

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Integrating probability and non-probability sampling data in presence of nonignorable participation

Yu Liu , Yukun Liu & Jing Qin

To cite this article: Yu Liu , Yukun Liu & Jing Qin (09 Oct 2026): Integrating probability and non-probability sampling data in presence of nonignorable participation, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2736943

To link to this article: https://doi.org/10.1080/24754269.2026.2736943

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 09 Oct 2026.

Submit your article to this journal

Article views: 267

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2736943

Taylor & Francis
by informa

OPEN ACCESS Check for updates

Integrating probability and non-probability sampling data in presence of nonignorable participation

Yu Liu[a], Yukun Liu[a] and Jing Qin[b]

[a]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, People's Republic of China;
[b]National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA

ABSTRACT

Non-probability samples are widely adopted in surveys for cost and efficiency, yet the absence of inclusion probabilities challenges finite population inference. A typical solution is integrating them with a reference probability sample using auxiliary information, but existing methods usually assume ignorable selection given covariates. In practice, participation often directly depends on the study variable, making selection nonignorable and standard adjustments invalid. This paper presents a unified likelihood-based framework for estimating finite population means under nonignorable non-probability sampling. The approach integrates empirical likelihood, pseudo-likelihood, and parametric modelling of the participation mechanism, and covers two common survey scenarios: (1) a reference sample with both auxiliary and study variables; (2) a reference sample with auxiliary variables only. Within this framework, we propose three estimators for the finite population mean and establish their asymptotic normality under suitable regularity conditions for survey contexts. Simulations confirm that the proposed estimators outperform existing methods in robustness and reliability when the ignorability assumption is violated. An application to the ESPACOV survey demonstrates the practical utility of the methodology.

ARTICLE HISTORY
Received 28 April 2026
Revised 2 September 2026
Accepted 16 September 2026

KEYWORDS
Non-probability sampling; empirical likelihood; nonignorable missing; probability sampling

1. Introduction

Probability sampling has long provided the theoretical foundation for survey inference, offering a principled framework for estimating finite population parameters with quantifiable uncertainty. By assigning known inclusion probabilities to population units, probability samples support unbiased or approximately unbiased estimation, valid variance estimation, and transparent measures of precision. Despite these strengths, the practical implementation of probability sampling has become increasingly challenging due to declining response rates, rising costs, and logistical constraints in large-scale surveys (Wu & Thompson, 2020).

As a result, non-probability samples, such as convenience samples, volunteer panels, quota samples, and web-based surveys, have become increasingly prevalent in survey practice. These samples are attractive for their speed, flexibility, and cost efficiency, but they lack a

CONTACT Yukun Liu ykliu@sfs.ecnu.edu.cn KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai 200062, People's Republic of China

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/24754269.2026.2736943.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

known sampling design and therefore do not support direct design-based inference. From a finite population perspective, the absence of inclusion probabilities renders naive estimators based on non-probability samples generally biased, with no straightforward means of assessing sampling variability (Baker et al., 2013).

To address these limitations, recent developments in survey methodology have focused on integrating non-probability samples with a reference probability sample drawn from the same target population. This integration framework, first advocated by Rivers (2007) and subsequently formalized by Elliott and Valliant (2017) and Chen et al. (2020), treats the probability sample as an anchor that provides auxiliary population information, or, in some settings, outcome information, needed to correct selection bias in the non-probability sample. The inferential target under this framework remains a finite population parameter, most commonly the population mean or total.

Within this setting, three broad classes of methods have emerged. The first class consists of propensity score-based adjustment methods, which model the participation mechanism of the non-probability sample using auxiliary variables observed in both samples. Representative approaches include inverse probability weighting (Chen et al., 2022; Elliott & Valliant, 2017; Valliant & Dever, 2011; Wang et al., 2021), subclassification (Lee & Valliant, 2009), kernel weighting (Wang et al., 2020), and calibration weighting (Chen et al., 2020; Liu & Valliant, 2023). These methods aim to reweight the non-probability sample so that its auxiliary variable distribution aligns with that of the finite population.

The second class comprises outcome modelling and imputation-based approaches, in which a working model for the study variable given auxiliary covariates is estimated using the non-probability sample and then projected to the probability sample or the population. Examples include mass imputation (Chen et al., 2020; Kim et al., 2021), sample matching (Rivers, 2007; Yang et al., 2021), and predictive mean matching (Chlebicki et al., 2024; Yang & Kim, 2020, PMM). While attractive in settings where outcome data are rich in the non-probability sample, these methods rely heavily on correct specification of the outcome model.

The third class consists of doubly robust estimators that combine propensity score adjustment and outcome modelling to enhance robustness to model misspecification (Chen et al., 2020). These estimators remain consistent for finite population parameters if either the participation model or the outcome model is correctly specified. Extensions to multiply robust procedures in the context of non-probability sampling have also been developed (Chen & Haziza, 2023). For comprehensive reviews, see Rao (2021) and Wu (2022).

A critical assumption underlying nearly all existing methods is that the participation mechanism governing the non-probability sample is ignorable given observed auxiliary variables. In survey applications, this assumption is often questionable. Participation may depend directly on the study variable itself, even after conditioning on observed covariates, for example, when individuals with extreme outcomes are more likely to self-select into a survey. Such behaviour induces a nonignorable, or missing-not-at-random, selection mechanism that invalidates standard weighting, calibration, and doubly robust approaches, leading to biased finite population inference.

Despite its importance, methodological work on nonignorable non-probability sampling remains limited. Kim and Morikawa (2023) studied this problem under restrictive conditions, including the availability of full population auxiliary information. Liu et al. (2024) proposed a Bayesian approach, while Liu et al. (2026) introduced a pseudo-likelihood framework for modelling nonignorable participation probabilities. However, general methods

that integrate probability and non-probability samples under nonignorable selection while retaining a clear finite population interpretation are still scarce.

In this paper, we develop a unified empirical likelihood-based hybrid likelihood framework for finite population mean estimation under nonignorable non-probability sampling. We consider two survey scenarios of practical importance: (a) a reference probability sample that observes both auxiliary variables and the study variable, and (b) a reference probability sample that observes auxiliary variables only. In both scenarios, we specify a parametric model for the non-probability sample participation mechanism, and in Scenario (b) additionally model the conditional distribution of the study variable given auxiliary covariates. By integrating these components within an empirical likelihood framework, we construct estimators that remain valid under nonignorable selection. We establish the asymptotic normality of the proposed estimators and demonstrate their finite-sample performance through simulation studies and an application to the ESPACOV survey.

The rest of the article is organized as follows. Section 2 introduces the necessary notation and the mean estimation problem for a finite population. Sections 2 and 3 present the proposed empirical likelihood-based hybrid likelihood estimation methods for the population mean under Scenarios (a) and (b), respectively. We also develop algorithms for their numerical implementation and establish their asymptotic normality. Numerical results, including a simulation study and a real data analysis, are presented in Sections 4 and 5. For clarity, all proofs and technical details are postponed to the supplementary material.

2. Methodology

2.1. Problem setup

Consider a finite population consisting of N units with $\mathcal{U} = \{1, 2, \ldots, N\}$ being the set of their indices, and the population size N is assumed known. Let Y and $\boldsymbol{X}$ denote a response variable and a vector of covariates or auxiliary variables. For each $i \in \mathcal{U}$, unit i is associated with a covariate vector $\boldsymbol{X}_i$, and a response Y_i. We wish to estimate $\mu_y = N^{-1} \sum_{i \in \mathcal{U}} Y_i$, the finite population mean of the response variable.

Suppose that we have a non-probability sample $\mathcal{S}_A$ with n_A units from $\mathcal{U}$. For each unit in the non-probability sample $\mathcal{S}_A$, both the covariate vector and the response are available. The corresponding dataset is denoted as $\{\boldsymbol{Z}_i, i \in \mathcal{S}_A\}$, where $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. Let $D_i = I(i \in \mathcal{S}_A)$ be the indicator variable for unit i included in the non-probability sample $\mathcal{S}_A$, i.e. $D_i = 1$ if $i \in \mathcal{S}_A$ and $D_i = 0$ if $i \notin \mathcal{S}_A$, $i = 1, 2, \ldots, N$. Denote

$$\pi_i^A = P(D_i = 1 | \boldsymbol{Z}_i), \quad i = 1, 2, \ldots, N,$$

the probability of unit $i \in \mathcal{S}_A$ being drawn in the non-probability sample $\mathcal{S}_A$ or its propensity score. We make the following assumptions on the selection mechanism of $\mathcal{S}_A$.

Assumption 2.1: (1) *All units have a nonzero propensity score, that is* $\pi_i^A > 0$ *for all* $1 \leq i \leq N$. (2) *The indicator variables* D_i *and* D_j *are independent given* $\boldsymbol{Z}_i$ *and* $\boldsymbol{Z}_j$ *for* $i \neq j$.

Chen et al. (2020) assumed that the propensity score π_i^A depends not on Y_i but only on $\boldsymbol{X}_i$, which corresponds to the missing at random or ignorable mechanism in missing data problems (Rubin, 1976). However, this assumption is violated when D_i remains correlated with Y_i

even after conditioning on $\boldsymbol{X}_i$, which is often the case in practice. This is referred to as non-ignorable, corresponding to the missing not at random (MNAR) mechanism (Rubin, 1976), and is also the focus of the current paper.

We impose a parametric model on the propensity score of the non-probability sample, i.e.,

$$\pi_i^A = \pi(\boldsymbol{Z}_i; \boldsymbol{\theta}), \quad i = 1, 2, \ldots, N, \tag{1}$$

where $\boldsymbol{\theta}$ is unknown, and $\pi(\cdot, \cdot)$ has a known parametric form, such as the logistic function. In the design-based framework, the finite set $\mathcal{F}_N = \{\boldsymbol{Z}_i, i \in \mathcal{U}\}$ is viewed as fixed. We treat the observations of the finite population $\mathcal{F}_N = \{\boldsymbol{Z}_i, i \in \mathcal{U}\}$ as independent and identically distributed (i.i.d.) observations from a super population with distribution function $F(\boldsymbol{z})$ and probability density/mass function $f(\boldsymbol{z})$ with $\boldsymbol{z} = (\boldsymbol{x}, y)$. Let n_A denote the cardinality of $\mathcal{S}_A$. Together with the size N of the finite population, the non-probability sample $\{\boldsymbol{Z}_i : i \in \mathcal{S}_A\}$ can be regarded as a missing dataset of the super population $\mathcal{F}_N$, where the attributes of the n_A units in $\mathcal{S}_A$ are completely observed and those of the rest $N - n_A$ units are completely missing. The probability of a unit in the finite population $\mathcal{F}_N$ being included in $\mathcal{S}_A$ is

$$\eta = P(D_i = 1) = \mathbb{E}\{P(D_i = 1 | \boldsymbol{Z}_i)\} = \int \pi(\boldsymbol{z}; \boldsymbol{\theta}) \, \mathrm{d}F(\boldsymbol{z}). \tag{2}$$

Consequently, the full likelihood based on this missing dataset is

$$(1 - \eta)^{N - n_A} \times \prod_{i \in \mathcal{S}_A} \{\pi(\boldsymbol{Z}_i; \boldsymbol{\theta}) \, \mathrm{d}F(\boldsymbol{Z}_i)\},$$

where we write $\mathrm{d}F(\boldsymbol{Z}_i)$ instead of $f(\boldsymbol{Z}_i)$ for technical convenience. The log-likelihood contribution of the non-probability sample $\mathcal{S}_A$ is

$$\ell_{A1}(F, \boldsymbol{\theta}, \eta) = \sum_{i \in \mathcal{S}_A} \{\log \, \mathrm{d}F(\boldsymbol{Z}_i) + \log \pi(\boldsymbol{Z}_i; \boldsymbol{\theta})\} + (N - n_A) \log(1 - \eta).$$

However, because the non-probability $\mathcal{S}_A$ is a biased-sample of the finite population $\mathcal{F}_N$, the propensity score $\pi(\boldsymbol{z}; \boldsymbol{\theta})$ cannot be consistently estimated from $\mathcal{S}_A$ or $\ell_{A1}(F, \boldsymbol{\theta}, \eta)$ itself; additional information or extra data are required to overcome this problem.

Suppose that a reference probability sample, denoted by $\mathcal{S}_B$, is available from an existing survey or can be obtained without much difficulty. We assume that the probability sample $\mathcal{S}_B$ is independent of the non-probability sample $\mathcal{S}_A$. Let $R_i = I(i \in \mathcal{S}_B)$ be the indicator of whether unit i is included in $\mathcal{S}_B$, with $R_i = 1$ if $i \in \mathcal{S}_B$ and $R_i = 0$ otherwise, for $i = 1, 2, \ldots, N$. The survey weight is defined as $d_i^B = 1/\pi_i^B$, where $\pi_i^B = \mathbb{E}(R_i | \mathcal{F}_N)$ is the inclusion probability of unit i under the sampling design $\mathcal{S}_B$. Let n_B denote the cardinality of $\mathcal{S}_B$. The inclusion probabilities π_i^B for a reference probability sample are known a priori.

Depending on whether the responses in the probability sample are available, we consider two scenarios of sample $\mathcal{S}_B$: Scenario (a): both $\boldsymbol{X}_i$ and Y_i are available for the units in $\mathcal{S}_B$, and Scenario (b): only $\boldsymbol{X}_i$ are available for the units in $\mathcal{S}_B$.

2.2. Scenario (a)

We first consider Scenario (a), where both $\boldsymbol{X}_i$ and Y_i are available for the units in the probability sample $\mathcal{S}_B$. This immediately implies that $F(\boldsymbol{z}) = P(\boldsymbol{Z} \leq \boldsymbol{z})$ or $P(\boldsymbol{Z})$ is identifiable because

it can be consistently estimated by $\tilde{F}(\boldsymbol{z}) = (1/N)\sum_{i\in\mathcal{S}_B} I(\boldsymbol{Z}_i \le \boldsymbol{z})/\pi_i^B$ from the probability sample. Because N is known, $\eta = P(D=1)$ is identifiable. Because $\mathcal{S}_A$ consists of $\boldsymbol{Z}_i$ with $D_i = 1$, $P(\boldsymbol{Z}|D=1)$ is identifiable. Therefore, $P(D=1|\boldsymbol{Z}) = P(\boldsymbol{Z}|D=1)P(D=1)/P(\boldsymbol{Z})$ is nonparametrically identifiable. For $\boldsymbol{\theta}$ in the parameter model $\pi(\boldsymbol{Z};\boldsymbol{\theta})$ to be identifiable, it is necessary to require that for each $P(D=1|\boldsymbol{Z})$, there exists a $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{Z};\boldsymbol{\theta}) = P(D=1|\boldsymbol{Z})$. See Condition 2.3(iv).

In this scenario, the log-likelihood contribution of $\mathcal{S}_B$ or a pseudo likelihood based on $\mathcal{S}_B$ is

$$\ell_{B1}(F) = \sum_{i\in\mathcal{S}_B} d_i^B \log\{\,\mathrm{d}F(\boldsymbol{Z}_i)\}.$$

Combining the likelihood contributions $\ell_{A1}(F,\boldsymbol{\theta},\eta)$ and $\ell_{B1}(F)$ of the non-probability and probability samples, we define a hybrid log-likelihood function

$$\ell_1(F,\boldsymbol{\theta},\eta) = \ell_{A1}(F,\boldsymbol{\theta},\eta) + \rho_B\ell_{B1}(F),$$

where $\rho_B > 0$ is a user-specified scale factor. Our method works if ρ_B is any positive constant. In this paper, we recommend taking $\rho_B = n_B/\sum_{i\in\mathcal{S}_B} d_i^B$, so that the weights of the likelihood contributions of the units in $\mathcal{S}_B$ and $\mathcal{S}_A$ are all the same.

For convenience, we assume that the observations in $\mathcal{S}_A$ and $\mathcal{S}_B$ are all different and let $\mathcal{S} = \mathcal{S}_A \cup \mathcal{S}_B$. In the principle of empirical likelihood (Liu & Fan, 2023; Owen, 1990; Qin & Lawless, 1994, EL), we propose to model F by a discrete distribution $F(\boldsymbol{z}) = \sum_{i\in\mathcal{S}} p_i I(\boldsymbol{Z}_i \le \boldsymbol{z})$, where the inequality holds element-wise. The hybrid log-EL function becomes

$$\ell_1(F,\boldsymbol{\theta},\eta) = \sum_{i\in\mathcal{S}_B} \rho_B d_i^B \log p_i + \sum_{i\in\mathcal{S}_A}\{\log p_i + \log\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})\} + (N-n_A)\log(1-\eta),$$

where

$$p_i \ge 0, \quad i\in\mathcal{S}, \quad \sum_{i\in\mathcal{S}} p_i = 1, \quad \sum_{i\in\mathcal{S}} p_i\{\pi(\boldsymbol{Z}_i;\boldsymbol{\theta}) - \eta\} = 0. \tag{3}$$

The last constraint comes from the definition in (2) of η.

For ease of presentation, we define $w_i = 1$ for $i\in\mathcal{S}_A$ and $w_i = \rho_B d_i^B$ for $i\in\mathcal{S}_B$. Then the log-likelihood can be written as

$$\ell_1(F,\boldsymbol{\theta},\eta) = \sum_{i\in\mathcal{S}_A} \log\pi(\boldsymbol{Z}_i;\boldsymbol{\theta}) + (N-n_A)\log(1-\eta) + \sum_{i\in\mathcal{S}} w_i\log(p_i).$$

Given $(\boldsymbol{\theta},\eta)$, by the Lagrange multiplier method, the hybrid log-EL $\ell_1(F,\boldsymbol{\theta},\eta)$ is maximized when

$$p_i = \frac{w_i}{n_t}\cdot\frac{1}{1+\lambda\{\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})-\eta\}}, \quad i\in\mathcal{S}, \tag{4}$$

where $n_t = n_A + n_B$ and $\lambda = \lambda(\boldsymbol{\theta},\eta)$ is the solution to

$$\sum_{i\in\mathcal{S}} \frac{w_i}{n_t}\cdot\frac{\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})-\eta}{1+\lambda\{\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})-\eta\}} = 0. \tag{5}$$

Putting (4) into $\ell_1(F,\boldsymbol{\theta},\eta)$, we have the profile log-EL (up to a constant) of $(\boldsymbol{\theta},\eta)$

$$\ell_{1p}(\boldsymbol{\theta},\eta) = \sum_{i\in\mathcal{S}_A} \log\pi(\boldsymbol{Z}_i;\boldsymbol{\theta}) + (N-n_A)\log(1-\eta) - \sum_{i\in\mathcal{S}} w_i\log[1+\lambda\{\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})-\eta\}],$$

where $\lambda = \lambda(\boldsymbol{\theta},\eta)$ is an implicit function determined by Equation (5).

Based on $\ell_{1p}(\boldsymbol{\theta}, \eta)$, we propose to estimate $(\boldsymbol{\theta}, \eta)$ by the maximum EL estimator $(\hat{\boldsymbol{\theta}}, \hat{\eta}) = \arg\max_{\boldsymbol{\theta},\eta} \ell_{1p}(\boldsymbol{\theta}, \eta)$. Let $\hat{\lambda} = \lambda(\hat{\boldsymbol{\theta}}, \hat{\eta})$ be the solution to (5) when $(\boldsymbol{\theta}, \eta)$ is replaced by $(\hat{\boldsymbol{\theta}}, \hat{\eta})$. Putting these estimators back into (4), we have

$$\hat{p}_i = \frac{w_i}{n_t} \cdot \frac{1}{1 + \hat{\lambda}\{\pi(\boldsymbol{Z}_i; \hat{\boldsymbol{\theta}}) - \hat{\eta}\}}, \quad i \in \mathcal{S}.$$

Accordingly, the proposed EL estimator for $\mu_y = N^{-1} \sum_{i \in \mathcal{U}} Y_i$ is

$$\hat{\mu}_{\mathrm{EL},1} = \sum_{i \in \mathcal{S}} \hat{p}_i Y_i = \frac{1}{n_t} \sum_{i \in \mathcal{S}} \cdot \frac{w_i Y_i}{1 + \hat{\lambda}\{\pi(\boldsymbol{Z}_i; \hat{\boldsymbol{\theta}}) - \hat{\eta}\}}. \tag{6}$$

2.3. Implementation

We present an algorithm to facilitate the implementation of our EL estimation procedure. This algorithm depends on several observations about our method. Given $\boldsymbol{\theta}$, let $\eta(\boldsymbol{\theta}) = \arg\max_{\eta} \ell_{1p}(\boldsymbol{\theta}, \eta)$ and $\lambda(\boldsymbol{\theta})$ be the solution to Equation (5) when η is replaced by $\eta(\boldsymbol{\theta})$. Also, let $\boldsymbol{\theta}^{(t)}$ denote the value of $\boldsymbol{\theta}$ in the t-th iteration, $\pi^A_{i(t)} = \pi(\boldsymbol{Z}_i; \boldsymbol{\theta}^{(t)})$ for each $i \in \mathcal{S}$, and denote $\pi^{(t)}_{\mathrm{low}} = \min_{i \in \mathcal{S}} \pi^A_{i(t)}$ and $\pi^{(t)}_{\mathrm{up}} = \max_{i \in \mathcal{S}} \pi^A_{i(t)}$.

Setting the partial derivative of $\ell_{1p}(\boldsymbol{\theta}, \eta)$ with respect to η to zero, we have

$$0 = -\frac{N - n_A}{1 - \eta} + \sum_{i \in \mathcal{S}} w_i \frac{\lambda}{1 + \lambda\{\pi(\boldsymbol{Z}_i; \boldsymbol{\theta}) - \eta\}}.$$

This together with Equation (4) and $\sum_{i \in \mathcal{S}} p_i = 1$ implies

$$0 = -\frac{N - n_A}{1 - \eta} + n_t \lambda \Longrightarrow \lambda = \frac{N - n_A}{n_t(1 - \eta)}.$$

Putting this λ back into Equation (4), it follows that $\eta(\boldsymbol{\theta})$ must be the solution to

$$\sum_{i \in \mathcal{S}} \frac{w_i}{n_t} \frac{1}{1 + \frac{N - n_A}{n_t(1-\eta)}\{\pi(\boldsymbol{Z}_i; \boldsymbol{\theta}) - \eta\}} = 1 \tag{7}$$

in the interval $(\pi^{(t)}_{\mathrm{low}}, \pi^{(t)}_{\mathrm{up}})$. Then the profile log-likelihood of $\boldsymbol{\theta}$ is $\ell_{1p}(\boldsymbol{\theta}, \eta(\boldsymbol{\theta}))$. It can be verified that the partial derivative of $\ell_{1p}(\boldsymbol{\theta}, \eta(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ is

$$\nabla_{\boldsymbol{\theta}} \ell_{1p}(\boldsymbol{\theta}, \eta(\boldsymbol{\theta})) = \sum_{i \in \mathcal{S}_A} \frac{\nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{Z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{Z}_i; \boldsymbol{\theta})} - \sum_{i \in \mathcal{S}} w_i \frac{\lambda(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{Z}_i; \boldsymbol{\theta})}{1 + \lambda(\boldsymbol{\theta})\{\pi(\boldsymbol{Z}_i; \boldsymbol{\theta}) - \eta(\boldsymbol{\theta})\}}. \tag{8}$$

The proposed EL estimation procedure for μ_y can be implemented via Algorithm 1.

2.4. Asymptotics

When studying the large-sample properties of the proposed EL estimator $\hat{\mu}_{\mathrm{EL}}$, we adopt the following asymptotic framework that is commonly used in the literature of survey sampling. Assume there is a sequence of finite populations, U_v, indexed by v, with population

Algorithm 1 The proposed EL estimation procedure when $(\boldsymbol{Z}_i)$ are available in $\mathcal{S}_B$

Require: $\{\boldsymbol{Z}_i : i \in \mathcal{S}_A\}$, $\{(\boldsymbol{Z}_i, \pi_i^B) : i \in \mathcal{S}_B\}$, N, an initial value $\boldsymbol{\theta}^{(0)}$ and a tolerance.
Ensure: EL estimate $\hat{\mu}_{\mathrm{EL},1}$ for $\mu = (1/N)\sum_{i\in\mathcal{F}_\mathcal{N}} Y_i$.

1: Set $t = 0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$, and $\mathcal{S} = \mathcal{S}_A \cup \mathcal{S}_B$.
2: **while** the tolerance is not met **do**
3: Calculate $\pi_{i(t)}^A = \pi(\boldsymbol{Z}_i; \boldsymbol{\theta}^{(t)})$ for each $i \in \mathcal{S}$.
4: Set $\pi_{\mathrm{low}}^{(t)} = \min_{i\in\mathcal{S}} \pi_{i(t)}^A$ and $\pi_{\mathrm{up}}^{(t)} = \max_{i\in\mathcal{S}} \pi_{i(t)}^A$.
5: Calculate $\eta(\boldsymbol{\theta}^{(t)})$ by solving Equation (7) in the interval $(\pi_{\mathrm{low}}^{(t)}, \pi_{\mathrm{up}}^{(t)})$.
6: Calculate the gradient in (8) and update $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \alpha \nabla_{\boldsymbol{\theta}} \ell_{1p}(\boldsymbol{\theta}^{(t)}, \eta(\boldsymbol{\theta}^{(t)}))$. Here α is chosen such that $\nabla_{\boldsymbol{\theta}} \ell_{1p}(\boldsymbol{\theta}^{(t+1)}, \eta(\boldsymbol{\theta}^{(t+1)})) > \nabla_{\boldsymbol{\theta}} \ell_{1p}(\boldsymbol{\theta}^{(t)}, \eta(\boldsymbol{\theta}^{(t)}))$.
7: $t \leftarrow t + 1$
8: **end while**
9: Denote the final $\boldsymbol{\theta}^{(t)}$ as $\hat{\boldsymbol{\theta}}$. Let $\hat{\eta} = \eta(\hat{\boldsymbol{\theta}})$ and $\hat{\lambda} = \lambda(\hat{\boldsymbol{\theta}})$.
10: Calculate $\hat{p}_i = w_i n_t^{-1}[1 + \hat{\lambda}\{\pi(\boldsymbol{Z}_i; \hat{\boldsymbol{\theta}}) - \hat{\eta}\}]^{-1}$ for $i \in \mathcal{S}$.
11: **return** $\hat{\mu}_{\mathrm{EL},1} = \sum_{i\in\mathcal{S}} \hat{p}_i Y_i$.

size N_ν, $\nu = 1, 2, \ldots$. For a particular population U_ν, a non-probability sample $\mathcal{S}_{A,\nu}$ of size $n_{A,\nu}$ and a probability sample $\mathcal{S}_{B,\nu}$ of size $n_{B,\nu}$ are taken from the population. The population sizes $N_\nu \to \infty$ and the sample sizes $n_{A,\nu} \to \infty$ and $n_{B,\nu} \to \infty$ as $\nu \to \infty$. The index ν is suppressed throughout for notational simplicity; all limiting processes are understood as governed by $\nu \to \infty$. This convention is maintained in the remainder of the paper. See Wu & Thompson (2020) for further discussion.

When both the covariates $\boldsymbol{X}_i$ and the responses Y_i are available in $\mathcal{S}_B$, we establish the asymptotic normality of $\hat{\mu}_{\mathrm{EL},1}$ in Theorem 2.1 under the following regularity conditions.

Condition 2.1: (i) $\mathbb{E}(D_1) = \eta_0 \in (0, 1)$. (ii) The finite population size N and the sample size n_B satisfy $n_B/N = f_B$ for a constant $f_B \in (0, 1)$.

Condition 2.2: $\min_{i\in\mathcal{S}_B} \pi_i^B \geq \pi_0 > 0$ and $\limsup_{\nu\to\infty}\{n_B \max_{i\neq j\in\mathcal{S}_B} |\pi_{ij}^B - \pi_i^B \pi_j^B|\} < \infty$.

Condition 2.3: (i) The function $\pi(\boldsymbol{z}; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$; (ii) There exists $M_0, M_1 > 0$ such that $\pi(\boldsymbol{z}; \boldsymbol{\theta}_0) \geq M_0$ and $\|\nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{z}; \boldsymbol{\theta}_0)\| \leq M_1$ for all $\boldsymbol{z}$. (iii) $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\epsilon_0} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} \pi(\boldsymbol{Z}; \boldsymbol{\theta})\| < M_2$ for some $M_2 > 0$ and $\epsilon_0 > 0$. (iv) For any $P(\boldsymbol{z})$, there exists a unique $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{z}; \boldsymbol{\theta}) = P(\boldsymbol{z})$ for all $\boldsymbol{z}$.

Condition 2.1 implies that the sizes of the non-probability sample $\mathcal{S}_A$ and the probability sample $\mathcal{S}_B$ are comparable with that of the finite population under study. Condition 2.2 is a common condition imposed on the first-order and second-order inclusion probabilities on the probability sample. The same conditions were used in Breidt and Opsomer (2002), where further comments on Condition 2.2 are provided. Condition 2.3 is imposed for technical convenience.

For ease of presentation, we use $(\boldsymbol{\theta}_0, \eta_0)$ and μ_0 as the true values of $(\boldsymbol{\theta}, \eta)$ and μ_y, respectively. We denote $A^{\otimes 2} = AA^\top$ for a vector or matrix A and $\pi_{ij}^B = \mathbb{E}(R_i R_j | \mathcal{F}_N)$ for all $1 \leq i, j \leq N$, which implies $\pi_{ii}^B = \pi_i^B$. Let $\pi(\boldsymbol{Z}) = \pi(\boldsymbol{Z}, \boldsymbol{\theta}_0)$, $\dot{\pi}(\boldsymbol{Z}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{Z}, \boldsymbol{\theta})$, and

$\dot{\pi}(\boldsymbol{Z}) = \dot{\pi}(\boldsymbol{Z}, \boldsymbol{\theta}_0)$. Define

$$H_{11} = \mathbb{E}\left\{\frac{1}{\pi(\boldsymbol{Z}) + f_B}\right\}, \quad H_{\dot{\pi}1} = \mathbb{E}\left\{\frac{\dot{\pi}(\boldsymbol{Z})}{\pi(\boldsymbol{Z}) + f_B}\right\}, \quad H_{\dot{\pi}\dot{\pi}1} = \mathbb{E}\left[\frac{\{\dot{\pi}(\boldsymbol{Z})\}^{\otimes 2}}{\pi(\boldsymbol{Z}) + f_B}\right],$$

and

$$H_{\dot{\pi}\dot{\pi}0} = \mathbb{E}\left[\frac{\{\dot{\pi}(\boldsymbol{Z})\}^{\otimes 2}}{\pi(\boldsymbol{Z})}\right], \quad H_{\pi y} = \frac{1}{1-\eta_0}\mathbb{E}\left\{\frac{1-\pi(\boldsymbol{Z})}{f_B + \pi(\boldsymbol{Z})}Y\right\}, \quad H_{\dot{\pi}y} = \mathbb{E}\left\{\frac{\dot{\pi}(\boldsymbol{Z})}{f_B + \pi(\boldsymbol{Z})}Y\right\}.$$

Define $\boldsymbol{a}(\boldsymbol{Z}) = (1, a_{11}(\boldsymbol{Z}), a_{\dot{\pi}1}(\boldsymbol{Z}), a_{y1}(\boldsymbol{Z}))$, where

$$a_{11}(\boldsymbol{Z}) = \frac{1}{f_B + \pi(\boldsymbol{Z})}, \quad a_{\dot{\pi}1}(\boldsymbol{Z}) = \frac{\dot{\pi}(\boldsymbol{Z})}{f_B + \pi(\boldsymbol{Z})}, \quad a_{y1}(\boldsymbol{Z}) = \frac{Y}{f_B + \pi(\boldsymbol{Z})},$$

and $\Sigma_{N1} = (1/N)\sum_{i,j=1}^{N} \boldsymbol{a}(\boldsymbol{Z}_i)\boldsymbol{a}^{\top}(\boldsymbol{Z}_j)(\pi_{ij}^B - \pi_i^B\pi_j^B)/(\pi_i^B\pi_j^B)$.

Condition 2.4: (i) $\mathcal{S}_A$ and $\mathcal{S}_B$ are conditionally independent of each other given $\mathcal{F}_N$. (ii) As $\nu \to \infty$, $\Sigma_{N1}^{-1/2}N^{-1/2}\sum_{i=1}^{N}\boldsymbol{a}(\boldsymbol{Z}_i)(R_i - \pi_i^B)/\pi_i^B \xrightarrow{d} N(0, I)$. (iii) There exists a positive definite matrix Σ_1 such that $\Sigma_{N1} = \Sigma_1 + o_p(1)$.

Condition 2.4(i) is mild as the non-probability sample $\mathcal{S}_A$ and the probability sample $\mathcal{S}_B$ were usually obtained independently. Condition 2.4(ii) is commonly adopted in the community of survey sampling (Kim et al., 2023), since it is nontrivial to establish a finite-dimensional central limit theorem for general sampling designs (Han & Wellner, 2021). Condition 2.4(ii) guarantees that any linear transformation of $N^{-1/2}\sum_{i=1}^{N}\boldsymbol{a}(\boldsymbol{Z}_i)(R_i - \pi_i^B)/\pi_i^B$ also follows an asymptotically normal distribution. Condition 2.4(iii) implies that the matrix Q_1 defined in Theorem 2.1 is finite and well defined.

Theorem 2.1: *Under Conditions* 2.1–2.4, *as* $N \to \infty$, $\sqrt{N}(\hat{\mu}_{\mathrm{EL},1} - \mu_y) \xrightarrow{d} N(0, Q_1)$, *where* $\xrightarrow{d}$ *stands for 'convergence in distribution' and*

$$Q_1 = \mathbb{E}\left[\{q_{12}(\boldsymbol{Z})\}^2\frac{1-\pi(\boldsymbol{Z})}{\pi(\boldsymbol{Z})}\right] + q_{13}^2\mathbb{E}\left\{(\pi(\boldsymbol{Z}_i) - \eta_0)^2\right\} + \lim_{N\to\infty}\frac{1}{N}\sum_{i,j\in\mathcal{F}_N}\frac{\pi_{ij}^B - \pi_i^B\pi_j^B}{\pi_i^B\pi_j^B}\{q_{11}(\boldsymbol{Z}_i)\}^2. \tag{9}$$

Here $q_{11}(\boldsymbol{Z})$, $q_{12}(\boldsymbol{Z})$ *and* q_{13} *are defined in* (A1)–(A3), *respectively.*

Theorem 2.1 indicates that the proposed EL estimator asymptotically follows a normal distribution under certain regularity conditions. The complicated asymptotic variance Q_1 of $\hat{\mu}_{\mathrm{EL},1}$ consists of three parts, where the first two parts are attributed to the non-probability sample and the last part is attributed to the probability sample.

3. Scenario (b)

Next, we consider Scenario (b), where only $\boldsymbol{X}$'s are available in $\mathcal{S}_B$. Let F_x denote the marginal distribution of $\boldsymbol{X}$. In this scenario, the log-likelihood contribution of $\mathcal{S}_B$ is

$$\ell_{B2}(F_x) = \sum_{i\in\mathcal{S}_B} d_i^B \log\{\mathrm{d}F_x(\boldsymbol{X}_i)\}.$$

In addition to the propensity model (1) in Section 2.1, we impose a parametric model $f(y|\boldsymbol{x},\boldsymbol{\zeta})$ on $P(Y=y|\boldsymbol{X}=\boldsymbol{x})$, the conditional density/mass function of Y given $\boldsymbol{X}$, where $\boldsymbol{\zeta}$ is a finite-dimensional parameter. Then the log-likelihood contribution of the non-probability sample $\mathcal{S}_A$ in this scenario can be written as

$$\ell_{A2}(F_x,\boldsymbol{\theta},\boldsymbol{\zeta},\eta) = \sum_{i\in\mathcal{S}_A}\{\log\,\mathrm{d}F_x(\boldsymbol{X}_i) + \log f(Y_i|\boldsymbol{X}_i,\boldsymbol{\zeta}) + \log\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})\} + (N-n_A)\log(1-\eta).$$

Similar to Scenario (a), we define a hybrid log-likelihood function

$$\ell_2(F_x,\boldsymbol{\theta},\boldsymbol{\zeta},\eta) = \ell_{A2}(F_x,\boldsymbol{\theta},\eta) + \rho_B\ell_{B2}(F_x).$$

When F_x is modelled by $F_x(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} q_i I(\boldsymbol{X}_i \le \boldsymbol{x})$, the hybrid log-EL function becomes

$$\ell_2(F_x,\boldsymbol{\theta},\boldsymbol{\zeta},\eta) = \sum_{i\in\mathcal{S}_A}\{\log f(Y_i|\boldsymbol{X}_i,\boldsymbol{\zeta}) + \log\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})\} + (N-n_A)\log(1-\eta) + \sum_{i\in\mathcal{S}} w_i\log q_i,$$

where w_i were defined in Section 2.2 and feasible q_i satisfy

$$q_i \ge 0, \quad i\in\mathcal{S}, \quad \sum_{i\in\mathcal{S}} q_i = 1, \quad \sum_{i\in\mathcal{S}} q_i\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta}) - \eta\} = 0. \tag{10}$$

Here $\varphi(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\zeta}) = \int \pi(\boldsymbol{x},y;\boldsymbol{\theta}) f(y|\boldsymbol{x},\boldsymbol{\zeta})\,\mathrm{d}y$ and the last constraint in (10) follows from

$$\eta = \int P(D=1|\boldsymbol{X}=x,Y=y)P(Y=y|\boldsymbol{X}=\boldsymbol{x})\,\mathrm{d}y\,\mathrm{d}F_x(\boldsymbol{x}) = \int \varphi(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\zeta})\,\mathrm{d}F_x(\boldsymbol{x}).$$

It is clear that $P(\boldsymbol{X})$ is identifiable based on the probability sample $\mathcal{S}_B$, where only $\boldsymbol{X}$ is available, and that $P(\boldsymbol{X}, Y \mid D=1)$ and η are identifiable based on N and the non-probability sample $\mathcal{S}_A$. Note that

$$P(\boldsymbol{X}, Y \mid D=1) = \frac{P(D=1 \mid \boldsymbol{X}, Y)P(Y \mid \boldsymbol{X})P(\boldsymbol{X})}{P(D=1)} = \pi(\boldsymbol{X}, Y;\boldsymbol{\theta}) f(Y|\boldsymbol{X},\boldsymbol{\zeta}) \times \frac{P(\boldsymbol{X})}{\eta}. \tag{11}$$

Therefore, for $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ to be identifiable, it is necessary to assume that for any $P(\boldsymbol{X}, Y \mid D=1)$, there exists only one pair of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ such that Equation (11) holds. See Condition 3.1.

Given $(\boldsymbol{\theta},\boldsymbol{\zeta},\eta)$, $\ell_2(F_x,\boldsymbol{\theta},\boldsymbol{\zeta},\eta)$ is maximized when

$$q_i = \frac{w_i}{n_t}\frac{1}{1+\lambda\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta}) - \eta\}}, \quad i\in\mathcal{S},$$

where $\lambda = \lambda(\boldsymbol{\theta},\boldsymbol{\zeta},\eta)$ is the solution to

$$\sum_{i\in\mathcal{S}} \frac{w_i}{n_t}\frac{\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta}) - \eta\}}{1+\lambda\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta}) - \eta\}} = 0. \tag{12}$$

Accordingly, the profile log-EL function of $(\boldsymbol{\theta},\boldsymbol{\zeta},\eta)$ is (up to a constant)

$$\ell_{2p}(\boldsymbol{\theta},\boldsymbol{\zeta},\eta) = \sum_{i\in\mathcal{S}_A}\{\log f(Y_i|\boldsymbol{X}_i,\boldsymbol{\zeta}) + \log\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})\} + (N-n_A)\log(1-\eta)$$

Algorithm 2 The proposed EL estimation procedure when only $\boldsymbol{X}_i$ are available in $\mathcal{S}_B$

Require: $\{\boldsymbol{Z}_i : i \in \mathcal{S}_A\}$, $\{(\boldsymbol{X}_i, \pi_i^B) : i \in \mathcal{S}_B\}$, N, initial value $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\zeta}^{(0)})$ and a tolerance.
Ensure: EL estimates $\hat{\mu}_{\mathrm{EL},2}$ and $\hat{\mu}_{\mathrm{EL},3}$ for $\mu = (1/N)\sum_{i\in\mathcal{F}_\mathcal{N}} Y_i$.

1: Set $t = 0$, $(\boldsymbol{\theta}, \boldsymbol{\zeta}) = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\zeta}^{(0)})$ and $\mathcal{S} = \mathcal{S}_A \cup \mathcal{S}_B$.
2: **while** the tolerance criterion is not met **do**
3: Calculate $\varphi_{i(t)}^A = \varphi(\boldsymbol{X}_i; \boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)})$ for each $i \in \mathcal{S}$.
4: Set $\varphi_{\mathrm{low}}^{(t)} = \min_{i\in\mathcal{S}} \varphi_{i(t)}^A$ and $\varphi_{\mathrm{up}}^{(t)} = \max_{i\in\mathcal{S}} \varphi_{i(t)}^A$.
5: Calculate $\eta(\boldsymbol{\theta}, \boldsymbol{\zeta})$ by solving Equation (15) in the interval $(\varphi_{\mathrm{low}}^{(t)}, \varphi_{\mathrm{up}}^{(t)})$.
6: Calculate the gradient in (16); Update $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ by $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \alpha \nabla_{\boldsymbol{\theta}} \ell_{2p}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)}, \eta(\boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)}))$ and $\boldsymbol{\zeta}^{(t+1)} = \boldsymbol{\zeta}^{(t)} + \alpha \nabla_{\boldsymbol{\zeta}} \ell_{2p}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)}, \eta(\boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)}))$ with a carefully chosen α.
7: $t \leftarrow t + 1$
8: **end while**
9: Denote the final $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\zeta}^{(t)}$ as $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\zeta}}$, respectively. Let $\hat{\eta} = \eta(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}})$ and $\hat{\lambda} = \lambda(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}})$.
10: Calculate $\hat{q}_i = w_i n_t^{-1}[1 + \hat{\lambda}\{\varphi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}}) - \hat{\eta}\}]^{-1}$ for $i \in \mathcal{S}$.
11: **return** $\hat{\mu}_{\mathrm{EL},2}$ in (13) and $\hat{\mu}_{\mathrm{EL},3}$ in (14).

$$-\sum_{i\in\mathcal{S}} w_i \log[1 + \lambda\{\varphi(\boldsymbol{X}_i; \boldsymbol{\theta}, \boldsymbol{\zeta}) - \eta\}].$$

Then we estimate $(\boldsymbol{\theta}, \boldsymbol{\zeta}, \eta)$ by its MLE $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}}, \hat{\eta}) = \arg\max_{\boldsymbol{\theta},\boldsymbol{\zeta},\eta} \ell_{2p}(\boldsymbol{\theta}, \boldsymbol{\zeta}, \eta)$. The corresponding q_i are

$$\hat{q}_i = \frac{w_i}{n_t} \frac{1}{1 + \hat{\lambda}\{\varphi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}}) - \hat{\eta}\}}, \quad i \in \mathcal{S},$$

where $\hat{\lambda} = \lambda(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}})$. Finally, we propose to estimate μ by

$$\hat{\mu}_{\mathrm{EL},2} = \sum_{i\in\mathcal{S}} \hat{q}_i \int y f(y|\boldsymbol{X}_i, \hat{\boldsymbol{\zeta}})\, \mathrm{d}y, \tag{13}$$

or

$$\hat{\mu}_{\mathrm{EL},3} = \sum_{i\in\mathcal{S}_A} \hat{q}_i Y_i + \sum_{i\in\mathcal{S}_B} \hat{q}_i \int y f(y|\boldsymbol{X}_i, \hat{\boldsymbol{\zeta}})\, \mathrm{d}y. \tag{14}$$

The above estimators both depend crucially on the function $\varphi(x; \boldsymbol{\theta}, \boldsymbol{\zeta})$. In general, $\varphi(x; \boldsymbol{\theta}, \boldsymbol{\zeta})$ involves an integral and does not admit a closed-form expression; therefore, it must be evaluated numerically. For example, under the working normal model $Y \mid X = x \sim N(\mu(x; \boldsymbol{\zeta}), \sigma^2)$, we employ the Gauss-Hermite quadrature to approximate the integral.

3.1. Implementation

Given $(\boldsymbol{\theta}, \boldsymbol{\zeta})$, let $\eta(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \arg\max_\eta \ell_{2p}(\boldsymbol{\theta}, \boldsymbol{\zeta}, \eta)$ and $\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ be the solution to Equation (12) when η is replaced by $\eta(\boldsymbol{\theta}, \boldsymbol{\zeta})$. Again, let $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)})$ denote the value of $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ in the t-th iteration, $\varphi_{i(t)}^A = \varphi(\boldsymbol{X}_i; \boldsymbol{\theta}^{(t)}, \boldsymbol{\zeta}^{(t)})$ for each $i \in \mathcal{S}$, and denote $\varphi_{\mathrm{low}}^{(t)} = \min_{i\in\mathcal{S}} \varphi_{i(t)}^A$ and $\varphi_{\mathrm{up}}^{(t)} =$

$\max_{i\in\mathcal{S}}\varphi_{i(t)}^{A}$. With similar algebra in Section 2.3, it can be found that $\eta(\boldsymbol{\theta},\boldsymbol{\zeta})$ is the solution to

$$\sum_{i\in\mathcal{S}}\frac{w_i}{n_t}\frac{1}{1+\frac{N-n_A}{n_t(1-\eta)}\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta})-\eta\}}=1 \tag{15}$$

in the interval $(\varphi_{\text{low}}^{(t)},\ \varphi_{\text{up}}^{(t)})$. Then the profile log-likelihood of $(\boldsymbol{\theta},\boldsymbol{\zeta})$ is $\ell_{2p}(\boldsymbol{\theta},\boldsymbol{\zeta},\eta(\boldsymbol{\theta},\boldsymbol{\zeta}))$. It can be verified that the partial derivatives of $\ell_{2p}(\boldsymbol{\theta},\boldsymbol{\zeta},\eta(\boldsymbol{\theta},\boldsymbol{\zeta}))$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are

$$\begin{cases}\nabla_{\boldsymbol{\theta}}\ell_{2p}(\boldsymbol{\theta},\boldsymbol{\zeta},\eta(\boldsymbol{\theta},\boldsymbol{\zeta}))=\displaystyle\sum_{i\in\mathcal{S}_A}\frac{\nabla_{\boldsymbol{\theta}}\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{Z}_i;\boldsymbol{\theta})}-\sum_{i\in\mathcal{S}}w_i\frac{\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta})}{1+\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta})-\eta(\boldsymbol{\theta},\boldsymbol{\zeta})\}},\\ \nabla_{\boldsymbol{\zeta}}\ell_{2p}(\boldsymbol{\theta},\boldsymbol{\zeta},\eta(\boldsymbol{\theta},\boldsymbol{\zeta}))=\displaystyle\sum_{i\in\mathcal{S}_A}\frac{\nabla_{\boldsymbol{\zeta}}f(Y_i|\boldsymbol{X}_i,\boldsymbol{\zeta})}{f(Y_i|\boldsymbol{X}_i,\boldsymbol{\zeta})}-\sum_{i\in\mathcal{S}}w_i\frac{\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta})}{1+\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})\{\varphi(\boldsymbol{X}_i;\boldsymbol{\theta},\boldsymbol{\zeta})-\eta(\boldsymbol{\theta},\boldsymbol{\zeta})\}},\end{cases} \tag{16}$$

respectively. The proposed estimators $\hat{\mu}_{\text{EL},2}$ and $\hat{\mu}_{\text{EL},3}$ can be calculated via Algorithm 2.

3.2. Asymptotics

Let $\varphi(\boldsymbol{X})=\varphi(\boldsymbol{X};\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$, $\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X})=\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X};\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$ and $\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X})=\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X};\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$. Define $\boldsymbol{b}(\boldsymbol{X})=(1,b_{11}(\boldsymbol{X}),b_{\varphi\dot{\boldsymbol{\theta}}}(\boldsymbol{X}),b_{\varphi\dot{\boldsymbol{\zeta}}}(\boldsymbol{X}),b_{y1}(\boldsymbol{X}),b_{y2}(\boldsymbol{X}))$, where

$$b_{11}(\boldsymbol{X})=\frac{1}{f_B+\varphi(\boldsymbol{X})},\quad b_{\varphi\dot{\boldsymbol{\theta}}}(\boldsymbol{X})=\frac{\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X})}{f_B+\varphi(\boldsymbol{X})},\quad b_{\varphi\dot{\boldsymbol{\zeta}}}(\boldsymbol{X})=\frac{\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X})}{f_B+\varphi(\boldsymbol{X})},$$
$$b_{y1}(\boldsymbol{X})=\frac{Y}{f_B+\varphi(\boldsymbol{X})},\quad b_{y2}(\boldsymbol{X})=\frac{\int yf(y|\boldsymbol{X},\boldsymbol{\zeta}_0)\,\mathrm{d}y}{f_B+\varphi(\boldsymbol{X})}.$$

Also, define $\Sigma_{N2}=(1/N)\sum_{i,j=1}^{N}\boldsymbol{b}(\boldsymbol{X}_i)\boldsymbol{b}^{\top}(\boldsymbol{X}_j)(\pi_{ij}^{B}-\pi_i^{B}\pi_j^{B})/(\pi_i^{B}\pi_j^{B})$. Our asymptotic results on $\hat{\mu}_{\text{EL},2}$ and $\hat{\mu}_{\text{EL},3}$ are built on Conditions 3.1, 3.2 and 3.3.

Condition 3.1: For any $P(\boldsymbol{X},Y\mid D=1)$, there exists only one pair of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ such that (11) holds.

Condition 3.2: (i) The function $\varphi(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\zeta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. (ii) There exists $N_0,N_1,N_2>0$ such that $\varphi(\boldsymbol{x};\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\geq N_0$, $\|\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{x};\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\|\leq N_1$ and $\|\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{x};\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\|\leq N_2$ for all $(\boldsymbol{x},y)$. (iii) $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\epsilon_0}\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^{\top}}\varphi(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{\zeta})\|<N_3$ and $\sup_{\boldsymbol{\zeta}:\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|\leq\epsilon_0}\|\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}}\varphi(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{\zeta})\|<N_4$ for some $N_3>0$, $N_4>0$ and $\epsilon_0>0$.

Condition 3.3: (i) As $\nu\to\infty$, $\Sigma_{N2}^{-1/2}N^{-1/2}\sum_{i=1}^{N}\boldsymbol{b}(\boldsymbol{X})(R_i-\pi_i^{B})/\pi_i^{B}\xrightarrow{d}N(0,I)$. (ii) There exists a positive definite matrix Σ_2 such that $\Sigma_{N2}=\Sigma_2+o_p(1)$.

Condition 3.1 guarantees that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are identifiable. Condition 3.2 requires that the function $\varphi(x;\boldsymbol{\theta},\boldsymbol{\zeta})$ has bounded first-order partial derivatives at $(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$ and uniformly bounded second-order partial derivatives in a neighbourhood of $(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$. These conditions are used to derive linear approximations of $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\zeta}})$, which lay the foundation for deriving the limiting distributions of $\hat{\mu}_{\text{EL},2}$ and $\hat{\mu}_{\text{EL},3}$. Condition 3.3 guarantees that the asymptotic variances Q_2 and Q_3 in Theorem 3.1 are well defined.

Theorem 3.1: *Under Conditions* 2.1, 2.2, 3.1, 3.2 *and* 3.3, *as* $N \to \infty$,

(i) $\sqrt{N}(\hat{\mu}_{\mathrm{EL},2} - \mu_y) \xrightarrow{d} N(0, Q_2)$, *where* $\xrightarrow{d}$ *stands for 'convergence in distribution' and*

$$Q_2 = \mathbb{E}\left\{\{1 - \pi(\boldsymbol{Z})\}\pi(\boldsymbol{Z})\{q_{22}(\boldsymbol{X})\}^{\otimes 2}\right\} + \mathbb{E}\{q_{23}(\boldsymbol{X})\}^{\otimes 2} + \lim_{N\to\infty} \frac{1}{N} \sum_{i,j\in\mathcal{F}_N} \frac{\pi_{ij}^B - \pi_i^B \pi_j^B}{\pi_i^B \pi_j^B} \{q_{21}(\boldsymbol{X}_i)\}^{\otimes 2}. \tag{17}$$

Here $q_{21}(\boldsymbol{X})$, $q_{22}(\boldsymbol{X})$ *and* $q_{23}(\boldsymbol{X})$ *are defined in* (A4)–(A6), *respectively.*

(ii) $\sqrt{N}(\hat{\mu}_{\mathrm{EL},3} - \mu_y) \xrightarrow{d} N(0, Q_3)$, *where*

$$Q_3 = \mathbb{E}\left\{\{1 - \pi(\boldsymbol{Z})\}\pi(\boldsymbol{Z})\{q_{32}(\boldsymbol{X})\}^{\otimes 2}\right\} + \mathbb{E}\{q_{33}(\boldsymbol{X})\}^{\otimes 2} + \lim_{N\to\infty} \frac{1}{N} \sum_{i,j\in\mathcal{F}_N} \frac{\pi_{ij}^B - \pi_i^B \pi_j^B}{\pi_i^B \pi_j^B} \{q_{31}(\boldsymbol{X}_i)\}^{\otimes 2}. \tag{18}$$

Here $q_{31}(\boldsymbol{X})$, $q_{32}(\boldsymbol{X})$ *and* $q_{33}(\boldsymbol{X})$ *are defined in* (A7)–(A9), *respectively.*

Similar to Theorem 2.1, the asymptotic variances Q_2 and Q_3 of $\hat{\mu}_{\mathrm{EL},2}$ and $\hat{\mu}_{\mathrm{EL},3}$ admit a decomposition into three parts, with the first two parts contributed by the non-probability sample and the last part contributed by the probability sample.

4. Simulation study

In this section, we carry out simulations to investigate the finite-sample performance of the proposed EL estimators.

4.1. Simulation settings

We consider a trivariate covariate vector $\boldsymbol{X} = (X_1, X_2, X_3)^\top$, where X_1 follows the Bernoulli distribution with success probability 0.4, X_2 follows a normal distribution $N(1, 0.5)$, and X_3 follows a uniform distribution $U(0, 4)$. Let ϵ denote a standard normal random variable. Given $\boldsymbol{X}$, we may generate the response Y from

(1) Model ξ_1, $Y = 2 + 1.2X_1 + 0.6\exp(X_2/2) + 0.8\log(X_3^2 + 1) + 0.3\epsilon$, or
(2) Model ξ_2, $Y = 2 + 1.2X_1 + 0.6X_2 + 0.8X_3 + 0.3\epsilon$.

From the above super population, we generate $N = 20000$ i.i.d. observations, say $\{(Y_i, \boldsymbol{X}_i), 1 \le i \le N\}$ with $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})^\top$, to form a finite population $\mathcal{U}$.

We may set the true propensity scores $\pi_i^A = P(D_i = 1|\boldsymbol{X}_i, Y_i)$ for the non-probability sample $\mathcal{S}_A$ to satisfy one of the following:

(1) Model q_1, $\mathrm{logit}(\pi_i^A) = \alpha + 0.4\cos(X_{i1}) - 0.5X_{i2} - 0.4X_{i3} + 0.1X_{i1}X_{i3} + Y_i$,
(2) Model q_2, $\mathrm{logit}(\pi_i^A) = \alpha - 0.5X_{i2} - 0.4X_{i3} + Y_i$,
(3) Model q_3, $\mathrm{logit}(\pi_i^A) = \alpha + 0.3X_{i1} + 0.2X_{i2} + 0.4X_{i3}$,

Table 1. Data generating processes considered in our simulation study.

Case	TT1	FT1	TT	FT	TF	FF
$P(Y=y\|\boldsymbol{X}=\boldsymbol{x})$	ξ_1	ξ_2	ξ_1	ξ_2	ξ_1	ξ_2
$P(D=1\|\boldsymbol{X}=\boldsymbol{x},Y=y)$	q_1	q_1	q_2	q_2	q_3	q_3

where α is chosen such that $\sum_{i=1}^{N}\pi_i^A = n_A$. Then a non-probability sample is selected by the Poisson sampling method with inclusion probabilities specified by π_i^A and the target sample size n_A. A probability sample $\mathcal{S}_B$ with size n_B is drawn from the finite population using simple random sampling without replacement. In this case, the first-order inclusion probabilities are $\pi_i^B = n_B/N$ and the second-order inclusion probabilities are $\pi_{i,j}^B = n_B(n_B-1)/N(N-1)$.

We generate data from the six data generating processes in Table 1. The missingness mechanisms in the first two data generating processes are MAR, while those in the remaining four are MNAR. In the proposed estimation procedure, we set $f(y|\boldsymbol{x},\zeta)$ to be the density of $N(\zeta_0 + \zeta_1 X_1 + \zeta_2\exp(X_2/2) + \zeta_3\log(X_3^2+1),\ \zeta_4)$ and choose $\pi(\boldsymbol{z},\theta)$ such that $\text{logit}\{\pi(\boldsymbol{Z}_i,\theta)\} = \theta_0 + \theta_1\cos(X_{i1}) + \theta_2 X_{i2} + \theta_3 X_{i3} + \theta_4 X_{i1}X_{i3} + \theta_5 Y_i$. Therefore, in case (TT), both assumed models are correct (True), whereas in case (TF), the assumed model for $f(y|\boldsymbol{x},\zeta)$ is correct (True) but that for $\pi(\boldsymbol{z},\theta)$ is mis-specified (False). The parameter of interest is the mean $\mu_y = N^{-1}\sum_{i=1}^{N} Y_i$ of the finite population. We set the sample size pair (n_A, n_B) to (500, 500) or (1000, 1000).

4.2. Methods under comparison

Let $\tilde{\pi}_i^A$ denote Chen et al. (2020)'s estimates of π_i^A under the logistic model $\text{logit}(\pi_i^A) = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3}$. In this simulation study, we compare the proposed EL estimators $\hat{\mu}_{\text{EL},1}$, $\hat{\mu}_{\text{EL},2}$ and $\hat{\mu}_{\text{EL},3}$ with the following competitors.

(1) The naive estimator based on $\mathcal{S}_A$: $\hat{\mu}_{\text{naive}} = (1/n_A)\sum_{i\in\mathcal{S}_A} Y_i$.
(2) The IPW estimator based on $\mathcal{S}_A$: $\hat{\mu}_{\text{IPWA1}} = (1/N)\sum_{i\in\mathcal{S}_A} Y_i/\tilde{\pi}_i^A$.
(3) The stabilized IPW estimator based on $\mathcal{S}_A$: $\hat{\mu}_{\text{IPWA2}} = (\sum_{i\in\mathcal{S}_A} Y_i/\tilde{\pi}_i^A)/(\sum_{j\in\mathcal{S}_A} 1/\tilde{\pi}_j^A)$.
(4) The IPW estimator based on $\mathcal{S}_B$: $\hat{\mu}_{\text{IPWB}} = (1/N)\sum_{i\in\mathcal{S}_B} Y_i/\pi_i^B$.
(5) Chen et al. (2020)'s double robust estimator $\hat{\mu}_{\text{CHEN}}$.

The above estimators can be categorized into three groups. Group 1 consists of $\hat{\mu}_{\text{EL},1}$ and $\hat{\mu}_{\text{IPWB}}$, which are designed for Scenario (a) and utilize the responses in $\mathcal{S}_B$. Group 2 comprises $\hat{\mu}_{\text{EL},2}$, $\hat{\mu}_{\text{EL},3}$, and $\hat{\mu}_{\text{CHEN}}$, which are designed for Scenario (b). Group 3 includes $\hat{\mu}_{\text{IPWA1}}$, $\hat{\mu}_{\text{IPWA2}}$, and $\hat{\mu}_{\text{naive}}$, which are applicable to both scenarios.

For a generic estimator $\hat{\mu}$, we evaluate its performance by the relative bias (%RB) and the root mean square error (%RMSE), i.e.,

$$\%\text{RB} = \frac{1}{B}\sum_{j=1}^{B}\frac{\hat{\mu}^{(j)}-\mu_y}{\mu_y}\times 100, \quad \%\text{RMSE} = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\hat{\mu}^{(j)}-\mu_y)^2}\times 100,$$

where $\hat{\mu}^{(j)}$ is the estimate based on the j-th simulated sample and $B = 500$ is the number of simulation repetitions.

4.3. Interval estimation

Besides point estimation, it is important to construct confidence intervals for the population mean. We may do this by using Theorems 2.1 and 3.1 and constructing reasonable estimates for the asymptotic variances therein. However, this is formidable due to the extremely complicated form of the variances. We instead construct variance estimates for the proposed EL estimators by using the first bootstrap weights-generation method of Chauvet (2025).

We use the following procedure to construct variance estimates for our EL estimators.

(1) Draw n_A indices from $\mathcal{S}_A$ by simple random sampling with replacement (SRSWR), and let m_i denote the number of times that the index i was drawn. Set $G_i^A = n_A m_i/(n_A - 1)$.
(2) For $i \in \mathcal{S}_B$, let G_i^B be random variables satisfying $\mathbb{E}^*(G_i^B) = 1$ and $\mathbb{C}\mathrm{ov}^*(G_i^B, G_j^B) = (\pi_{ij}^B - \pi_i^B \pi_j^B)/\pi_{ij}^B$, at least approximately. The above covariance structure is chosen to mimic the dependence structure induced by the probability sampling design. In particular, it is determined by the first- and second-order inclusion probabilities and is commonly used in survey bootstrap methods to preserve the design-based variance characteristics of the probability sample.
(3) Using the weights G_i^A and G_i^B to reweight the profile log-likelihood, we shall obtain a bootstrap version of $\hat{\mu}_{\mathrm{EL,j}}, j = 1, 2, 3$. The corresponding bootstrap reweighted objective function is

$$\ell_1^*(F, \boldsymbol{\theta}, \eta) = \sum_{i \in \mathcal{S}_A} G_i^A \log \pi(\boldsymbol{Z}_i; \boldsymbol{\theta}) + (N - n_A) \log(1 - \eta) + \sum_{i \in \mathcal{S}} G_i w_i \log(p_i),$$

in Scenario (a), and

$$\ell_{2p}^*(\boldsymbol{\theta}, \boldsymbol{\zeta}, \eta) = \sum_{i \in \mathcal{S}_A} G_i^A \left\{ \log f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\zeta}) + \log \pi(\boldsymbol{Z}_i; \boldsymbol{\theta}) \right\} + (N - n_A) \log(1 - \eta) - \sum_{i \in \mathcal{S}} G_i w_i \log[1 + \lambda\{\varphi(\boldsymbol{X}_i; \boldsymbol{\theta}, \boldsymbol{\zeta}) - \eta\}],$$

in Scenario (b), where $G_i = G_i^A I(i \in \mathcal{S}_A) + G_i^B I(i \in \mathcal{S}_B)$.
(4) For $j = 1, 2, 3$, repeating the above steps B times, we shall have B bootstrap estimates, $\hat{\mu}_{\mathrm{EL,j}}^{(b)}, b = 1, 2, \ldots, B$.
(5) A $(1 - \alpha)$-level Wald-type confidence interval is $\hat{\mu} \pm 1.96\sqrt{\hat{v}_j}$, where $\hat{v}_j$ is the sample variance of $\hat{\mu}_{\mathrm{EL,j}}^{(b)}$, i.e. $\hat{v}_j = \frac{1}{B-1} \sum_{b=1}^B (\hat{\mu}_{\mathrm{EL,j}}^{(b)} - \frac{1}{B} \sum_{s=1}^B \hat{\mu}_{\mathrm{EL,j}}^{(s)})^2$.

The G_i^B in Step (2) can be generated from a probability distribution with appropriate moments; see Chauvet (2025).

4.4. Simulation results

Tables 2–4 report the simulated RB (unit: 0.01), RMSE (unit: 0.01), coverage probabilities (CP) and average interval lengths (Len) of the Wald-type intervals of all estimators under comparison. The naive estimator $\hat{\mu}_{\mathrm{naive}}$, which relies solely on the non-probability sample, consistently exhibits severe RB and extremely large RMSE. The accompanying Wald-type interval must have unacceptably low coverage probabilities and hence we do not report such

Table 2. Simulated relative bias (RB, unit: 0.01), root mean squared error (RMSE, unit: 0.01), coverage probabilities (CP) and average interval length (Len) of various estimators of μ_y when data were generated from Cases (TT1) and (FT1).

		$(n_a, n_b) = (500, 500)$				$(n_a, n_b) = (1000, 1000)$			
Case	Estimator	RB	RMSE	CP	Len	RB	RMSE	CP	Len
TT1	$\hat{\mu}_{\text{EL},1}$	**0.09**	**4.24**	**0.946**	**0.161**	**0.03**	**2.92**	**0.956**	**0.115**
	$\hat{\mu}_{\text{IPWB}}$	0.04	4.41	0.964	0.175	−0.01	3.11	0.952	0.122
	$\hat{\mu}_{\text{EL},2}$	**0.23**	**4.35**	**0.952**	**0.175**	**0.17**	**3.15**	**0.914**	**0.122**
	$\hat{\mu}_{\text{EL},3}$	**0.23**	**4.35**	**0.888**	**0.156**	**0.17**	**3.13**	**0.914**	**0.116**
	$\hat{\mu}_{\text{CHEN}}$	0.02	4.55	0.956	0.178	−0.04	3.21	0.942	0.124
	$\hat{\mu}_{\text{IPWA1}}$	0.14	4.90	0.958	0.193	0.01	3.23	0.946	0.130
	$\hat{\mu}_{\text{IPWA2}}$	−0.02	5.76	0.940	0.229	−0.07	3.97	0.954	0.158
	$\hat{\mu}_{\text{naive}}$	8.87	41.92	–	–	8.69	40.98	–	–
FT1	$\hat{\mu}_{\text{EL},1}$	**0.09**	**4.24**	**0.946**	**0.162**	**0.03**	**2.92**	**0.960**	**0.115**
	$\hat{\mu}_{\text{IPWB}}$	0.04	4.41	0.964	0.175	−0.01	3.31	0.952	0.122
	$\hat{\mu}_{\text{EL},2}$	**0.36**	**4.46**	**0.956**	**0.175**	**0.30**	**3.28**	**0.932**	**0.122**
	$\hat{\mu}_{\text{EL},3}$	**0.35**	**4.45**	**0.886**	**0.155**	**0.29**	**3.26**	**0.890**	**0.114**
	$\hat{\mu}_{\text{CHEN}}$	−0.02	4.67	0.932	0.178	0.01	3.20	0.930	0.125
	$\hat{\mu}_{\text{IPWA1}}$	0.05	4.73	0.954	0.192	0.06	3.28	0.942	0.130
	$\hat{\mu}_{\text{IPWA2}}$	−0.07	6.14	0.946	0.229	−0.01	3.96	0.962	0.158
	$\hat{\mu}_{\text{naive}}$	8.89	42.00	–	–	8.66	40.80	–	–

Table 3. Simulated relative bias (RB, unit: 0.01), root mean squared error (RMSE, unit: 0.01), coverage probabilities (CP) and average interval length (Len) of various estimators of μ_y when data were generated from Cases (TT) and (FT).

		$(n_a, n_b) = (500, 500)$				$(n_a, n_b) = (1000, 1000)$			
Case	Estimator	RB	RMSE	CP	Len	RB	RMSE	CP	Len
TT	$\hat{\mu}_{\text{EL},1}$	**0.09**	**4.40**	**0.930**	**0.170**	**0.03**	**3.16**	**0.956**	**0.138**
	$\hat{\mu}_{\text{IPWB}}$	0.04	4.41	0.964	0.175	−0.01	3.11	0.952	0.122
	$\hat{\mu}_{\text{EL},2}$	**0.19**	**4.56**	**0.946**	**0.175**	**0.07**	**3.18**	**0.954**	**0.122**
	$\hat{\mu}_{\text{EL},3}$	**0.89**	**6.15**	**0.890**	**0.189**	**0.76**	**4.79**	**0.825**	**0.135**
	$\hat{\mu}_{\text{CHEN}}$	1.85	9.70	0.538	0.181	1.80	9.07	0.232	0.126
	$\hat{\mu}_{\text{IPWA1}}$	2.61	13.51	0.464	0.238	2.38	11.76	0.138	0.146
	$\hat{\mu}_{\text{IPWA2}}$	1.51	9.76	0.782	0.265	1.44	8.33	0.644	0.183
	$\hat{\mu}_{\text{naive}}$	13.97	65.81	–	–	13.64	64.17	–	–
FT	$\hat{\mu}_{\text{EL},1}$	**0.04**	**4.38**	**0.946**	**0.178**	**−0.02**	**3.24**	**0.952**	**0.130**
	$\hat{\mu}_{\text{IPWB}}$	0.01	4.40	0.956	0.175	−0.01	3.32	0.948	0.122
	$\hat{\mu}_{\text{EL},2}$	**0.21**	**4.56**	**0.948**	**0.189**	**0.12**	**3.31**	**0.952**	**0.133**
	$\hat{\mu}_{\text{EL},3}$	**0.98**	**6.41**	**0.840**	**0.188**	**0.88**	**5.25**	**0.766**	**0.132**
	$\hat{\mu}_{\text{CHEN}}$	1.90	10.15	0.506	0.181	1.88	9.41	0.222	0.126
	$\hat{\mu}_{\text{IPWA1}}$	2.46	12.84	0.488	0.236	2.29	11.36	0.178	0.145
	$\hat{\mu}_{\text{IPWA2}}$	1.46	9.83	0.760	0.266	1.45	7.97	0.686	0.182
	$\hat{\mu}_{\text{naive}}$	14.01	65.98	–	–	13.59	63.92	–	–

results. Because the non-probability sample is biased for the target population, incorporating the probability sample is necessary to correct this bias.

Cases (TT1) and (FT1) are designed such that the missingness in $\mathcal{S}_A$ is MAR, under which Chen et al. (2020)'s estimator is constructed. Table 2 indicates that all estimators except $\hat{\mu}_{\text{naive}}$ have very close RMSEs. This is probably because the propensity model underlying Chen et al. (2020)'s estimator is correctly specified and the propensity scores $\tilde{\pi}_i^A$ are reasonably estimated. Even so, our $\hat{\mu}_{\text{EL},1}$ usually has smaller RMSEs than $\hat{\mu}_{\text{IPWB}}$ when responses are available in $\mathcal{S}_B$ (Scenario (a)), and our $\hat{\mu}_{\text{EL},2}$ usually has smaller RMSEs than $\hat{\mu}_{\text{CHEN}}$ when responses are unavailable in $\mathcal{S}_B$ (Scenario (b)). The IPW estimator $\hat{\mu}_{\text{IPWB}}$ consistently yields reasonable estimates (but requires the responses in $\mathcal{S}_B$) as it is independent of working models for $P(Y = y|\boldsymbol{X} = \boldsymbol{x})$ and $P(D = 1|\boldsymbol{X} = \boldsymbol{x}, Y = y)$. In contrast, $\hat{\mu}_{\text{CHEN}}$ does

Table 4. Simulated relative bias (RB, unit: 0.01), root mean squared error (RMSE, unit: 0.01), coverage probabilities (CP) and average interval length (Len) of various estimators of μ_y when data were generated from Cases (TF) and (FF).

		$(n_a, n_b) = (500, 500)$				$(n_a, n_b) = (1000, 1000)$			
Case	Estimator	RB	RMSE	CP	Len	RB	RMSE	CP	Len
TF	$\hat{\mu}_{\text{EL},1}$	**0.37**	**4.62**	**0.958**	**0.175**	**0.25**	**3.63**	**0.940**	**0.131**
	$\hat{\mu}_{\text{IPWB}}$	−0.10	4.47	0.966	0.174	0.07	3.06	0.947	0.122
	$\hat{\mu}_{\text{EL},2}$	**0.42**	**4.81**	**0.948**	**0.187**	**0.29**	**3.76**	**0.922**	**0.136**
	$\hat{\mu}_{\text{EL},3}$	**1.16**	**6.71**	**0.852**	**0.185**	**1.04**	**6.29**	**0.646**	**0.134**
	$\hat{\mu}_{\text{CHEN}}$	1.89	10.04	0.502	0.176	1.76	8.88	0.226	0.123
	$\hat{\mu}_{\text{IPWA1}}$	9.18	43.41	0	–	8.95	42.18	0	–
	$\hat{\mu}_{\text{IPWA2}}$	9.45	44.65	0	–	9.18	43.28	0	–
	$\hat{\mu}_{\text{naive}}$	14.04	66.13	–	–	13.62	64.09	–	–
FF	$\hat{\mu}_{\text{EL},1}$	**0.32**	**4.53**	**0.944**	**0.175**	**0.29**	**3.29**	**0.938**	**0.130**
	$\hat{\mu}_{\text{IPWB}}$	−0.01	4.51	0.948	0.175	0.03	3.14	0.948	0.122
	$\hat{\mu}_{\text{EL},2}$	**0.52**	**4.92**	**0.934**	**0.183**	**0.46**	**3.87**	**0.900**	**0.129**
	$\hat{\mu}_{\text{EL},3}$	**1.29**	**7.36**	**0.756**	**0.182**	**1.26**	**6.76**	**0.544**	**0.128**
	$\hat{\mu}_{\text{CHEN}}$	1.99	10.57	0.448	0.176	1.93	9.56	0.170	0.123
	$\hat{\mu}_{\text{IPWA1}}$	9.18	43.43	0	–	8.97	42.26	0	–
	$\hat{\mu}_{\text{IPWA2}}$	9.46	44.75	0	–	9.22	43.44	0	–
	$\hat{\mu}_{\text{naive}}$	14.02	66.04	–	–	13.60	64.01	–	–

not require the responses in $\mathcal{S}_B$, and depends on working models for $P(Y = y|\boldsymbol{X} = \boldsymbol{x})$ and $P(D = 1|\boldsymbol{X} = \boldsymbol{x}, Y = y)$. Both $\hat{\mu}_{\text{IPWA1}}$ and $\hat{\mu}_{\text{IPWA2}}$ have larger RMSEs than all the above estimators except $\hat{\mu}_{\text{naive}}$.

In the remaining four cases, the missingness in $\mathcal{S}_A$ is MNAR, which is more challenging than MAR. The proposed EL estimators continue to perform very well: $\hat{\mu}_{\text{EL},1}$ has smaller RMSEs than the benchmark estimator $\hat{\mu}_{\text{IPWB}}$ in Scenario (a), and $\hat{\mu}_{\text{EL},2}$ outperforms $\hat{\mu}_{\text{IPWA1}}$ and $\hat{\mu}_{\text{IPWA2}}$ and $\hat{\mu}_{\text{CHEN}}$ in Scenario (b). As the propensity scores are estimated under the MAR assumption, which is incorrect in Cases (TT), (FT), (TF) and (FF), $\hat{\mu}_{\text{CHEN}}$, $\hat{\mu}_{\text{IPWA1}}$ and $\hat{\mu}_{\text{IPWA2}}$ all have remarkably larger RMSEs than our EL estimators. Regarding interval estimation, $\hat{\mu}_{\text{EL},1}$, $\hat{\mu}_{\text{IPWB}}$ and $\hat{\mu}_{\text{EL},2}$ produce satisfactory coverage probabilities and average interval lengths. In contrast, $\hat{\mu}_{\text{EL},3}$ usually has larger RMSEs than $\hat{\mu}_{\text{EL},2}$ and the accompanying interval estimator usually has severe undercoverage even if the working models for $P(Y = y|\boldsymbol{X} = \boldsymbol{x})$ and $P(D = 1|\boldsymbol{X} = \boldsymbol{x}, Y = y)$ are correctly specified (Case (TT)). Note that the difference between $\hat{\mu}_{\text{EL},2}$ and $\hat{\mu}_{\text{EL},3}$ is that $\int yf(y|\boldsymbol{X}_i, \hat{\boldsymbol{\xi}})\,\mathrm{d}y$ in $\hat{\mu}_{\text{EL},2}$ is replaced by Y_i in $\hat{\mu}_{\text{EL},3}$ for all units in $\mathcal{S}_A$. As $\int yf(y|\boldsymbol{X}_i, \hat{\boldsymbol{\xi}})\,\mathrm{d}y$ is an estimator of the conditional expectation $\mathbb{E}(Y_i|\boldsymbol{X}_i)$, a possible explanation for the worse performance of $\hat{\mu}_{\text{EL},3}$ is that $\mathbb{V}\text{ar}\{\mathbb{E}(Y_i|\boldsymbol{X}_i)\} \leq \mathbb{V}\text{ar}(Y_i) = \mathbb{V}\text{ar}\{\mathbb{E}(Y_i|\boldsymbol{X}_i)\} + \mathbb{E}\{\mathbb{V}\text{ar}(Y_i|\boldsymbol{X}_i)\}$ and the proposed bootstrap variance estimator may underestimate the variance of $\hat{\mu}_{\text{EL},3}$.

Overall, we recommend $\hat{\mu}_{\text{EL},1}$ under Scenario (a) and $\hat{\mu}_{\text{EL},2}$ under Scenario (b) for estimating the population mean μ_y. For confidence interval construction, we recommend the Wald-type interval estimators associated with $\hat{\mu}_{\text{EL},1}$ and $\hat{\mu}_{\text{EL},2}$ under Scenarios (a) and (b), respectively.

5. Real data analysis

In this section, we analyse a dataset from the second wave of the ESPACOV survey (Estudio Social sobre la Pandemia de COVID-19), conducted by the Institute for Advanced Social Studies at the Spanish National Research Council (IESA-CSIC). The dataset is available at

https://digital.csic.es/handle/10261/233224. Both waves of the ESPACOV survey were conducted online and employed a mixed multi-phase sampling design. A probability sample ($\mathcal{S}_B$) was obtained through SMS invitations sent to a randomly generated list of mobile phone numbers. In January 2021, a total of 66,439 SMS invitations containing a link to the questionnaire were issued for the second edition of the ESPACOV Survey, of which 51.3% were successfully delivered. After eight days of fieldwork, the resulting effective sample size was $n = 973$, corresponding to 2.97% of delivered invitations. A non-probability sample ($\mathcal{S}_A$) was recruited in parallel via targeted advertisements on Facebook, Instagram, and Google. The advertising campaign was designed to oversample sociodemographic groups that were underrepresented in the probability-based sample. Running from January 18 to January 22, the campaign generated 1,054,301 impressions and 7,647 clicks, yielding a total of 671 completed interviews. Additional details on the survey design and implementation can be found in Rinken et al. (2020) and Rueda et al. (2023).

The questionnaires assessed the opinions and attitudes of the Spanish population regarding the COVID-19 crisis, including evaluations of its management and perceived consequences. This study focuses on the self-assessment of mood among Spanish residents aged 18 and older during the pandemic. Mood was originally measured on a five-point scale (very bad or very good) and recoded into a binary variable indicating good mood (1 = good, very good and neither good nor bad; 0 = otherwise). Our analysis incorporates 9 covariates that may affect participants' real-time mood self-assessments; complete variable descriptions are available in Table 5. Responses such as 'not answered' and 'refused to answer' among the covariates were removed. After processing, we consider two scenarios for $\mathcal{S}_B$ according to whether responses are available or not. The final sample consists of 910 observations from the probability sample and 616 observations from the non-probability sample.

In fact, a negative mood may be associated with emotional distress that respondents are reluctant to report. This increases the likelihood of missingness conditional on mood, suggesting that the missing-data mechanism is potentially nonignorable. We consider modelling the inclusion of an observation in the non-probability sample by a logistic model

$$P_q(D_i = 1|\boldsymbol{X} = \boldsymbol{x}_i, Y = y_i) = \frac{\exp(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i + \gamma\, y_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i + \gamma\, y_i)},$$

where D_i is the inclusion indicator, and α, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9)$, and γ are unknown parameters. As the response variable is binary, we assume a logistic regression model $\exp(\boldsymbol{x}_i^\top \boldsymbol{\eta})/\{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\eta})\}$ for $P(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i)$.

We apply the estimation methods considered in our simulation study to analyse this dataset. Table 6 reports point estimates, bootstrap standard errors (computed from $B = 500$ bootstrap replications), and $(1 - \alpha)$ normal-approximation confidence intervals of the form $\hat{\theta} \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta})$, where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ quantile of the standard normal distribution. When responses are available in $\mathcal{S}_B$, the resulting estimates are $\hat{\mu}_{\mathrm{IPWB}} = 0.724$ and $\hat{\mu}_{\mathrm{EL},1} = 0.734$. When responses are unavailable in $\mathcal{S}_B$, the resulting estimates are $\hat{\mu}_{\mathrm{EL},2} = 0.725$, $\hat{\mu}_{\mathrm{EL},3} = 0.740$, $\hat{\mu}_{\mathrm{naive}} = 0.766$, $\hat{\mu}_{\mathrm{IPWA1}} = 3.836$, $\hat{\mu}_{\mathrm{IPWA2}} = 0.999$, and $\hat{\mu}_{\mathrm{DR}} = 3.649$. We see that the three EL estimates are all around 0.74; however, the competitors are either much smaller than 0.74 or much greater than 0.74. Notably, these three estimators $\hat{\mu}_{\mathrm{IPWA1}}$, $\hat{\mu}_{\mathrm{IPWA2}}$, and $\hat{\mu}_{\mathrm{DR}}$ either approach or significantly exceed 1. This is highly inconsistent with the nature of a binary dependent variable (restricted to 0 or 1), as the results violate the fundamental probability constraints. According to our simulation experience, the proposed EL estimators usually produce the most reliable estimates for the finite population of interest.

Table 5. Description of the nine variables in the ESPACOV survey.

Notation	Variable	Level	$\mathcal{S}_A$	$\mathcal{S}_B$	$\mathcal{S}_A \cup \mathcal{S}_B$
X_1	Gender	Male (1)	41%	48%	46%
		Female (2)	59%	52%	54%
X_2	Age	18-29 (1)	3%	18%	12%
		30-44 (2)	15%	34%	26%
		45-64 (3)	38%	41%	40%
		65 or older (4)	44%	7%	22%
	Average Age		58.54	44.12	49.94
X_3	Education level	First degree (1)	19%	20%	20%
		Secondary (2)	27%	18%	21%
		Superiors (3)	54%	62%	59%
X_4	Labour Status	Employed (1)	42%	71%	59%
		Unemployed (2)	6%	9%	8%
		Inactive (3)	52%	21%	33%
X_5	Social class	Low/Low medium (1)	31%	31%	31%
		Others (2)	69%	69%	69%
X_6	Health self-assessment	Bad/Very bad (1)	8%	6%	7%
		Others (2)	92%	94%	93%
X_7	Infect coronavirus	No (1)	91%	87%	89%
		Infected (2)	9%	13%	11%
X_8	Household income level	Increase (1)	4%	4%	4%
		Stay the same (2)	65%	53%	58%
		Decrease (3)	31%	43%	38%
X_9	Close-contact infection	Yes (1)	29%	31%	30%
		No (2)	71%	69%	70%
Y	Mood self-assessment $^{\perp}$	Bad/Very bad (0)	24%		
		Others (1)	76%		

When y is available in $\mathcal{S}_B$ and $\mathcal{S}_A \cup \mathcal{S}_B$, the proportions of responses coded as '0' and '1' are 28% and 72%, 26% and 74%, respectively.

Table 6. Point estimates, bootstrap standard errors, and confidence intervals.

Method	Point Estimate	Bootstrap SE	Confidence Interval
$\hat{\mu}_{\text{IPWB}}$	0.724	0.009	[0.708, 0.741]
$\hat{\mu}_{\text{EL},1}$	0.734	0.017	[0.702, 0.767]
$\hat{\mu}_{\text{EL},2}$	0.725	0.022	[0.682, 0.767]
$\hat{\mu}_{\text{EL},3}$	0.740	0.022	[0.698, 0.783]

The associated bootstrap standard errors were modest, ranging from 0.017 to 0.022. Overall, the strong agreement among the three EL estimators indicates that the proposed procedures deliver stable inference for the target population quantity. We believe that the average Mood self-assessment of Spanish residents aged 18 and older during the pandemic is around 0.74.

6. Discussion

Integrating both a probability sample and a non-probability sample, we have developed three EL estimation methods for the finite population mean, depending on whether responses are available or unavailable in the probability sample. We establish the asymptotic normalities of the proposed EL estimators under certain regularity conditions. According to our simulation study, the proposed EL estimators $\hat{\mu}_{\text{EL},1}$ and $\hat{\mu}_{\text{EL},2}$ generally have more reliable performance than existing competitors depending on whether the responses in the probability sample are available.

One crucial requirement for our method is the availability of the finite population size N. Without additional conditions, our method is not applicable when N is unknown. Even

so, we believe that extending our method to the unknown N case is possible if two or more non-probability samples are available from the same finite population. This setting bears close resemblance to the maximum full likelihood method for capture-recapture data developed by Liu et al. (2017). We leave this problem as an interesting topic for future research.

Author contributions

CRediT: **Yu Liu:** Data curation, Formal analysis, Writing – original draft; **Yukun Liu:** Formal analysis, Methodology, Writing – review & editing; **Jing Qin:** Methodology, Writing – review & editing

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China [grant number JYB2025XDXM904] and the National Natural Science Foundation of China [grant number 12571283].

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143.

Breidt, F. J., & Opsomer, J. D. (2002). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, *28*(4), 1026–1053.

Chauvet, G. (2025). A review of bootstrap methods for survey sampling. *The Survey Statistician*, *91*, 14–25.

Chen, S., & Haziza, D. (2023). General purpose multiply robust data integration procedures for handling nonprobability samples. *Scandinavian Journal of Statistics*, *50*(2), 697–724.

Chen, Y., Li, P., Rao, J., & Wu, C. (2022). Pseudo empirical likelihood inference for nonprobability survey samples. *Canadian Journal of Statistics*, *50*(4), 1166–1185.

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021.

Chlebicki, P., Chrostowski, L., & Beresewicz, M. (2024). Data integration of non-probability and probability samples with predictive mean matching. *arXiv:2403.13750*.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264.

Han, Q., & Wellner, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. *Annals of Statistics*, *49*, 459–485.

Kim, J. K., & Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin*, *75*(1), 8–27.

Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(3), 941–963.

Kim, J. K., Rao, J. N. K., & Wang, Z. (2023). Hypotheses testing from complex survey data using bootstrap weights: A unified approach. *Journal of the American Statistical Association*, *119*(546), 1229–1239.

Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, *37*(3), 319–343.

Liu, Y., & Fan, Y. (2023). Biased-sample empirical likelihood weighting for missing data problems: An alternative to inverse probability weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(1), 67–83.
Liu, Y., Li, P., & Qin, J. (2017). Maximum empirical likelihood estimation for abundance in a closed population from capture-recapture data. *Biometrika*, *104*, 527–543.
Liu, Y., Yuan, M., Li, P., & Wu, C. (2026). Statistical inference with nonignorable non-probability survey samples. *Electronic Journal of Statistics*, *20*(1), 2009–2048.
Liu, Z., Chen, X., Li, R., & Hou, L. (2024). Bayesian inference for nonprobability samples with non-ignorable missingness. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *17*(1), e11667.
Liu, Z., & Valliant, R. (2023). Investigating an alternative for estimation from a nonprobability sample: Matching plus calibration. *Journal of Official Statistics*, *39*(1), 45–78.
Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, *18*, 90–120.
Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.
Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, *83*(1), 242–272.
Rinken, S., Domínguez-Álvarez, J. A., Trujillo, M., Lafuente, R., Sotomayor, R., & Serrano-del Rosal, R. (2020). Combined mobile-phone and social-media sampling for web survey on social effects of COVID-19 in Spain. *Survey Research Methods*, *14*(2), 165–170.
Rivers, D. (2007). Sampling for web surveys. In *Joint statistical meetings* (Vol. 4, pp. 1320).
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.
Rueda, M. D. M., Pasadas-del Amo, S., Rodríguez, B. C., Castro-Martín, L., & Ferri-García, R. (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques: An application to a survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, *65*(2), 2200035.
Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*(1), 105–137.
Wang, L., Graubard, B. I., Katki, H. A., & Li, Y. (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(3), 1293–1311.
Wang, L., Valliant, R., & Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, *40*(24), 5237–5250.
Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, *48*(2), 283–311.
Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice*. Springer.
Yang, S., & Kim, J. K. (2020). Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scandinavian Journal of Statistics*, *47*(3), 839–861.
Yang, S., Kim, J. K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, *47*(1), 29–58.

Appendices

Appendix 1. Definitions of $q_{11}(Z)$, $q_{12}(Z)$ and q_{13}

Define

$$q_{11}(\boldsymbol{Z}) = f_B a_{11}(\boldsymbol{Z}_i) Y_i - H_y \Sigma_a^{-1} \boldsymbol{a}_1(\boldsymbol{Z}_i), \tag{A1}$$

$$q_{12}(\boldsymbol{Z}) = \pi(\boldsymbol{Z}_i) a_{11}(\boldsymbol{Z}_i) Y_i - \pi(\boldsymbol{Z}_i) H_{\pi y} - H_y \Sigma_a^{-1} \boldsymbol{a}_2(\boldsymbol{Z}_i), \tag{A2}$$

$$q_{13} = -H_y \Sigma_a^{-1} e_* / (1 - \eta_0) - H_{\pi y}, \tag{A3}$$

where $e_* = (0, \ldots, 0, 1)^\top$, $H_y = (H_{\dot\pi y}^\top, -H_{\pi y})^\top$,

$$\boldsymbol{a}_1(\boldsymbol{Z}) = \begin{pmatrix} -a_{\dot\pi 1}(\boldsymbol{Z}) + H_{\dot\pi 1} \\ a_{11}(\boldsymbol{Z}) - H_{11} \end{pmatrix} f_B, \quad \boldsymbol{a}_2(\boldsymbol{Z}) = \begin{pmatrix} f_B a_{\dot\pi 1}(\boldsymbol{Z}) \\ \pi(\boldsymbol{Z}) a_{11}(\boldsymbol{Z}) + \frac{\pi(\boldsymbol{Z})}{1-\eta_0} \end{pmatrix},$$

and

$$\Sigma_a = \frac{1}{(f_B + \eta_0)H_{11} - 1} \begin{pmatrix} H_{\dot\pi 1} \\ H_{11} \end{pmatrix} H_{11}^{-1} \begin{pmatrix} H_{\dot\pi 1} \\ H_{11} \end{pmatrix}^\top + \begin{pmatrix} H_{\dot\pi\dot\pi 0} - H_{\dot\pi\dot\pi 1} + H_{\dot\pi 1} H_{11}^{-1} H_{\dot\pi 1}^\top & 0 \\ 0 & (1-\eta_0)^{-1} \end{pmatrix}.$$

Appendix 2. Definition of $q_{21}(X)$, $q_{22}(X)$, $q_{23}(X)$ and $q_{31}(X)$, $q_{32}(X)$, $q_{33}(X)$

Further let $\hat{Y}_i = Y_i I(i \in \mathcal{S}_A) + \tilde{Y}_i I(i \in \mathcal{S}_B)$ with $\tilde{Y}_i = \int y f(y|X_i, \boldsymbol{\zeta}_0)\, \mathrm{d}y$.

Define

$$J_{11} = \mathbb{E}\left[\frac{1}{f_B + \varphi(\boldsymbol{X})}\right], \quad J_{\varphi\dot{\boldsymbol{\theta}}} = \mathbb{E}\left[\frac{\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X})}{f_B + \varphi(\boldsymbol{X})}\right], \quad J_{\varphi\dot{\boldsymbol{\zeta}}} = \mathbb{E}\left[\frac{\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X})}{f_B + \varphi(\boldsymbol{X})}\right],$$

$$J_{\dot{\boldsymbol{\theta}}\dot{\boldsymbol{\theta}}1} = \mathbb{E}\left[\frac{\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X})\nabla_{\boldsymbol{\theta}}{}^\top\varphi(\boldsymbol{X})}{f_B + \varphi(\boldsymbol{X})}\right], \quad J_{\dot{\boldsymbol{\theta}}\dot{\boldsymbol{\zeta}}1} = \mathbb{E}\left[\frac{\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X})\nabla_{\boldsymbol{\zeta}}{}^\top\varphi(\boldsymbol{X})}{\varphi(\boldsymbol{X}) + f_B}\right],$$

$$J_{\dot{\boldsymbol{\zeta}}\dot{\boldsymbol{\zeta}}1} = \mathbb{E}\left[\frac{\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X})\nabla_{\boldsymbol{\zeta}}{}^\top\varphi(\boldsymbol{X})}{f_B + \varphi(\boldsymbol{X})}\right], \quad J_{\boldsymbol{\theta}\boldsymbol{\zeta}^\top} = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\zeta}^\top}\varphi(\boldsymbol{X})\right], \quad J_{\boldsymbol{\zeta}\boldsymbol{\zeta}^\top} = \mathbb{E}\left[\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}^\top}\varphi(\boldsymbol{X})\right],$$

$$J_f = \mathbb{E}[\pi(\boldsymbol{Z})\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}^\top}\log f(Y|\boldsymbol{X}, \boldsymbol{\zeta}_0)], \quad J_{\varphi y} = \frac{1}{1-\eta_0}\mathbb{E}\left\{\frac{1-\varphi(\boldsymbol{X}_i)}{f_B + \varphi(\boldsymbol{X}_i)}\tilde{Y}_i\right\},$$

$$J_{\varphi\dot{\boldsymbol{\theta}}y} = \mathbb{E}\left\{\frac{\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X}_i)}{f_B + \varphi(\boldsymbol{X}_i)}\tilde{Y}_i\right\}, \quad J_{\varphi\dot{\boldsymbol{\zeta}}y} = \mathbb{E}\left\{\frac{\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X}_i)}{f_B + \varphi(\boldsymbol{X}_i)}\tilde{Y}_i\right\}, \quad J_{fy} = \mathbb{E}\{Y\nabla_{\boldsymbol{\zeta}}\log f(Y|X, \boldsymbol{\zeta}_0)\}.$$

$$J_{\varphi\hat{y}} = \frac{1}{1-\eta_0}\mathbb{E}\left\{\frac{1-\varphi(\boldsymbol{X}_i)}{f_B + \varphi(\boldsymbol{X}_i)}\hat{Y}_i\right\}, \quad J_{\varphi\dot{\boldsymbol{\theta}}\hat{y}} = \mathbb{E}\left\{\frac{\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{X}_i)}{f_B + \varphi(\boldsymbol{X}_i)}\hat{Y}_i\right\},$$

$$J_{\varphi\dot{\boldsymbol{\zeta}}\hat{y}} = \mathbb{E}\left\{\frac{\nabla_{\boldsymbol{\zeta}}\varphi(\boldsymbol{X}_i)}{f_B + \varphi(\boldsymbol{X}_i)}\hat{Y}_i\right\}, \quad J_{f\hat{y}} = f_B\mathbb{E}\left\{\frac{Y\nabla_{\boldsymbol{\zeta}}\log f(Y|X, \boldsymbol{\zeta}_0)}{f_B + \varphi(\boldsymbol{X}_i)}\right\}.$$

Define

$$q_{21}(\boldsymbol{X}) = f_B b_{11}(\boldsymbol{X}_i)\tilde{Y}_i - J_y\Sigma_b^{-1}\boldsymbol{b}_1(\boldsymbol{X}_i), \tag{A4}$$

$$q_{22}(\boldsymbol{X}) = b_{11}(\boldsymbol{X}_i)\tilde{Y}_i - J_{\varphi y} - J_y\Sigma_b^{-1}\boldsymbol{b}_2(\boldsymbol{X}_i), \tag{A5}$$

$$q_{23}(\boldsymbol{X}) = f_B b_{11}(\boldsymbol{X}_i)\tilde{Y}_i - J_y\Sigma_b^{-1}\boldsymbol{b}_3(\boldsymbol{X}_i) - Y_i + \eta_0 J_{\varphi y}. \tag{A6}$$

And define

$$q_{31}(\boldsymbol{X}) = f_B b_{11}(\boldsymbol{X}_i)\tilde{Y}_i - J_{\hat{y}}\Sigma_b^{-1}\boldsymbol{b}_1(\boldsymbol{X}_i), \tag{A7}$$

$$q_{32}(\boldsymbol{X}) = b_{11}(\boldsymbol{X}_i)Y_i - J_{\varphi\hat{y}} - J_{\hat{y}}\Sigma_b^{-1}\boldsymbol{b}_2(\boldsymbol{X}_i), \tag{A8}$$

$$q_{33}(\boldsymbol{X}) = f_B b_{11}(\boldsymbol{X}_i)\tilde{Y}_i - J_{\hat{y}}\Sigma_b^{-1}\boldsymbol{b}_3(\boldsymbol{X}_i) - Y_i + \eta_0 J_{\varphi\hat{y}}, \tag{A9}$$

where $J_y = (J_{\varphi\dot{\boldsymbol{\theta}}y}^\top, (J_{\varphi\dot{\boldsymbol{\zeta}}y} - J_{fy})^\top, J_{\varphi y})^\top$, $J_{\hat{y}} = (J_{\varphi\dot{\boldsymbol{\theta}}\hat{y}}^\top, (J_{\varphi\dot{\boldsymbol{\zeta}}\hat{y}} - J_{f\hat{y}})^\top, J_{\varphi\hat{y}})^\top$,

$$\boldsymbol{b}_1(\boldsymbol{X}_i) = \begin{pmatrix} -b_{\varphi\dot{\boldsymbol{\theta}}}(\boldsymbol{X}_i) + J_{\varphi\dot{\boldsymbol{\theta}}} \\ -b_{\varphi\dot{\boldsymbol{\zeta}}}(\boldsymbol{X}_i) + J_{\varphi\dot{\boldsymbol{\zeta}}} \\ b_{11}(\boldsymbol{X}_i) - J_{11} \end{pmatrix} f_B, \quad \boldsymbol{b}_2(\boldsymbol{X}_i) = \begin{pmatrix} \frac{\dot\pi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{Z}_i)} - b_{\varphi\dot{\boldsymbol{\theta}}}(\boldsymbol{X}_i) \\ \frac{\nabla_{\boldsymbol{\zeta}} f(Y_i|\boldsymbol{X}_i, \boldsymbol{\zeta}_0)}{f(Y_i|\boldsymbol{X}_i, \boldsymbol{\zeta}_0)} - b_{\varphi\dot{\boldsymbol{\zeta}}}(\boldsymbol{X}_i) \\ b_{11}(\boldsymbol{X}_i) + \frac{1}{1-\eta_0} \end{pmatrix},$$

and

$$\boldsymbol{b}_3(\boldsymbol{X}_i) = \begin{pmatrix} -f_B b_{\varphi\dot{\boldsymbol{\theta}}}(\boldsymbol{X}_i) \\ -f_B b_{\varphi\dot{\boldsymbol{\zeta}}}(\boldsymbol{X}_i) \\ f_B b_{11}(\boldsymbol{X}_i) - \frac{1}{1-\eta_0} \end{pmatrix}.$$

And

$$\Sigma_b = \Sigma_{b1} + \Sigma_{b2} + \Sigma_{b3},$$

where

$$\Sigma_{b1} = \frac{1}{(f_B + \eta_0)J_{11} - 1} \begin{pmatrix} J_{\varphi\dot{\boldsymbol{\theta}}} \\ J_{\varphi\dot{\boldsymbol{\zeta}}} \\ J_{11} \end{pmatrix} J_{11}^{-1} \begin{pmatrix} J_{\varphi\dot{\boldsymbol{\theta}}} \\ J_{\varphi\dot{\boldsymbol{\zeta}}} \\ J_{11} \end{pmatrix}^{\top},$$

$$\Sigma_{b2} = \begin{pmatrix} J_{\varphi\dot{\boldsymbol{\theta}}} J_{11}^{-1} J_{\varphi\dot{\boldsymbol{\theta}}}^{\top} - J_{\dot{\boldsymbol{\theta}}\dot{\boldsymbol{\theta}}1} & J_{\varphi\dot{\boldsymbol{\theta}}} J_{11}^{-1} J_{\varphi\dot{\boldsymbol{\zeta}}}^{\top} - J_{\dot{\boldsymbol{\theta}}\dot{\boldsymbol{\zeta}}1} & 0 \\ J_{\varphi\dot{\boldsymbol{\zeta}}} J_{11}^{-1} J_{\varphi\dot{\boldsymbol{\theta}}}^{\top} - J_{\dot{\boldsymbol{\theta}}\dot{\boldsymbol{\zeta}}1} & J_{\varphi\dot{\boldsymbol{\zeta}}} J_{11}^{-1} J_{\varphi\dot{\boldsymbol{\zeta}}}^{\top} - J_{\dot{\boldsymbol{\zeta}}\dot{\boldsymbol{\zeta}}1} & 0 \\ 0 & 0 & (1-\eta_0)^{-1} \end{pmatrix},$$

$$\Sigma_{b3} = \begin{pmatrix} H_{\dot{\pi}\dot{\pi}0} & J_{\boldsymbol{\theta}\boldsymbol{\zeta}} & 0 \\ J_{\boldsymbol{\theta}\boldsymbol{\zeta}} & J_{\boldsymbol{\zeta}\boldsymbol{\zeta}}{}^{\top} - J_f & 0 \\ 0 & 0 & \end{pmatrix}.$$