A general multiply robust framework for combining probability and non-probability samples in surveys

David Haziza

Department of mathematics and statistics
University of Ottawa

Joint work with

Sixia Chen (University of Oklahoma)

Workshop on Methodologies for Official Statistics
Rome, Italy

December 5, 2022

Combining probability and non-probability samples

- Traditionally, National Statistical Offices have collected data by means of probability sampling procedures $\longrightarrow$ Design-based inference
- In recent years, there has been a shift of paradigm in NSOs that can be explained by three main factors:
  (i) a dramatic decrease in response rates;
  (ii) increasing data collection costs;
  (iii) the availability of various types of non-probabilistic data sources that include administrative files, opt-in panels, social medias and satellite information.
- Non-probabilistic data sources provide timely data but they often fail to represent the target population of interest because of inherent selection biases.

Combining probability and non-probability samples

- How to integrate data from non-probability samples has attracted a lot of attention in recent years; e.g., Rivers (2007), Bethlehem (2016), Elliot and Vaillant (2017), Lohr and Raghunathan (2017), Kim et al. (2019), Chen et al. (2020), Beaumont (2020) and Rao (2020).
- Estimation procedures may be classified into three broad classes :
  (i) Calibration weighting of a nonprobability sample to estimated benchmarks from a probability survey;
  (ii) Statistical matching or mass imputation;
  (iii) Propensity score weighting of a nonprobability sample;
- Focus of this presentation: (ii) and (iii).

Parameters of interest

- Consider a finite population $\mathcal{P}$ 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_0 defined as the solution of the census estimating equation:

$$\frac{1}{N} \sum_{i \in \mathcal{P}} U(y_i; \theta_0) = 0.$$

Parameter	$U(y_i; \theta_0)$	Explicit form of θ_0
Mean	$y_i - \theta_0$	$\overline{Y} = \sum_{i \in \mathcal{P}} y_i / N$
Population $\tau_\alpha -$th percentile	$1(y_i \leq \theta_0) - \alpha$	$\tau_\alpha = F_N^{-1}(\alpha)$

The setup

- S_A : sample, of size n_A, selected from $\mathcal{P}$ according to a probability sampling design with first-order inclusion probabilities π_i (Known).
- S_B: Non-probability sample, of size n_B, from $\mathcal{P}$.
- Typically, we would expect $n_B > n_A$.
- The data:

Data	$\mathbf{x} = (x_1, \cdots, x_p)$	y	
S_A	$\checkmark$	$\emptyset$	Probability
S_B	$\checkmark$	$\checkmark$	Non-Probability

- I_i : sample selection indicator such that $I_i = 1$ if $i \in S_A$ and $I_i = 0$, otherwise.
- δ_i : a participation indicator such that $\delta_i = 1$ if $i \in S_B$ and $\delta_i = 0$, otherwise.

	y	x_1	$\ldots$	x_p	I_i	$1/\pi_i$
1	X	✓	$\ldots$	✓	1	$1/\pi_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_A	X	✓	$\ldots$	✓	1	$1/\pi_{n_A}$
n_A+1	X	✓	$\ldots$	✓	0	$1/\pi_{n_A+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X	✓	$\ldots$	✓	0	$1/\pi_N$

Table 1: Probability sample S_A (n_A)

	y	x_1	$\ldots$	x_p	δ_i
1	✓	✓	$\ldots$	✓	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_B	✓	✓	$\ldots$	✓	1
n_B+1	X	X	$\ldots$	X	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X	X	$\ldots$	X	0

Table 2: Non-probability sample S_B (n_B)

The setup

- $\pi_i = P(I_i = 1) = P(i \in S_A)$ is known for all $i \in \mathcal{P}$.
- Unknown probability of participation on the non-probability source:

  $\Pr(\delta_i = 1 | \mathbf{x}_i, y_i) = \Pr(\delta_i = 1 | \mathbf{x}_i) \triangleq p(\mathbf{x}_i; \boldsymbol{\alpha}) \longrightarrow$ participation model

- Positivity assumption:

$$p(\mathbf{x}_i; \boldsymbol{\alpha}) > 0 \quad \text{for all } i \in \mathcal{P}.$$

- Outcome regression model:

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

  where $\mathbb{E}(\epsilon_i | \mathbf{x}_i) = 0$ and $\mathbb{V}(\epsilon_i | \mathbf{x}_i) = \sigma^2$.

The setup

- Statistical matching (or mass imputation)
  - Specification of an outcome regression model
  - The resulting estimator may be biased if the outcome regression model is misspecified.
- Propensity score weighting
  - Specification of a participation model
  - The resulting estimator may be biased if the participation model is misspecified.
- Regardless of the approach, the validity of point estimators relies on the validity of an assumed model $\longrightarrow$ point estimators are vulnerable to model misspecification.
- Multiply robust estimation procedures are attractive because they provide some protection against misspecification of the model.

Two classes of models

- Class of potential outcome regression models:

$$\mathcal{M}_1 = \left\{ m^{(j)}(\mathbf{x}; \boldsymbol{\beta}^{(j)}), j = 1, 2, \ldots, J \right\}$$

  J models for the survey variable y

- Class of potential participation models:

$$\mathcal{M}_2 = \left\{ p^{(k)}(\mathbf{x}; \boldsymbol{\alpha}^{(k)}), k = 1, 2, \ldots, K \right\}$$

  K models for the participation probability

- The models in $\mathcal{M}_1$ (respectively in $\mathcal{M}_2$) may be based on different functionals and/or different vectors of explanatory variables.

Estimation of the $\boldsymbol{\beta}$'s

- Estimators of $\boldsymbol{\beta}^{(j)}, j = 1, 2, \ldots, J$: obtained by solving the sample estimating equations

$$\frac{1}{N}\sum_{i \in S_B}\left\{y_i - m^{(j)}(\mathbf{x}_i; \boldsymbol{\beta}^{(j)})\right\}\left\{\frac{\partial m^{(j)}(\mathbf{x}_i; \boldsymbol{\beta}^{(j)})}{\partial \boldsymbol{\beta}^{(j)}}\right\}^\top = 0.$$

- Special case: The jth model is a linear regression model $\longrightarrow$

$$\widehat{\boldsymbol{\beta}}^{(j)} = \left(\sum_{i \in S_B}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\sum_{i \in S_B}\mathbf{x}_i y_i.$$

- For each unit i, we obtain J predicted values:

$$m^{(1)}(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(1)}), m^{(2)}(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(2)}), \ldots, m^{(J)}(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(J)})$$

	y	x_1	$\ldots$	x_p	I_i	$1/\pi_i$
1	X	✓	$\ldots$	✓	1	$1/\pi_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_A	X	✓	$\ldots$	✓	1	$1/\pi_{n_A}$
n_A+1	X	✓	$\ldots$	✓	0	$1/\pi_{n_A+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X	✓	$\ldots$	✓	0	$1/\pi_N$

Table 3: Probability sample S_A (n_A)

	y	x_1	$\ldots$	x_p	δ_i
1	✓	✓	$\ldots$	✓	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_B	✓	✓	$\ldots$	✓	1
n_B+1	X	X	$\ldots$	X	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X	X	$\ldots$	X	0

Table 4: Non-probability sample S_B (n_B)

Estimation of the $\boldsymbol{\alpha}$'s

- If $\mathbf{x}_i$ was available for $i \in \mathcal{P} - S_B$, we would estimate $\boldsymbol{\alpha}^{(k)}, k = 1 \cdots, K$, by solving the census estimating equations

$$\frac{1}{N} \sum_{i \in \mathcal{P}} \frac{\delta_i - p_i^{(k)}}{p_i^{(k)}(1 - p_i^{(k)})} \left\{ \frac{\partial p_i^{(k)}}{\partial \boldsymbol{\alpha}^{(k)}} \right\}^\top = 0,$$

where $p_i^{(k)} \equiv p^{(k)}(\mathbf{x}_i; \boldsymbol{\alpha}^{(k)})$.

- Idea in Chen et al. (2020):

$$\frac{1}{N} \sum_{i \in \mathcal{P}} \frac{\delta_i}{p_i^{(k)}(1 - p_i^{(k)})} \left\{ \frac{\partial p_i^{(k)}}{\partial \boldsymbol{\alpha}^{(k)}} \right\}^\top - \frac{1}{N} \sum_{i \in \mathcal{P}} \frac{1}{1 - p_i^{(k)}} \left\{ \frac{\partial p_i^{(k)}}{\partial \boldsymbol{\alpha}^{(k)}} \right\}^\top = 0.$$

Estimation of the $\boldsymbol{\alpha}$'s

- The estimators of $\boldsymbol{\alpha}^{(k)}, k = 1, 2, \ldots, K$ can be obtained by solving

$$\frac{1}{N}\sum_{i \in S_B} \frac{1}{p_i^{(k)}(1-p_i^{(k)})}\left\{\frac{\partial p_i^{(k)}}{\partial \boldsymbol{\alpha}^{(k)}}\right\}^\top - \frac{1}{N}\sum_{i \in S_A} \pi_i^{-1}\frac{1}{1-p_i^{(k)}}\left\{\frac{\partial p_i^{(k)}}{\partial \boldsymbol{\alpha}^{(k)}}\right\}^\top = 0.$$

- For each unit i, we obtain K estimated participation probabilities:

$$p^{(1)}(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^{(1)}), p^{(2)}(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^{(2)}), \ldots, p^{(K)}(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^{(K)})$$

Compressing the information

- For each i, define

$$\mathrm{v}_{1i} = (m^{(1)}(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(1)}), m^{(2)}(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(2)}), \ldots, m^{(J)}(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(J)}))^\top$$

and

$$\mathrm{v}_{2i} = (p^{(1)}(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^{(1)}), p^{(2)}(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^{(2)}), \ldots, p^{(K)}(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^{(K)}))^\top.$$

- Compress the information contained in the J outcome regression models in $\mathcal{M}_1$ by fitting a linear regression model based on the units in S_B with y as the dependent variable and v_1 as the vector of explanatory variables.

- The compressed score is $\widehat{m}_i = \mathrm{v}_{1i}^\top \widehat{\boldsymbol{\tau}}_1$ with

$$\widehat{\boldsymbol{\tau}}_1 = \left\{ \sum_{i \in S_B} \mathrm{v}_{1i} \mathrm{v}_{1i} \right\}^{-1} \sum_{i \in S_B} \mathrm{v}_{1i} y_i.$$

Compressing the information

- Compress the information contained in the K participation models in $\mathcal{M}_2$ by fitting a linear regression model with δ as the dependent variable and v_2 as the vector of explanatory variables.

- If v_{2i} was available for all $i \in \mathcal{P}$, the compressed score would be $\widetilde{p}_i = \mathrm{v}_{2i}^\top \widetilde{\boldsymbol{\tau}}_2$ with

$$\widetilde{\boldsymbol{\tau}}_2 = \left\{ \sum_{i \in \mathcal{P}} \mathrm{v}_{2i} \mathrm{v}_{2i}^\top \right\}^{-1} \sum_{i \in \mathcal{P}} \mathrm{v}_{2i} \delta_i .$$

- Solution:

$$\widehat{\boldsymbol{\tau}}_2 = \left\{ \sum_{i \in S_A} \pi_i^{-1} \mathrm{v}_{2i} \mathrm{v}_{2i}^\top \right\}^{-1} \sum_{i \in S_B} \mathrm{v}_{2i} .$$

- The compressed score is $\widehat{p}_i = \mathrm{v}_{2i}^\top \widehat{\boldsymbol{\tau}}_2$.

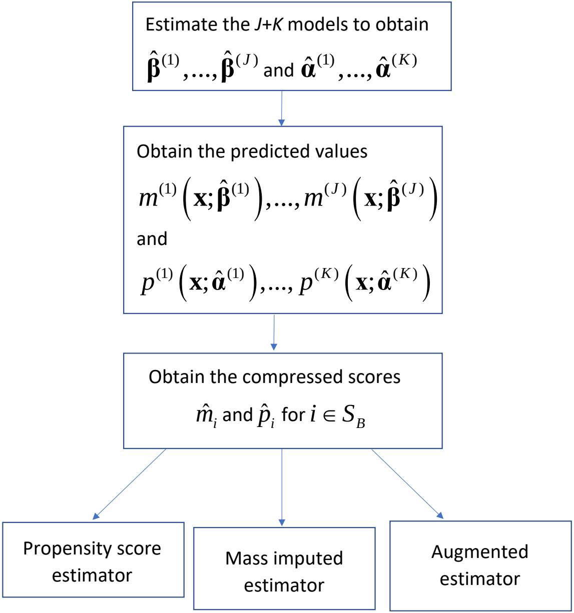

Figure 1: Steps for multiply robust estimation

Propensity score estimation

- Inverse probability weighting estimator $\widehat{\theta}_{IPW}$: obtained by solving the sample estimating equations:

$$\widehat{U}_{IPW}(\theta) = \frac{1}{N} \sum_{i \in S_B} \frac{1}{\widehat{p}_i} U(y_i; \theta) = 0.$$

- $\widehat{\theta}_{IPW}$:multiply robust in the sense that it remains consistent if one of the participation models in $\mathcal{M}_2$ is correctly specified.

- Special case: The population mean

$$\widehat{\theta}_{IPW} = \frac{\sum_{i \in S_B} \frac{y_i}{\widehat{p}_i}}{\sum_{i \in S_B} \frac{1}{\widehat{p}_i}}$$

Fractionally mass imputation

- Fractionally mass imputed estimator $\widehat{\theta}_{FMI}$:
- A consistent estimator of θ_0 is obtained by solving the following expected estimating equations:

$$\frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} \mathbb{E} \left\{ U(y_i; \theta_0) \mid \mathbf{x}_i \right\} = 0.$$

- The expectation is unknown as $f(y \mid \mathbf{x})$ is unknown.
- We want to approximate the conditional expectation by the weighted mean of the fractionally imputed estimating equations:

$$\mathbb{E} \left\{ U(y_i; \theta) \mid \mathbf{x}_i \right\} \approx \sum_{j \in S_B} w_{ij}^* U(y_i^{*(j)}; \theta),$$

where w_{ij}^* are the fractional weights such that $\sum_{j \in S_B} w_{ij}^* = 1$ and the $y_i^{*(j)}$'s denote the imputed values for unit i.

	y	x_1	$\dots$	x_p	I_i	$1/\pi_i$
1	X	✓	$\dots$	✓	1	$1/\pi_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_A	X	✓	$\dots$	✓	1	$1/\pi_{n_A}$
$n_A + 1$	X	✓	$\dots$	✓	0	$1/\pi_{n_A+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X	✓	$\dots$	✓	0	$1/\pi_N$

Table 5: Probability sample S_A (n_A)

	y	x_1	$\dots$	x_p	δ_i
1	✓	✓	$\dots$	✓	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_B	✓	✓	$\dots$	✓	1
$n_B + 1$	X	X	$\dots$	X	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X	X	$\dots$	X	0

Table 6: Non-probability sample S_B (n_B)

Fractionally mass imputation

- Fractionally mass imputed estimator $\widehat{\theta}_{FMI}$:

(Step1). Obtain the weights $\widehat{w}_i$ by maximizing the empirical likelihood function

$$l = \sum_{i \in S_B} \log(w_i),$$

subject to

$$\sum_{i \in S_B} w_i = 1, \quad \sum_{i \in S_B} w_i \widehat{\epsilon}_i = 0,$$

where $\widehat{\epsilon}_i = y_i - \widehat{m}_i$ denotes the residual attached to $i \in S_B$.

(Step2). Obtain $\widehat{\boldsymbol{\theta}}_{FMI}$ by solving the sample estimating equations

$$\widehat{U}_{FMI}(\theta) = \frac{1}{N} \sum_{i \in S_A} \pi_i^{-1} \sum_{j \in S_B} w_{ij}^* U(y_i^{*(j)}; \theta) = 0,$$

with $w_{ij}^* = \widehat{w}_j$ denote the fractional weight and $y_i^{*(j)} = \widehat{m}_i + \widehat{\epsilon}_j$.

Fractionally mass imputation

- $\widehat{\theta}_{FMI}$:multiply robust in the sense that it remains consistent if one of the outcome regression models in $\mathcal{M}_1$ is correctly specified.

- Special case: The population mean

$$\widehat{\theta}_{FMI} = \frac{\sum_{i \in S_A} \pi_i^{-1} \sum_{j \in S_B} w_{ij}^* y_i^{*(j)}}{\sum_{i \in S_A} \frac{1}{\pi_i}}$$

Augmented estimator

- Augmented estimator $\widehat{\theta}_{AMR}$: can be obtained by solving the sample estimating equations

$$\begin{aligned} \widehat{U}_{AMR}(\theta) &= \frac{1}{N}\sum_{i\in S_B}\frac{1}{\widehat{p}_i}U(y_i;\theta)+\frac{1}{N}\sum_{i\in S_A}\pi_i^{-1}\sum_{j\in S_B}w_{ij}^{*}U(y_i^{*(j)};\theta) \\ &- \frac{1}{N}\sum_{i\in S_B}\frac{1}{\widehat{p}_i}\sum_{j\in S_B}w_{ij}^{*}U(y_i^{*(j)};\theta)=0. \end{aligned}$$

- $\widehat{\theta}_{AMR}$: multiply robust in the sense that it remains consistent if one of the models in either $\mathcal{M}_1$ or $\mathcal{M}_2$ is correctly specified.

Simulation study

- We generated $B = 1,000$ finite populations of size $N = 20,000$.
- Two auxiliary variables x_1 and x_2: generated from a $\mathcal{N}(1, 2)$
- The variable of interest y was generated according to

$$y = 0.3 + 2x_1 + 2x_2 + \epsilon,$$

  where $\epsilon \sim \mathcal{N}(0, 1)$.
- From each finite population, a sample S_A, of size n_A, was selected using SRSWOR. We used $n_A = 500$ and $n_A = 1000$.
- A non-probability sample S_B was generated using a Poisson sampling design with probability

$$p(\mathbf{x}_i; \boldsymbol{\alpha}) = \frac{\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})}{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})}.$$

- The values of α_0, α_1, and α_2 were chosen so as to lead to n_B approximately equal to 500 and 1000.

Simulation study

- To assess the performance of the proposed methods in the presence of model misspecification, we defined the transformed explanatory variables as

$$z_1 = \exp(x_1/2) \text{ and } z_2 = x_2 \left\{1 + \exp(x_1)\right\}^{-1}.$$

- The correct outcome regression and participation models were fitted using a linear regression model and a logistic model, respectively, based on the set of explanatory variables $\mathbf{x} = (x_1, x_2)^\top$.
- The incorrect outcome regression and participation models were fitted using a linear regression model and a logistic model, respectively, based on the set of transformed explanatory variables $\mathbf{z} = (z_1, z_2)^\top$.
- We assume that only the variables $\mathbf{x}$ and $\mathbf{z}$ were available in S_A, whereas the variables $\mathbf{x}$, $\mathbf{z}$ and y were available in S_B.

Simulation study

- Goal: estimate
  - The population mean of y;
  - The population 25th percentile of y.
- We computed several estimators:

(1) The (unfeasible) design-weighted estimators (Benchmark) based on S_A obtained as a solution of the following estimating equations

$$\frac{1}{N}\sum_{i\in S_A}\pi_i^{-1}U(y_i;\theta)=0.$$

(2) The naive estimators (Naive) based on S_B obtained as a solution of the following estimating equations

$$\frac{1}{N}\sum_{i\in S_B}U(y_i;\theta)=0.$$

Simulation study

(3) The parametric mass imputed estimators considered in Kim et al. (2019) using correct outcome regression model (PFMI(1000)) and the incorrect outcome regression model (PFMI(0100)).

(4) The doubly robust estimators proposed by Chen et al. (2020): DR(1010), DR(1001), DR(0110) and DR(0101).

(5) The MR inverse probability weighting estimator: MRIPW(0011).

(6) The MR fractionally mass imputed estimator: MRFMI(1100).

(7) The augmented multiply robust estimators: AMR(1110), AMR(1101), AMR(1011), AMR(0111) and AMR(1111).

Simulation results

Parameter	Method	$(n_A, n_B = 500)$			$(n_A, n_B = 1000)$		
		RB (%)	RSE	RRMSE	RB (%)	RSE	RRMSE
	Benchmark	-0.04	1.59	1.59	-0.04	1.14	1.14
	Naive	9.37	1.56	9.50	9.14	1.16	9.21
	PFMI(1000)	0.02	1.59	1.59	-0.03	1.13	1.13
	PFMI(0100)	5.06	1.57	5.29	4.84	1.11	4.96
	DR(1010)	0.01	1.59	1.59	-0.03	1.13	1.13
	DR(1001)	0.02	1.60	1.60	-0.03	1.13	1.13
	DR(0110)	0.10	1.89	1.89	-0.04	1.34	1.34
Mean	DR(0101)	4.60	1.69	4.90	4.42	1.19	4.58
	MRIPW(0011)	0.48	1.98	2.04	0.12	1.37	1.37
	MRFMI(1100)	0.01	1.59	1.59	-0.03	1.12	1.12
	AMR(1110)	0.01	1.59	1.59	-0.03	1.13	1.13
	AMR(1101)	0.02	1.60	1.60	-0.03	1.13	1.13
	AMR(1011)	0.01	1.59	1.59	-0.03	1.13	1.13
	AMR(0111)	0.66	1.93	2.04	0.21	1.35	1.37
	AMR(1111)	0.01	1.59	1.59	-0.03	1.13	1.13

Simulation results

Parameter	Method	$(n_A, n_B = 500)$ RB(%)	RSE	RRMSE	$(n_A, n_B = 1000)$ RB(%)	RSE	RRMSE
	Benchmark	0.02	2.85	2.85	-0.05	2.04	2.04
	Naive	12.54	2.89	12.87	12.15	2.1	12.33
	PFMI(1000)	1.96	2.84	3.45	1.84	2.02	2.73
	PFMI(0100)	21.98	2.48	22.12	21.74	1.81	21.82
	MRIPW(0011)	0.44	3.69	3.72	-0.06	2.55	2.55
25th percentile	MRFMI(1100)	0.03	2.52	2.52	-0.04	1.78	1.78
	AMR(1110)	0.83	3.03	3.14	0.19	2.14	2.15
	AMR(1101)	0.39	2.88	2.91	0.05	2.08	2.08
	AMR(1011)	0.78	2.99	3.09	0.19	2.14	2.15
	AMR(0111)	1.46	3.24	3.55	0.5	2.3	2.35
	AMR(1111)	0.77	2.99	3.08	0.2	2.15	2.16

Simulation results: Bootstrap variance estimation

Table 7: Monte Carlo Percent Relative Bias (RB) of the bootstrap variance estimator, Coverage Rate (CR) %, and Average Length (AL) of confidence intervals for the proposed estimators with $n_A = n_B = 500$.

Parameter	Method	RB(%)	CR(%)	AL
Mean	MRIPW(0011)	6.56	95.7	0.66
	MRFMI(1100)	2.28	95.3	0.52
	AMR(1110)	2.29	95.3	0.52
	AMR(1101)	2.25	95.4	0.52
	AMR(1011)	2.30	95.3	0.52
	AMR(0111)	9.00	96.4	0.66
	AMR(1111)	2.30	95.3	0.52
25th percentile	MRIPW(0011)	9.37	95.5	0.94
	MRFMI(1100)	-0.42	94.7	0.63
	AMR(1110)	7.79	95.0	0.78
	AMR(1101)	4.41	95.3	0.74
	AMR(1011)	7.58	95.1	0.77
	AMR(0111)	9.28	96.3	0.89
	AMR(1111)	7.87	95.1	0.77

Final remarks

- We considered the case of parametric/semi-parametric models.
- Mass imputation: we can easily replace these models with machine learning procedures. However, establishing the properties of the resulting point estimators is challenging.
- Propensity score weighting: Use of nonparametric participation models has been considered:
  - Kernel regression: Yuan, Li and Wu (2022): same idea as in Chen, Li and Wu (2020) $\longrightarrow$ Curse of dimensionality $\longrightarrow$ Extension to Generalized Additive Models maybe envisioned.
  - Regression trees: Beaumont, Bosa, Brennan, Charlebois and Chu (2022): again, same idea as in Chen, Li and Wu (2020) $\longrightarrow$ Dimension is less of an issue

Final remarks

- Wang, Valliant and Li (2021) proposed an alternative to the method of Chen, Li and Wu (2020)
- Simulations suggest that their method is more efficient than that of Chen, Li and Wu (2020)
- They came up with a different estimating equation;
- The extension of our method based on the method of Wang, Valliant and Li (2021) may be interesting $\longrightarrow$ Gains in efficiency?