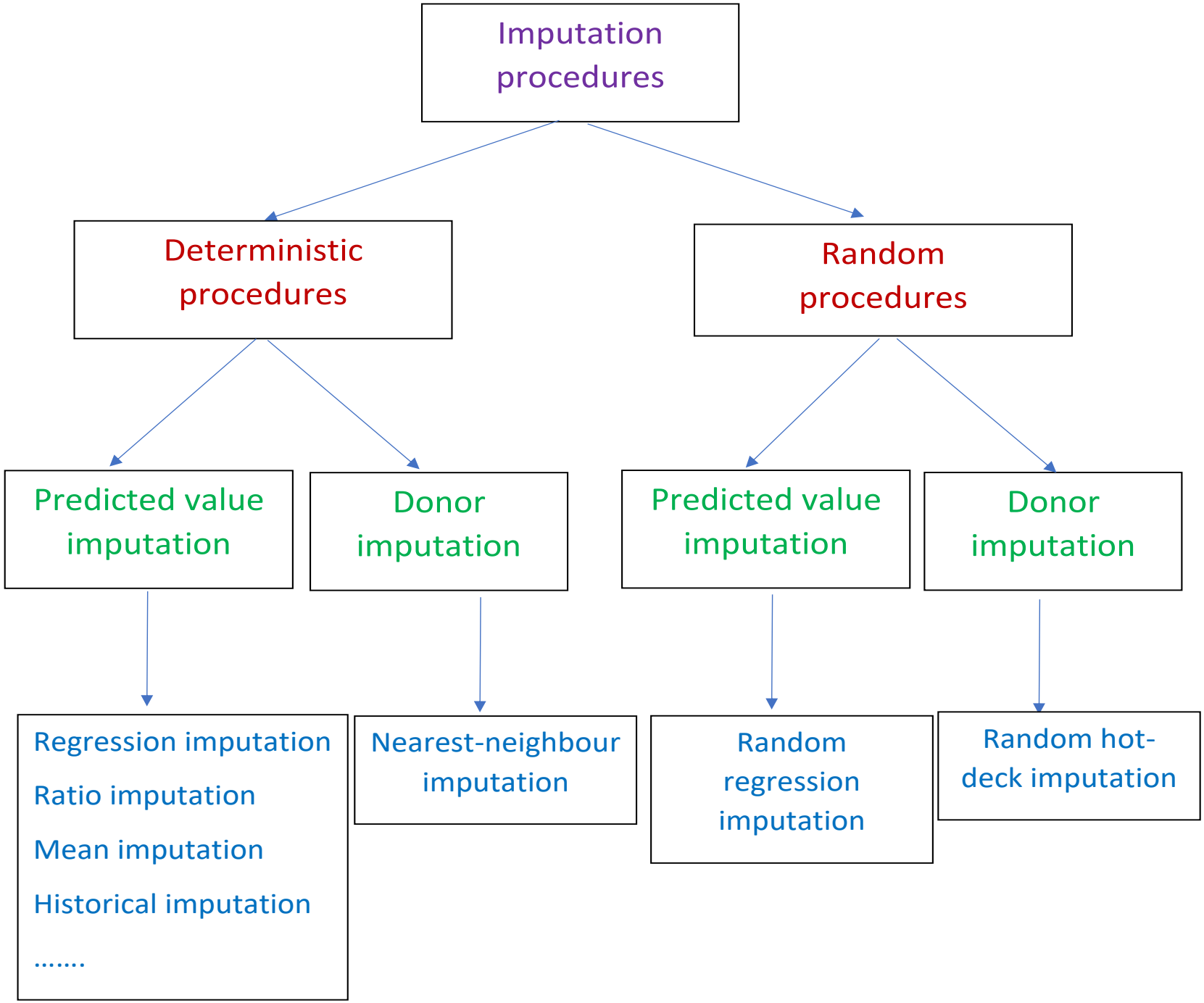
$$\widehat{U}_I(\theta_N) = \frac{1}{N} \sum_{i \in S} w_i u(\widetilde{y}_i; \theta_N) = 0.$$

Estimation after imputation

Parameter	$u(y_i;\theta_N)$	Imputed estimator $\widehat{\theta}_I$
Total	$y_i - n^{-1}\pi_i\theta_N$	$\widehat{t}_I = \sum_{i\in S} w_i\widetilde{y}_i$
Mean	$y_i - \theta_N$	$\widehat{\overline{Y}}_I = \sum_{i\in S} w_i\widetilde{y}_i / \sum_{i\in S} w_i$
Distribution function	$1(y_i \leq t) - \theta_N$	$\widehat{F}_I(t) = \frac{\sum_{i\in S} w_i 1(\widetilde{y}_i \leq t)}{\sum_{i\in S} w_i}$

- Imputed estimators:
  - obtained by applying complete data point estimation procedures to $\widetilde{y}_i$ instead of y_i;
  - attractive from a data user point of view.
- How to obtain the imputed values y_i^*?

Outline


Imputation procedures
Deterministic procedures
Random procedures
Predicted value imputation
Donor imputation
Predicted value imputation
Donor imputation
Regression imputation
Ratio imputation
Mean imputation
Historical imputation
.....
Nearest-neighbour imputation
Random regression imputation
Random hot-deck imputation

The imputation model

- We assume that the y-variable obeys the following model:

$$y_i = m(\mathbf{x}_i; \boldsymbol{\beta}) + \epsilon_i, \tag{1}$$

where

  - $m(\mathbf{x}_i; .)$ is an unknown function;
  - $\mathbf{x}_i$: vector of fully observed variables associated with unit i;
  - $\mathbb{E}_m(\epsilon_i) = 0$, $\mathbb{E}_m(\epsilon_i \epsilon_j) = 0$ for $i \neq j$ and $\mathbb{V}_m(\epsilon_i) = \sigma^2 c_i$, where c_i is a known coefficient attached to unit i.

- Model (1) is called an imputation model or an outcome regression model.

- $\mathbb{E}_m(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i; \boldsymbol{\beta})$: first moment of the imputation model.
- $\mathbb{V}_m(y_i \mid \mathbf{x}_i) = \sigma^2 c_i$: second moment of the imputation model.
- No distributional assumptions about the ϵ_i's

The nonresponse model

- We make the Missing At Random (MAR) assumption (Rubin, 1976):

$$\Pr(r_i = 1 \mid \mathbf{x}_i, y_i) = \Pr(r_i = 1 \mid \mathbf{x}_i) \equiv p(\mathbf{x}_i; \boldsymbol{\alpha}), \tag{2}$$

  where $p(\mathbf{x}_i; .)$ is an unknown function.

- Model (2) is called a nonresponse model.

- Consequence of the MAR assumption:

$$f(y \mid \mathbf{x}_i, r_i = 1) = f(y \mid \mathbf{x}_i, r_i = 0).$$

- Under this assumption, we can generate the imputed values from $f(y \mid \mathbf{x}_i, r_i = 1)$, which can be estimated from the observed data.

Semi-parametric imputation procedures

- Deterministic:

$$y_i^* = m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}_r), \quad i \in S_m,$$

where $\widehat{\boldsymbol{\beta}}_r$ is a suitable estimator (e.g., MLE) of $\boldsymbol{\beta}$ based on the respondents.

- Special cases:
  - $m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}_r) = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_r \longrightarrow$ Regression imputation
  - $\mathbf{x}_i = c_i = 1$ for all $i \longrightarrow m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}_r) = \widehat{\beta}_r \equiv \widehat{\overline{Y}}_r \longrightarrow$ Mean imputation
- Random:

$$y_i^* = m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}_r) + \widehat{\sigma}\sqrt{c_i}e_i^*, \quad i \in S_m,$$

where e_i^* is randomly drawn with replacement from the empirical distribution of standardized residuals observed among the respondents, $\widehat{F}_e(t)$.

- Special case:
  - Mean imputation + added random residuals $\longrightarrow$ Random hot-deck imputation

Nonparametric imputation procedures

- Nearest-neighbour imputation :

$$y_i^* = y_j,$$

where j is the index of the nearest-neighbour of unit i, which satisfies

$$D(\mathbf{x}_j, \mathbf{x}_i) \leq D(\mathbf{x}_k, \mathbf{x}_i), \quad k \in S_r$$

and $D(\cdot\,;\cdot)$ denotes a distance function.

- Predictive mean matching : Same as NNI except that the information contained $\mathbf{x}_i$ is compressed into a single score $\widehat{y}_i = m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}_r)$; see Little, (1988).
- Other nonparametric procedures:
  - Kernel methods: Zhong and Chen (2014)
  - Smoothing splines and additive models: Hasler and Craiu (2016)
  - B-splines: Goga and Haziza (2017)

Outline

Properties of imputed estimators

Back

- Decomposition of the total error for deterministic imputation:

$$\widehat{\theta}_I - \theta_N = \underbrace{\left(\widehat{\theta}_F - \theta_N\right)}_{\text{Sampling error}} + \underbrace{\left(\widehat{\theta}_I - \widehat{\theta}_F\right)}_{\text{Nonresponse error}}$$

- Sampling error: only affected by the population type, the sampling design, the sample size, etc.
- Nonresponse error: affected by the nonresponse mechanism, the response rate, the quality of the imputation model etc.
- Decomposition of the total error for random imputation:

$$\widehat{\theta}_I - \theta_N = \underbrace{\left(\widehat{\theta}_F - \theta_N\right)}_{\text{Sampling error}} + \underbrace{\left(\check{\theta}_I - \widehat{\theta}_F\right)}_{\text{Nonresponse error}} + \underbrace{\left(\widehat{\theta}_I - \check{\theta}_I\right)}_{\text{Imputation error}}$$

- Imputation error: purely artificial $\rightarrow$ parasitic error

Properties of imputed estimators

- Different inferential frameworks may be used:
  - Which quantities among y_i, $\mathbf{x}_i$, I_i, r_i are treated as fixed/random?
  - Usual framework: *mpq* (assuming MAR)
- Population means and totals
  - Deterministic methods such as regression imputation: $\widehat{\overline{Y}}_I - \overline{Y} = O_p(n^{-1/2})$ as long as the first moment of the imputation model is correctly specified; e.g., Chauvet, Deville and Haziza (2011).
  - NNI: we have $\widehat{\overline{Y}}_I - \overline{Y} = O_p(n^{1/2-1/q})$, where q is the size of $\mathbf{x}$; see Yang and Kim (2017a,b) $\longrightarrow$ Bias is not negligible for $q \geq 2$ $\longrightarrow$ curse of dimensionality
  - PMM: $\widehat{\overline{Y}}_I - \overline{Y} = O_p(n^{-1/2})$; see Yang and Kim (2017a,b).
  - Random methods such as random regression imputation: same as deterministic except that $\widehat{\overline{Y}}_I$ suffers from an additional variability due to the random selection of the residuals e_i^* $\longrightarrow$ Imputation variance

Properties of imputed estimators

- Distribution function and quantiles
  - Deterministic methods such as regression imputation: distort the distribution of the variable being imputed $\rightarrow$ biased estimators of quantiles
  - NNI: tend to preserve the distribution function; see Yang and Kim (2017a,b).
  - Random methods such as random regression imputation: tend to preserve the distribution function;

$$\widehat{F}_I(t) - F_N(t) = O_p(n^{-1/2});$$

see Chen, Rao and Sitter (2000), Chauvet, Deville and Haziza (2011) and Boistard, Chauvet and Haziza (2016).

Eliminating the imputation error

- Random imputation for population totals/means: Is it possible to preserve the distribution of the variable being imputed without paying the price of increased variance?
- Fractional imputation: e.g., Kalton and Kish (1981), Fay (1996), Kim and Fuller (2004) and Yang and Kim (2016).
- Balanced imputation: e.g., Kalton and Kish (1981), Chauvet, Deville and Haziza (2011) and Haziza, Nambeu and Chauvet (2014).
  - Idea: Select the residuals e_i^* at random with replacement so that

$$\underbrace{\left(\widehat{\boldsymbol{\theta}}_I - \check{\boldsymbol{\theta}}_I\right)}_{\text{Imputation error}} = 0.$$

  - One Option: Using the Cube algorithm (Deville and Tillé, 2004)

Outline

Variance estimation

- In the multiple imputation literature, single imputation is often criticized. The most common criticism is:

  The most obvious limitation of single imputation is the underlying assumption that the imputed value is the true value. This limitation leads to underestimation of the variance which affects confidence intervals and statistical tests.

- There is a wide literature on variance estimation procedures for singly imputed data in a survey sampling setting, developed in the last two decades.

Variance estimation: Two-phase framework

- $\mathcal{P} \longrightarrow S \longrightarrow S_r$
- Total variance of $\widehat{\theta}_I$ (Särndal, 1992): ▸ Decomposition

$$\mathrm{V}_{\mathrm{tot}} = \mathrm{V}_{\mathrm{sam}} + \mathrm{V}_{\mathrm{nr}} + \mathrm{V}_{\mathrm{mix}};$$

- For random imputation, add the extra variance due to imputation, $\mathrm{V}_{\mathrm{imp}}$.
- Obtaining consistent estimators of each term requires the first two moments of the imputation model to be correctly specified.
- Donor imputation: Brick, Kalton and Kim (2004);
- Nearest-neigbour imputation: Beaumont and Bocci (2009);
- Historical imputation: Beaumont, Haziza and Bocci (2011);
- Composite imputation: Beaumont and Bissonnette (2011).
- Generalized system SEVANI: Beaumont and Bissonnette (2011).

Variance estimation: The reverse framework

- $\mathcal{P} \longrightarrow (\mathcal{P}_r, \mathcal{P}_m) \longrightarrow (S_r, S_m)$
- Total variance of $\widehat{\theta}_I$:

$$\mathrm{V}_{tot} = \mathrm{V}_1 + \mathrm{V}_2,$$

where

$$\mathrm{V}_1 = E_q E_m V_p(\widehat{\theta}_I - \theta_N), \quad \mathrm{V}_2 = E_q V_m E_p(\widehat{\theta}_I - \theta_N).$$

- The contribution of V_2 to the total variance, $\mathrm{V}_2/(\mathrm{V}_1 + \mathrm{V}_2)$ is of order $O(n/N) \longrightarrow$ negligible when the sampling fraction, n/N, is negligible;
- For random imputation, add the extra variance due to imputation, $\mathrm{V}_{\mathrm{imp}}$.
- e.g., Fay (1991), Shao and Steel (1999), Haziza (2009) and Kim and Rao (2009).

Resampling methods

Adjusted jackknife (Rao and Shao, 1992; Rao and Sitter, 1995; Yung and Rao, 2000; Chen and Shao, 2001)

- Calculated in the usual way except that whenever the i-th unit is deleted, the imputed values, y_i^* are adjusted to reflect the fact that the set of respondents is changed.
- Consistent estimator of V_1 provided that the sampling fraction n/N is negligible

Bootstrap (Shao and Sitter, 1996)

- Performed in the usual way, except that the nonrespondents are re-imputed within each bootstrap samples
- Consistent estimator of V_1 provided that the sampling fraction n/N is negligible.
- Bootstrap for non-negligible sampling fractions: more challenging (Mashreghi, Léger and Haziza, 2014; Chen, Haziza, Léger and Mashreghi, 2019).

Outline

Multiply robust imputation procedures

- So far, the imputation procedures were based on a single imputation model $\longrightarrow$ Resulting estimators vulnerable to model misspecification
- To provide additional protection, it may be of interest of fitting multiple imputation models and/or multiple nonresponse models
- Each model may be based on different functionals and/or different sets of predictors.
- Idea: Develop imputation procedures that make use of multiple models (Chen and Haziza, 2017a)
- A procedure is said to be multiply robust if the resulting imputed estimator is consistent if all the models but one are misspecified.
- Concept introduced by Han and Wang (2013); see also Chan and Yam (2014)
- Can be viewed as an extension of double robustness

Two classes of models

- Two classes of models:
  - Nonresponse models:

$$\mathcal{C}_1 = \{p^j(\mathbf{x}_i; \alpha^j); j = 1, \ldots, J\}$$

  - Imputation models:

$$\mathcal{C}_2 = \{m^k(\mathbf{x}_i; \beta^k); k = 1, \ldots, K\}$$

- We fit each of the $J + K$ models to obtain

$$\widehat{\alpha}^1, \ldots, \widehat{\alpha}^J \longrightarrow p^1(\mathbf{x}_i; \widehat{\alpha}^1), \ldots, p^J(\mathbf{x}_i; \widehat{\alpha}^J)$$

and

$$\widehat{\beta}^1, \ldots, \widehat{\beta}^K \longrightarrow m^1(\mathbf{x}_i; \widehat{\beta}^1), \ldots, m^K(\mathbf{x}_i; \widehat{\beta}^K)$$

- Usual estimators: least squares estimators, maximum likelihood estimators etc.

Multiply robust imputation

	$p^1(\mathbf{x};\hat{\boldsymbol{\alpha}}^1)$	.	$p^J(\mathbf{x};\hat{\boldsymbol{\alpha}}^J)$	$m^1(\mathbf{x};\hat{\boldsymbol{\beta}}^1)$	.	$m^K(\mathbf{x};\hat{\boldsymbol{\beta}}^K)$	r	$\tilde{y}$
1	$p^1(\mathbf{x}_1;\hat{\boldsymbol{\alpha}}^1)$	.	$p^J(\mathbf{x}_1;\hat{\boldsymbol{\alpha}}^J)$	$m^1(\mathbf{x}_1;\hat{\boldsymbol{\beta}}^1)$	.	$m^K(\mathbf{x}_1;\hat{\boldsymbol{\beta}}^K)$	1	y_1
2	$p^1(\mathbf{x}_2;\hat{\boldsymbol{\alpha}}^1)$	.	$p^J(\mathbf{x}_2;\hat{\boldsymbol{\alpha}}^J)$	$m^1(\mathbf{x}_2;\hat{\boldsymbol{\beta}}^1)$	.	$m^K(\mathbf{x}_2;\hat{\boldsymbol{\beta}}^K)$	0	y_2^*
.	.			.			.	.
.	.			.			.	.
.	.			.			.	.
n	$p^1(\mathbf{x}_n;\hat{\boldsymbol{\alpha}}^1)$	.	$p^J(\mathbf{x}_n;\hat{\boldsymbol{\alpha}}^J)$	$m^1(\mathbf{x}_n;\hat{\boldsymbol{\beta}}^1)$	.	$m^K(\mathbf{x}_n;\hat{\boldsymbol{\beta}}^K)$	0	y_n^*

Construction of imputed values: Two steps

- Two-steps for constructing the imputed values y_i^*:
  - (1) First step: compress the information included in the J nonresponse models and the K imputation models
  - (2) Second step: Implementation step through linear regression
- Note 1: There are different ways to perform Step 1.
- Note 2:
  - $J = 0$ and $K = 1$: customary imputation based on a single imputation model.
  - $J = 1$ and $K = 1$: doubly robust imputation based on a single imputation model and a single nonresponse model.

Compressing through refitting

- For unit i, define

$$\widehat{\mathbf{Q}}_{pi} = \left(p^1(\mathbf{x}_i;\widehat{\alpha}^1),\ldots,p^J(\mathbf{x}_i;\widehat{\alpha}^J)\right), \quad \widehat{\mathbf{Q}}_{mi} = \left(m^1(\mathbf{x}_i;\widehat{\beta}^1),\ldots,m^K(\mathbf{x}_i;\widehat{\beta}^K)\right).$$

- We summarize the working models information by regressing

$$r_i \text{ on } \widehat{\mathbf{Q}}_{pi} \quad \text{and} \quad y_i \text{ on } \widehat{\mathbf{Q}}_{mi}.$$

This leads to the weighted least square regression coefficients

$$\widehat{\eta}_p = \left(\sum_{i\in S} w_i \widehat{\mathbf{Q}}_{pi}\widehat{\mathbf{Q}}_{pi}^\top\right)^{-1} \sum_{i\in S} w_i \widehat{\mathbf{Q}}_{pi} r_i$$

and

$$\widehat{\eta}_m = \left(\sum_{i\in S} w_i r_i \widehat{\mathbf{Q}}_{mi}\widehat{\mathbf{Q}}_{mi}^\top\right)^{-1} \sum_{i\in S} w_i r_i \widehat{\mathbf{Q}}_{mi} y_i.$$

Compressing through refitting

- For unit i, define

$$\widehat{p}_i = \widehat{\mathbf{Q}}_{pi}^{\top}\widehat{\boldsymbol{\eta}}_p \quad \text{and} \quad \widehat{m}_i = \widehat{\mathbf{Q}}_{mi}^{\top}\widehat{\boldsymbol{\eta}}_m.$$

- The scores $\widehat{p}_i$ and $\widehat{m}_i$ compress respectively the information contained in the J nonresponse models and the K imputation models.
- $\widehat{p}_i$: consistent estimator of $p(\mathbf{x}_i;\boldsymbol{\alpha})$ if one of the models in $\mathcal{C}_1$ is correctly specified.
- $\widehat{m}_i$: consistent estimator of $m(\mathbf{x}_i;\boldsymbol{\beta})$ if one of the models in $\mathcal{C}_2$ is correctly specified.
- see Duan and Yin (2017) and Chen and Haziza (2017b).

Implementation through linear regression

- We construct the imputed values y_i^* as follows:

$$y_i^* = \mathbf{h}_i^\top \widehat{\boldsymbol{\tau}}, \quad i \in S_m,$$

where $\mathbf{h}_i = (1, \widehat{m}_i)^\top$ and

$$\widehat{\boldsymbol{\tau}} = \left\{ \sum_{i \in S} w_i r_i (\widehat{p}_i^{-1} - 1) \mathbf{h}_i \mathbf{h}_i^\top \right\}^{-1} \sum_{i \in S} w_i r_i (\widehat{p}_i^{-1} - 1) \mathbf{h}_i y_i.$$

The resulting imputed estimator is given by

$$\widehat{t}_{MR} = \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) \mathbf{h}_i^\top \widehat{\boldsymbol{\tau}}.$$

Theoretical properties

Theorem

If one of the imputation models is true, then the estimator $\widehat{t}_{MR}$ *is consistent in the sense that* $\widehat{t}_{MR}/t_y \to^p 1$ *as* $n \to \infty$ *and* $N \to \infty$.

Theorem

If one of the nonresponse models is true, then the estimator $\widehat{t}_{MR}$ *is consistent in the sense that* $\widehat{t}_{MR}/t_y \to^p 1$ *as* $n \to \infty$ *and* $N \to \infty$.

Conclusion: $\widehat{t}_{MR}$ is multiply robust

Simulation study: the set-up

- We used the simulation setup of Kang and Schafer (2007), that was also used by Chan and Yam (2014) and Han (2014)
- We generated $B = 1,000$ finite populations of size $N = 10,000$ as follows:
  - For each unit, a vector $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top$ was generated from a standard multivariate normal distribution.
  - The survey variable y was generated according to

$$y = 210 + 27.4x_1 + 13.7(x_2 + x_3 + x_4) + \epsilon,$$

    where the errors ϵ were generated from a standard normal distribution.
  - For each population unit, we generated a size variable $\psi = 0.5\chi + 1$, where χ was drawn from a chi-square distribution with one degree of freedom.

Simulation study: the set-up

- Goal: estimate the population mean

$$\overline{Y} = N^{-1} \sum_{i \in \mathcal{P}} y_i.$$

- From each finite population, we selected a sample, of size $n = 800$, according to randomized systematic sampling with probability proportional to size; That is, $\pi_i = n \left(\psi_i / \sum_{i \in \mathcal{P}} \psi_i \right)$.

- In each sample, the response indicators r_i were generated independently from a Bernoulli distribution with probability

$$p(\mathbf{x}_i) = \{1 + \exp(x_{1i} - 0.5x_{2i} + 0.25x_{3i} + 0.1x_{4i})\}^{-1}.$$

This led to an overall response rate approximately equal to 50%.

Simulation study: the set-up

- As in Kang and Schafer (2007), we considered the following transformations of the $x-$variables:
  - $z_1 = \exp(x_1/2)$;
  - $z_2 = x_2\{1+\exp(x_1)\}^{-1} + 10$;
  - $z_3 = (x_1 x_3/25 + 0.6)^3$;
  - $z_4 = (x_2 + x_4 + 20)^2$.

2 imputation models

$$m^1(\mathbf{x};\beta^1) = \beta_0^1 + \beta_1^1 x_1 + \ldots + \beta_5^1 x_4.$$

$$m^2(\mathbf{x};\beta^1) = \beta_0^1 + \beta_1^1 z_1 + \ldots + \beta_5^1 z_4$$

2 nonresponse models

$$p^1(\mathbf{x};\alpha^1) = \left\{1+\exp\left(\alpha_0^1 + \alpha_1^1 x_1 + \ldots + \alpha_4^1 x_4\right)\right\}^{-1}$$

$$p^2(\mathbf{x};\alpha^1) = \left\{1+\exp\left(\alpha_0^1 + \alpha_1^1 z_1 + \ldots + \alpha_4^1 z_4\right)\right\}^{-1}$$

Simulation study: the set-up

We computed several estimators of $\overline{Y}$ of the form:

$$\widehat{\overline{Y}}_I = \frac{1}{N}\left(\sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i(1-r_i)y_i^*\right)$$

1. Two imputed estimators through nearest-neighbor imputation based on the set of matching variables x and z, respectively
2. Four doubly robust estimators (Haziza and Rao, 2006)
3. Nine multiply robust estimators

Simulation study: results

Estimator	Percent relative bias	Root mean square error
$\widehat{\overline{Y}}_{\mathrm{NN}}(00\|10)$	-0.8	2.21
$\widehat{\overline{Y}}_{\mathrm{NN}}(00\|01)$	-3.8	8.21
$\widehat{\overline{Y}}_{\mathrm{DR}}(10\|10)$	-0.0	1.40
$\widehat{\overline{Y}}_{\mathrm{DR}}(10\|01)$	0.0	1.97
$\widehat{\overline{Y}}_{\mathrm{DR}}(01\|10)$	0.0	1.45
$\widehat{\overline{Y}}_{\mathrm{DR}}(01\|01)$	-2.5	16.88
$\widehat{\overline{Y}}_{\mathrm{MR}}(10\|10)$	-0.0	1.40
$\widehat{\overline{Y}}_{\mathrm{MR}}(10\|01)$	0.1	1.63
$\widehat{\overline{Y}}_{\mathrm{MR}}(01\|10)$	-0.0	1.40
$\widehat{\overline{Y}}_{\mathrm{MR}}(01\|01)$	-1.2	3.05
$\widehat{\overline{Y}}_{\mathrm{MR}}(11\|10)$	-0.0	1.40
$\widehat{\overline{Y}}_{\mathrm{MR}}(11\|01)$	0.0	1.64
$\widehat{\overline{Y}}_{\mathrm{MR}}(10\|11)$	-0.0	1.40
$\widehat{\overline{Y}}_{\mathrm{MR}}(01\|11)$	-0.0	1.40
$\widehat{\overline{Y}}_{\mathrm{MR}}(11\|11)$	-0.0	1.40

Simulation study: More results

- How do multiply robust procedures perform when all the models are misspecified?

(MR1). $p^2(\mathbf{z};\alpha^2)$ and $m^2(\mathbf{z};\beta^2)$: $\widehat{\overline{Y}}_{\mathrm{MR1}}$.

(MR2). $p^2(\mathbf{z};\alpha^2)$ and $m^2(\mathbf{z};\beta^2)$: + one additional outcome regression model by including (z_1, z_2, z_3, z_4) and all their interactions as covariates: $\widehat{\overline{Y}}_{\mathrm{MR2}}$.

(MR3). One additional outcome regression model by using $(\sqrt{z_1}, \sqrt{z_2}, \sqrt{z_3}, \sqrt{z_4})$ and all their interactions : $\widehat{\overline{Y}}_{\mathrm{MR3}}$.

(MR4). One additional outcome regression model by using $(\log z_1, \log z_2, \log z_3, \log z_4)$ and all their interactions : $\widehat{\overline{Y}}_{\mathrm{MR4}}$.

(MR5). Three additional outcome regression models described in (MR2), (MR3) and (MR4): $\widehat{\overline{Y}}_{\mathrm{MR5}}$.

Simulation study: more results

Table: Percent relative bias (RB), and root mean squared error (RMSE) with a response rate of 50%

Estimator	$\widehat{\overline{Y}}_{\mathrm{MR1}}$	$\widehat{\overline{Y}}_{\mathrm{MR2}}$	$\widehat{\overline{Y}}_{\mathrm{MR3}}$	$\widehat{\overline{Y}}_{\mathrm{MR4}}$	$\widehat{\overline{Y}}_{\mathrm{MR5}}$
RB	-1.20	-0.49	-0.52	-0.55	-0.42
RMSE	3.05	2.03	1.91	1.90	1.92

Results suggest: Despite being inconsistent, multiply robust imputation procedures tend to have good numerical performance even if all the models are misspecified.

Outline

Multiple imputation, Rubin (1978, 1987)

- $M \geq 2$ values are imputed for each missing value, resulting in the creation of $M \geq 2$ imputed data files.
- The question is then: how to combine the M files to obtain a valid point estimate and a valid variance estimate?
- Although it is a popular method in many fields in statistics, single imputation is the norm in surveys. Why?
  - Survey statisticians and data analysts are used to work with a single imputed file;
  - There are some questions and debates about the validity of multiple imputation in a survey sampling setting.

Rubin's Rule

- Point estimator of θ_N:

$$\widehat{\theta}_{I,M} = M^{-1} \sum_{m=1}^{M} \widehat{\theta}_I^{(m)},$$

where $\widehat{\theta}_I^{(m)}$ the imputed estimator corresponding to the mth imputed data set.

- Variance estimator for $\widehat{\theta}_{I,M}$:

$$T_M = \overline{W}_M + \left(1 + \frac{1}{M}\right) B_M,$$

where

$$\overline{W}_M = \frac{1}{M} \sum_{m=1}^{M} W^{(m)}, \qquad B_M = \frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}_I^{(m)} - \widehat{\theta}_{I,M}\right)^2$$

and $W^{(m)}$ is the variance estimator computed from the mth imputed data set treating imputed values as if they were observed values.

Generating imputations

- Since multiple imputation was originally proposed under Bayesian considerations, Bayesian imputation procedures provide a natural option.
- Idea: generate imputed values from $P(\mathbf{y}_{mis} \mid \mathbf{y}_{obs})$, which can be expressed as

$$P(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}) = \int P(\mathbf{y}_{mis} \mid \beta, \mathbf{y}_{obs}) P(\beta \mid \mathbf{y}_{obs}) d\beta,$$

  where β is a parameter indexing the imputation model for the y-variable.
- Interpretation: For $1 \leq m \leq M$,
  - First, draw a value $\beta^{(m)}$ from $P(\beta \mid \mathbf{y}_{obs})$.
  - Then, sample $\mathbf{y}_{mis}^{(m)}$ from $P(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \beta^{(m)})$.

Bayesian validity of multiple imputation

- If the imputation model holds, then multiple imputation yields valid inferences provided M is not too small.
- In practice, the imputer's model is generally different from the analyst's model, a situation known as uncongeniality.
- For instance, in business surveys, imputation may be done at the NAICS3 level, whereas the analyst is interested in producing estimates at the NAICS6 level $\longrightarrow$ the analyst's model is more saturated than the imputer's model.
- Uncongeniality is especially problematic in this case; Xie and Meng (2017).
- Under uncongeniality, multiple imputation may lead to invalid inferences.
- What about the properties of multiple imputation from a frequentist perspective?

Validity of multiple imputation: frequentist *pq*-approach

- In the frequentist *pq* approach, the properties of point and variance estimators are evaluated with respect to the joint distribution induced by the sampling design and the nonresponse mechanism. The population y-values are treated as fixed.
- Here, the imputation must be proper for multiple imputation to yield valid inferences.
- To be proper, the imputation must meet the following three conditions:

  C1: $\mathbb{E}_q(\widehat{\theta}_{I,\infty}) = \widehat{\theta}_F$.

  C2: $\mathbb{E}_q(\overline{W}_\infty) = \mathbb{V}_p(\widehat{\theta}_F)$.

  C3: $\mathbb{E}_q(B_\infty) \approx \mathbb{V}_q(\widehat{\theta}_{I,\infty})$.

- Issue: proper imputation may be difficult to achieve for complex sampling designs (Binder and Sun, 1996).

Validity of multiple imputation: frequentist *pq*-approach

- To overcome the difficulty, Bjørnstad (2007) suggested a simple modification to the customary multiple imputation variance estimator:

$$T_M^* = \overline{W}_M + \left(c + \frac{1}{M}\right) B_M,$$

where c is such that

$$\mathbb{E}_{pql}\left(T_M^*\right) = \mathbb{V}_{pql}\left(\widehat{\theta}_{I,M}\right).$$

- For instance, in the context of simple random sampling without replacement and random hot-deck imputation, he found $c \approx \widehat{p}_r^{-1}$, where $\widehat{p}_r$ is the observed response rate.
- Drawback: the value of c depends on the sampling design, the parameter of interest and the imputation procedure.
- Extensions to more complex cases are needed.

Validity of multiple imputation: frequentist *mpq*-approach

- Recall that

$$\mathrm{V}_{\mathrm{tot}} = \mathrm{V}_{\mathrm{sam}} + \mathrm{V}_{\mathrm{nr}} + \mathrm{V}_{\mathrm{mix}} + \mathrm{V}_{\mathrm{imp}}$$

  and

$$T_M = \overline{W}_M + \left(1 + \frac{1}{M}\right) B_M.$$

  - $\overline{W}_M$ is asymptotically unbiased for $\mathbb{V}_{\mathrm{sam}}$.
  - $(1 + M^{-1})B_M$ is asymptotically unbiased for $\mathbb{V}_{\mathrm{nr}} + \mathbb{V}_{\mathrm{imp}}$.

- Issue: The estimator T_M does not track the component $\mathbb{V}_{\mathrm{mix}}$. Therefore,

$$\mathrm{Bias}(T_M) = -\mathbb{V}_{\mathrm{mix}};$$

  see Kott (1995) and Kim et al. (2006).

Validity of multiple imputation: frequentist *mpq*-approach

- The term $\mathbb{V}_{\text{mix}}$ is generally not equal to 0 and, in some cases, its contribution to the total variance may be large $\longrightarrow$ T_M may be considerably biased.
- Often (but not always), the term $\mathbb{V}_{\text{mix}}$ is negative leading to conservative variance estimator.
- The condition $\mathbb{V}_{\text{mix}} = 0$, is called the self-efficiency condition (Meng, 1994; Meng and Romero, 2003).
- Whether or not the self-efficiency condition holds depends on the sampling design, on the parameter of interest and on the imputation procedure used to fill in the missing values.

Validity of multiple imputation: survey features

- In the context of survey data, Reiter et al. (2006) showed that the point and variance estimators are generally biased when complex survey-design features are not accounted for in the imputation models. Survey-design features include unequal weighting, stratification and clustering.
- Reiter et al. (2006) and Schenker et al. (2006) suggested including design information (e.g., cluster membership indicators, stratum indicators and survey weights) as additional covariates in $\mathbf{x}$.
- Generally, incorporating design information into the imputation model may be challenging as the weighting process is typically complex.
- Kim and Yang (2017): Rubin's variance estimator may exhibit some bias even after incorporating design information in the imputation model $\longrightarrow$ proposed an approach, whereby Rubin's rule can be safely used.

Outline

Other topics

- Empirical likelihood confidence intervals under imputation: Cai, Rao and Malgorzata (2019)
- Imputation for multivariate parameters: challenge is to preserve the relationships; e.g., Shao and Wang (2002), Andridge and Little (2010), Chauvet and Haziza (2012), Kim and Fuller (2012), Chaput et al. (2018)
- NMAR: challenging problem; e.g., Siddique et al. (2012), van Buuren (2012) and Sullivan and Andridge (2015).

Grazie mille!