Statistica Sinica Preprint No: SS-2022-0295

Title	Improved Model-assisted Estimation via Probability Thresholding
Manuscript ID	SS-2022-0295
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0295
Complete List of Authors	Xianpeng Zong, Geoffrey K. F. Tso and Guohua Zou
Corresponding Authors	Guohua Zou
E-mails	ghzou@amss.ac.cn

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Statistica Sinica

IMPROVED MODEL-ASSISTED ESTIMATION VIA PROBABILITY THRESHOLDING

Xianpeng Zong, Geoffrey K. F. Tso and Guohua Zou

Beijing University of Technology, City University of Hong Kong and Capital Normal University

Abstract: In survey sampling, model-assisted approach is often used to improve the precision of survey estimators when auxiliary information is available. Generally, the model-assisted estimators are nonlinear functions of some classical Horvitz-Thompson estimators constructed via inverse probability weighting, which are seriously affected by the heterogeneous inclusion probabilities. In this paper, we improve the classical model-assisted estimation via probability thresholding, and propose the improved linear and nonparametric model-assisted estimators for finite populations. The proposed estimators are shown to be asymptotically design unbiased and design consistent. The corresponding design mean squared errors and their estimators are also derived. We theoretically prove that the new model-assisted estimators are asymptotically not worse than the commonly used model-assisted estimators. Two simulation examples and an empirical application indicate good finite sample performance of the proposed estimators.

Key words and phrases: Horvitz-Thompson estimator, Model-assisted estimation, Probability thresholding, Superpopulation model, Survey sampling.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

1. Introduction

In survey sampling, three kinds of frameworks are used to make statistical estimation and inference: design-based approach, model-based approach, and model-assisted approach. Generally, the last one is better than the first two when auxiliary information is available. Model-assisted method provides a convenient framework that uses a superpopulation model to describe the relationship between the variable of interest and the auxiliary variables. A model-assisted estimator improves the precision of the traditional survey estimators when the model is correct, and maintains desirable properties such as asymptotic design unbiasedness and design consistency when the model is incorrect. In the past decades, various superpopulations are considered. For instance, Särndal et al. (1992) detailed linear model-assisted estimation that assumes the superpopulation models are ratio or linear models. Based on the hypothesis of nonparametric model, Breidt and Opsomer (2000) used local polynomial regression to construct a nonparametric model-assisted estimator. Breidt et al. (2007) proposed a semiparametric model-assisted estimator. Moreover, Wang (2009) and Wang and Wang (2011) discussed single-index model-assisted estimation and nonparametric additive model-assisted estimation, respectively.

Typically, the above model-assisted estimators can be written as the nonlinear functions of some classical Horvitz-Thompson (HT) estimators proposed by Horvitz and Thompson (1952). The HT estimator is a design unbiased estimator constructed via inverse probability weighting. It has

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

been applied to many other fields such as treatment effect (Rosenbaum, 2002), functional data analysis (Cardot and Josserand, 2011) and optimal subsampling for big data (Wang et al., 2018). However, when the first-order inclusion probabilities of some units are relatively small, the variance of the HT estimator becomes large due to inverse probability weighting. In order to solve this problem, a simple and common method is to modify the large design weights (i.e., inverses of small inclusion probabilities) by trimming. Weight trimming, often with a threshold, can be explained as a shrinkage strategy which generally improves the estimation accuracy. Benrud (1978) trimmed all weights below some factor c of the square root of the mean of the squared weights, which is called the "NAEP procedure" because of its use in the National Assessment of Educational Progress. Potter (1988, 1990) assumed the distribution of weights and then truncated some "unlikely" weights by the estimated distribution. Further, Kokic and Bell (1994) and Rivest and Hurtubise (1995) selected the threshold by minimizing the mean squared error of the winsorized estimator that trims large weighted y-values. Beaumont et al. (2013) constructed a robust version of the HT estimator based on the concept of conditional bias. This method can be extended to the generalized regression estimator. For the winsorized estimator, Favre-Martinoz et al. (2015) suggested determining the threshold that minimizes the absolute estimated conditional bias. Chen et al. (2017) reviewed the weight trimming methods and commented that the thresholds obtained by the optimization methods are usually y-specific and the resulting weights

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

are not multipurpose. In addition, the calibration method can also improve the estimation accuracy by adjusting the weights. See, for example, Deville and Särndal (1992), Wu and Sitter (2001) and Montanari and Ranalli (2005). Recently, Zong et al. (2019) proposed a method of trimming the small inclusion probabilities, and determined a probability threshold by comparing the MSEs of the traditional estimators and the resulting estimators. An improved Horvitz-Thompson (IHT) estimator can be constructed based on the modified inclusion probabilities. It should be emphasized that the probability threshold for the IHT estimator does not need to assume the weights distribution and the resulting weights are not y-specific. In exploring theoretical properties of the IHT estimator, Zong et al. (2019) assumed that the first-order inclusion probabilities have a lower bound away from zero, which will be removed in this paper.

The main purpose of this paper is to improve the classical model-assisted estimation via using the modified first-order inclusion probabilities from Zong et al. (2019). The improved linear and nonparametric model-assisted estimators are proposed. Compared to Zong et al. (2019) who improved only the design-based estimator, our proposed estimators are generally more efficient and robust than the traditional design-based estimators and model-assisted estimators. Like the existing model-assisted estimators, we establish the design properties of the improved model-assisted estimators including calibration, design consistency and asymptotic design unbiasedness. The design mean squared errors and their estimators

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

for all the proposed estimators are also obtained. In addition, we theoretically compare the design mean squared errors of traditional model-assisted estimators and improved model-assisted estimators.

The remainder of this paper is organized as follows. Section 2 briefly introduces the IHT estimator, and the proofs of its theoretical properties are provided without the constraint that the first-order inclusion probabilities have a lower bound away from zero. Section 3 proposes the improved linear model-assisted estimator. Its design properties are established in Section 4. Section 5 develops the improved nonparametric model-assisted estimator based on local polynomial regression. In Section 6, we derive the design properties of such a nonparametric model-assisted estimator. Section 7 provides numerical examples from two simulations and a real data analysis. Section 8 concludes. Proofs of theoretical results are contained in the supplementary material.

2. Improved Horvitz-Thompson Estimator

Consider a finite population $U = \{1, \cdots, k, \cdots, N\}$. For each unit k, the value of target characteristic Y is denoted as y_k. A sample s of size n is randomly drawn from the population U according to a sampling design $p(\cdot)$, where $p(s)$ is the probability of drawing the sample s. We implement the unequal probability sampling without replacement. Denote $\pi_k = \Pr\{k \in s\} = \sum_{\{s:\ s \ni k\}} p(s)$ and $\pi_{kl} = \Pr\{k, l \in s\} = \sum_{\{s:\ s \ni k,l\}} p(s)$ for all $k, l \in U$ as the first-order inclusion probabilities and the second-order inclusion

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

probabilities, respectively. Our aim is to estimate the population mean, $\bar{t}_y = N^{-1}\sum_U y_k$.

Let $I_k = 1$ or 0, if the k^{th} unit is drawn or not, $k = 1, \ldots, N$. Note that $\mathrm{E}_p(I_k) = \pi_k$, where $\mathrm{E}_p(\cdot)$ denotes expectation with respect to the sampling design. A well-known design unbiased estimator of $\bar{t}_y$ (i.e. $\mathrm{E}_p(\hat{\bar{t}}_y) = \bar{t}_y$) is the HT estimator,

$$\hat{\bar{t}}_{\text{HT}} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} \tag{2.1}$$

with the variance given by

$$\mathrm{V}_p(\hat{\bar{t}}_{\text{HT}}) = \frac{1}{N^2}\sum_{k,l\in U}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}y_k y_l. \tag{2.2}$$

Note that when the first-order inclusion probabilities of some units are relatively small, the variance of the HT estimator will be large due to inverse probability weighting. Hence, Zong et al. (2019) used a hard-threshold method for the first-order inclusion probabilities to construct an IHT estimator.

Definition 1. Let $\pi_{(1)} \leq \pi_{(2)} \leq \cdots \leq \pi_{(N)}$ be the ordered values of the first-order inclusion probabilities $\{\pi_1, \pi_2, \cdots, \pi_N\}$. Assume that there exists an integer $K \geq 2$ such that $\pi_{(K)} \leq (K+1)^{-1}$. Define the modified first-order inclusion probabilities as follows

$$\pi_k^* = \begin{cases} \pi_k & \pi_k > \pi_{(K)}, \\ \pi_{(K)} & \pi_k \leq \pi_{(K)}. \end{cases} \qquad 1 \leq k \leq N,$$

From this definition, the finite population is divided into two parts: $U_1 = \{k : \pi_k > \pi_{(K)}\}$ with size $N - K$, and $U_2 = \{k : \pi_k \leq \pi_{(K)}\}$ with size

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

K. Obviously, the efficiency of the IHT estimator relies on the choice of K, which provides a control of the variance-bias tradeoff. Zong et al. (2019) chose $K^* = \max\{k \in U : \pi_{(k)} \leq (k+1)^{-1}\}$ as the threshold and gave an algorithm to find it.

Using the modified first-order inclusion probabilities $\{\pi_k^*\}_{k=1}^N$, the IHT estimator can be constructed by

$$\hat{\bar{t}}_{\text{IHT}} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k^*}. \tag{2.3}$$

Zong et al. (2019) showed that such an estimator is asymptotically design unbiased and design consistent, and compared it with the HT estimator in terms of design mean squared errors. However, they assumed that the first-order inclusion probabilities have a lower bound away from zero, which holds in many practical surveys but deviates from the purpose of improvement. Here, we break this constraint and establish the same theoretical properties. To this end, we need the following regularity conditions.

Condition 1. $\max_{i \in U} |y_i| \leq c$ with c being a positive constant.

Condition 2. $\min\limits_{i \in U} \pi_i \geq \lambda_N > 0$, $\min\limits_{i,j \in U} \pi_{ij} \geq \lambda_N^* > 0$, and

$$\limsup_{n \to \infty} n \max_{i \neq j \in U} \mid \pi_{ij} - \pi_i \pi_j \mid < \infty,$$

where λ_N and λ_N^* can depend on N.

Condition 1 states that the study variable is bounded. This is a reasonable assumption in many situations, but not when the variables are heavily skewed. In Condition 2, λ_N and λ_N^* are allowed to tend to zero as

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

$N \to \infty$. The last part of Condition 2 is a commonly used assumption, which captures sampling dependence between pairs of units (Robinson and Särndal, 1983). This assumption holds, say, for simple random sampling without replacement, Poisson sampling and rejective sampling (Hájek, 1964; Boistard et al., 2012), but not for the designs with strong dependencies, such as multistage sampling and systematic sampling (Delevoye and Sävje, 2020).

Theorem 1. *If Conditions 1 and 2 are satisfied and* $(n\lambda_N^2)^{-1} = o(1)$*, then the IHT estimator, defined in (2.3), is asymptotically design unbiased and design consistent.*

It is clear that Theorem 1 allows λ_N to tend to 0 at a rate slower than $n^{-1/2}$. Additionally, from the proof of Theorem 1, we see that this result is also right for the IHT estimator with the correction ratio satisfying $K/N = O\left(n^{-1/2}\lambda_N^{-1}\right)$.

The following theorem compares the design MSEs of the two estimators $\hat{\hat{t}}_{\mathrm{HT}}$ and $\hat{\hat{t}}_{\mathrm{IHT}}$.

Theorem 2. *If Conditions 1 and 2 are satisfied and* $(N\lambda_N^3)^{-1} = o(1)$*, then*

$$MSE_p(\hat{\hat{t}}_{IHT}) \leq MSE_p(\hat{\hat{t}}_{HT}) + o\left(n^{-1}\right).$$

Especially, for Poisson sampling, we have

$$MSE_p(\hat{\hat{t}}_{IHT}) \leq MSE_p(\hat{\hat{t}}_{HT}),$$

where the strict inequality holds if and only if there exist $k \neq l \in U_2$ *such that* $(\pi_k - \pi_{(K)})y_k \neq (\pi_l - \pi_{(K)})y_l$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Theorem 2 shows that the improved estimator is asymptotically not worse than the classic estimator. However, for Poisson sampling, the IHT estimator outperforms the HT estimator.

3. Improved Linear Model-Assisted Estimation

In this section, we use the modified first-order inclusion probabilities to construct an improved linear model-assisted estimator. Assume that there are J auxiliary variables, denoted by $z_1, \ldots, z_J$. The value of the j^{th} auxiliary variable for the k^{th} population unit is written as z_{jk}. Define $\boldsymbol{z}_k = (z_{1k}, \ldots, z_{Jk})'$. As before, the study variable Y takes the value y_k for the k^{th} unit. Our target is to estimate the population mean, $\bar{t}_y = N^{-1} \sum_U y_k$, assuming that we have observed $(y_k, \boldsymbol{z}_k)$ for $k \in s$, and $\bar{t}_{\boldsymbol{z}}$, the population mean of auxiliary vector, is also known.

We suppose that $\{(y_k, \boldsymbol{z}_k), k \in U\}$ are from the following superpopulation model ξ,

$$y_k = \boldsymbol{z}_k' \boldsymbol{\beta} + \epsilon_k, \quad k \in U, \tag{3.1}$$

where ϵ_k $(k \in U)$ are independent random variables with $E_\xi(\epsilon_k) = 0$ and $V_\xi(\epsilon_k) = \sigma^2$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ is the regression coefficients vector. Note that ξ introduces a new type of randomness, that is, E_ξ and V_ξ denote expectation and variance with respect to the model ξ, respectively.

Let $\boldsymbol{Z} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_N})$, $\boldsymbol{y} = (y_1, \ldots, y_N)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)'$. In matrix notation, (3.1) may be written as $\boldsymbol{y} = \boldsymbol{Z}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$, so the least-squares estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{B} = (\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{y} = \left(\frac{1}{N}\sum\nolimits_U \boldsymbol{z}_k\boldsymbol{z}_k'\right)^{-1}\left(\frac{1}{N}\sum\nolimits_U \boldsymbol{z}_k y_k\right) \triangleq \boldsymbol{T}^{-1}\boldsymbol{t}.$$

Note that $\boldsymbol{B}$ cannot be calculated in the present context, because only the y_k in s are known. A design-based estimator of $\boldsymbol{B}$ is given by

$$\hat{\boldsymbol{B}} = [\hat{B}_1, \ldots, \hat{B}_J]' = \left(\frac{1}{N}\sum\nolimits_s \frac{\boldsymbol{z}_k\boldsymbol{z}_k'}{\pi_k}\right)^{-1}\left(\frac{1}{N}\sum\nolimits_s \frac{\boldsymbol{z}_k y_k}{\pi_k}\right) \triangleq \hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{t}}. \tag{3.2}$$

Then the linear model-assisted estimator for the population mean is written as

$$\hat{\bar{t}}_{yr} = \hat{\bar{t}}_{y\pi} + \sum_{j=1}^{J}\hat{B}_j(\bar{t}_{z_j} - \hat{\bar{t}}_{z_j\pi}), \tag{3.3}$$

where $\bar{t}_{z_j}$ is the population mean of the j^{th} auxiliary variable z_j, and $\hat{\bar{t}}_{y\pi}$ and $\hat{\bar{t}}_{z_j\pi}$ are the HT estimators of $\bar{t}_y$ and $\bar{t}_{z_j}$, respectively, i.e.,

$$\hat{\bar{t}}_{y\pi} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}; \quad \hat{\bar{t}}_{z_j\pi} = \frac{1}{N}\sum_{k\in s}\frac{z_{jk}}{\pi_k}.$$

Define the J-vector $\bar{\boldsymbol{t}}_z = (\bar{t}_{z_1}, \ldots, \bar{t}_{z_J})'$ and its HT estimator as $\hat{\bar{\boldsymbol{t}}}_{z\pi} = (\hat{\bar{t}}_{z_1\pi}, \ldots, \hat{\bar{t}}_{z_J\pi})'$. With these notations we can write

$$\hat{\bar{t}}_{yr} = \hat{\bar{t}}_{y\pi} + (\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}_{z\pi})'\hat{\boldsymbol{B}} \triangleq f(\hat{\bar{t}}_{y\pi}, \hat{\bar{\boldsymbol{t}}}_{z\pi}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{t}}). \tag{3.4}$$

Further, the approximate design variance of the linear model-assisted estimator is

$$\text{AV}_p(\hat{\bar{t}}_{yr}) = \frac{1}{N^2}\sum_{k,l\in U}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}(y_k - \boldsymbol{z}_k'\boldsymbol{B})(y_l - \boldsymbol{z}_l'\boldsymbol{B})$$

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

[Särndal et al. (1992), page 235]. Note that the linear model-assisted estimator is a nonlinear function of some classical HT estimators, and when the first-order inclusion probabilities of some units are relatively small, the approximate design variance $\mathrm{AV}_p(\hat{\bar{t}}_{yr})$ becomes large due to inverse probability weighting.

To overcome this shortcoming, we apply the modified first-order inclusion probabilities $\{\pi_k^*\}_{k=1}^N$ defined by Definition 1 to improve the linear model-assisted estimator. Accordingly, an improved design-based estimator of $\boldsymbol{B}$ is given by

$$\hat{\boldsymbol{B}}^* = [\hat{B}_1^*, \ldots, \hat{B}_J^*]' = \left(\frac{1}{N}\sum\nolimits_s \frac{\boldsymbol{z}_k \boldsymbol{z}_k'}{\pi_k^*}\right)^{-1} \left(\frac{1}{N}\sum\nolimits_s \frac{\boldsymbol{z}_k y_k}{\pi_k^*}\right) \triangleq (\hat{\boldsymbol{T}}^*)^{-1}\hat{\boldsymbol{t}}^*, \quad (3.5)$$

and then the improved linear model-assisted (ILMA) estimator for the population mean is written as

$$\hat{\bar{t}}^*_{yr} = \hat{\bar{t}}^*_{y\pi} + \sum_{j=1}^{J} \hat{B}_j^*(\bar{t}_{z_j} - \hat{\bar{t}}^*_{z_j\pi}), \quad (3.6)$$

where $\hat{\bar{t}}^*_{y\pi}$ and $\hat{\bar{t}}^*_{z_j\pi}$ are the IHT estimators of $\bar{t}_y$ and $\bar{t}_{z_j}$, respectively, i.e.,

$$\hat{\bar{t}}^*_{y\pi} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k^*}; \quad \hat{\bar{t}}^*_{z_j\pi} = \frac{1}{N}\sum_{k\in s}\frac{z_{jk}}{\pi_k^*}.$$

Denote the IHT estimator of $\bar{\boldsymbol{t}}_z$ as $\hat{\bar{\boldsymbol{t}}}^*_{z\pi} = (\hat{\bar{t}}^*_{z_1\pi}, \ldots, \hat{\bar{t}}^*_{z_J\pi})'$, then we have

$$\hat{\bar{t}}^*_{yr} = \hat{\bar{t}}^*_{y\pi} + (\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}^*_{z\pi})'\hat{\boldsymbol{B}}^* \triangleq f(\hat{\bar{t}}^*_{y\pi}, \hat{\bar{\boldsymbol{t}}}^*_{z\pi}, \hat{\boldsymbol{T}}^*, \hat{\boldsymbol{t}}^*). \quad (3.7)$$

In the next section, we will discuss some properties of the improved linear model-assisted estimator, and theoretically show its effectiveness compared to the traditional linear model-assisted estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

4. Properties of ILMA Estimation

In this section, we investigate the properties of the improved estimators $\hat{\boldsymbol{B}}^*$ and $\hat{\hat{t}}^*_{y\pi}$. To this end, we make the following assumptions.

Condition 3. $\max_{j=1,\ldots,J;k\in U} |z_{jk}| < c_1$ and $c_2 \leq \lambda_{\min}(\boldsymbol{T}) \leq \lambda_{\max}(\boldsymbol{T}) \leq c_3$, where c_1, c_2 and c_3 are some positive constants.

Condition 4. Let $\pi_{ijk} = \Pr\{i, j, k \in s\} = \sum_{\{s:\ s\ni i,j,k\}} p(s)$ for all $i, j, k \in U$ be the third-order inclusion probabilities, and denote $D_{t,N}$ as the set of all distinct t-tuples $(i_1, i_2, \ldots, i_t)$ from U. Then

$$\limsup_{n\to\infty} n \max_{(i,j,k)\in D_{3,N}} |\pi_{ijk} - \pi_{ij}\pi_k| < \infty,$$

$$\limsup_{n\to\infty} n^2 \max_{(i,j,k,l)\in D_{4,N}} |\mathrm{E}_p\{(I_i - \pi_i)(I_j - \pi_j)(I_k - \pi_k)(I_l - \pi_l)\}| < \infty,$$

and for some $\alpha \in (0,\, 1)$,

$$\limsup_{n\to\infty} n^\alpha \max_{(i,j,k,l)\in D_{4,N}} |\mathrm{E}_p\{(I_iI_j - \pi_{ij})(I_kI_l - \pi_{kl})\}| < \infty.$$

Similar to Conditions A2 and A8 in Breidt et al. (2007), Condition 3 is the common assumptions on the covariates and the minimum and maximum characteristic values of $\boldsymbol{T}$. Condition 4 extends Condition 2, which is a regular assumption for the higher-order inclusion probabilities. As shown in Boistard et al. (2012) and Breidt and Opsomer (2000), this condition holds for simple random sampling without replacement, Poisson sampling and rejective sampling, but not for the designs with nontrivial clustering. Conditions 2 and 4 have been used to prove the consistency and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

the asymptotic normality of complex estimators in many literatures. See, for example, Breidt et al. (2007), Wang (2009), Cardot et al. (2010), Wang and Wang (2011) and Zong et al. (2019).

Theorem 3. *If Conditions 1 - 4 are satisfied and* $(n\lambda_N^2)^{-1} = o(1)$*, then the improved estimator* $\hat{\boldsymbol{B}}^*$*, defined in (3.5), is asymptotically design unbiased and design consistent. Moreover, the design mean squared error matrix of* $\hat{\boldsymbol{B}}^*$ *is given by*

$$MSE_p(\hat{\boldsymbol{B}}^*) = E_p\left\{(\hat{\boldsymbol{B}}^* - \boldsymbol{B})(\hat{\boldsymbol{B}}^* - \boldsymbol{B})'\right\} = \boldsymbol{T}^{-1}\boldsymbol{V}^*\boldsymbol{T}^{-1} + O\left(n^{-3/2}\lambda_N^{-3}\right)$$

with the rate holding component-wise, and $\boldsymbol{V}^*$ *being a symmetric* $J \times J$ *matrix with elements*

$$v_{jj'}^* = \frac{1}{N^2}\sum_{k,l\in U}\frac{\Delta_{kl}^*}{\pi_k^*\pi_l^*}\left(z_{jk}E_k\right)\left(z_{j'l}E_l\right),$$

where $E_k = y_k - \boldsymbol{z}_k'\boldsymbol{B}$ *is the population fit residual and* $\Delta_{kl}^* = \pi_{kl} - \pi_k\pi_l^* - \pi_k^*\pi_l + \pi_k^*\pi_l^*$.

Theorem 3 establishes the design properties of the improved regression coefficients estimator $\hat{\boldsymbol{B}}^*$. The design mean squared error matrix of $\hat{\boldsymbol{B}}^*$ is derived, and its second term is negligible if $(n\lambda_N^6)^{-1} = o(1)$. This theorem is similar to Result 5.10.1 of Särndal et al. (1992) who discussed the design properties of $\hat{\boldsymbol{B}}$.

From Theorem 3, the estimator of the design mean squared error matrix of $\hat{\boldsymbol{B}}^*$ can be constructed as

$$\widehat{\mathrm{MSE}}_p(\hat{\boldsymbol{B}}^*) = \hat{\boldsymbol{T}}^{*-1}\hat{\boldsymbol{V}}^*\hat{\boldsymbol{T}}^{*-1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

with $\hat{\boldsymbol{V}}^*$ being a symmetric $J \times J$ matrix with elements

$$\hat{v}^*_{jj'} = \frac{1}{N^2} \sum_{k,l \in s} \frac{\check{\Delta}^*_{kl}}{\pi^*_k \pi^*_l} \left(z_{jk} e^*_k\right) \left(z_{j'l} e^*_l\right),$$

where $e^*_k = y_k - \boldsymbol{z}'_k \hat{\boldsymbol{B}}^*$ is the sample fit residual and $\check{\Delta}^*_{kl} = \Delta^*_{kl}/\pi_{kl}$.

Theorem 4. *If Conditions 1 - 4 hold and* $\left(n\lambda_N^3\right)^{-1} = o(1)$, *then*

$$\lim_{n\to\infty} n^{1+\kappa} \lambda_N^2 \lambda_N^* E_p \left|\hat{v}^*_{jj'} - v^*_{jj'}\right| = 0,$$

where $\kappa = \min\left\{\frac{\alpha}{2}, \frac{1}{4}\right\}$.

From Theorem 4, it is found that $\hat{v}^*_{jj'}$ is an asymptotically design unbiased and design consistent estimator for $v^*_{jj'}$ as long as $\left(n^{\kappa}\lambda_N^2\lambda_N^*\right)^{-1} = o(1)$. On the other hand, by Theorem 1, the improved estimator $\hat{\boldsymbol{T}}^*$ whose elements are some IHT estimators is a design consistent estimator of $\boldsymbol{T}$. Thus $\widehat{\mathrm{MSE}}_p(\hat{\boldsymbol{B}}^*)$ is consistent for estimating $\mathrm{MSE}_p(\hat{\boldsymbol{B}}^*)$ if $\left(n\lambda_N^6\right)^{-1} = o(1)$ and $\left(n^{\kappa}\lambda_N^2\lambda_N^*\right)^{-1} = o(1)$.

The following theorem theoretically compares the efficiency of the two estimators $\hat{\boldsymbol{B}}^*$ and $\hat{\boldsymbol{B}}$.

Theorem 5. *If Conditions 1 - 4 hold and* $\left(n\lambda_N^6\right)^{-1} = o(1)$, *then*

$$tr\left\{MSE_p(\hat{\boldsymbol{B}}^*)\right\} \leq tr\left\{MSE_p(\hat{\boldsymbol{B}})\right\} + o\left(n^{-1}\right).$$

Now we turn to study the design properties of the improved linear

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

model-assisted estimator $\hat{\bar{t}}^*_{yr}$. Note from (3.7) that

$$\begin{aligned}\hat{\bar{t}}^*_{yr} &= \hat{\bar{t}}^*_{y\pi} + (\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}^*_{z\pi})'\hat{\boldsymbol{B}}^* \\ &= \frac{1}{N}\sum_{k\in s}\left\{\frac{1}{\pi^*_k} + (\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}^*_{z\pi})'(\hat{\boldsymbol{T}}^*)^{-1}\frac{\boldsymbol{z}_k}{\pi^*_k}\right\} y_k \\ &= \frac{1}{N}\sum_{k\in s}\omega^*_{ks}y_k. \qquad (4.1)\end{aligned}$$

Thus $\hat{\bar{t}}^*_{yr}$ is a linear combination of the sample y_k's with the weights only depending on the auxiliary information. Further, when the weights are applied to the auxiliary information $\boldsymbol{z}_k$, we have

$$\begin{aligned}\sum_s \omega^*_{ks}\boldsymbol{z}'_k &= \sum_{k\in s}\left\{\frac{1}{\pi^*_k} + (\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}^*_{z\pi})'(\hat{\boldsymbol{T}}^*)^{-1}\frac{\boldsymbol{z}_k}{\pi^*_k}\right\}\boldsymbol{z}'_k \\ &= \sum_{k\in s}\frac{\boldsymbol{z}'_k}{\pi^*_k} + N(\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}^*_{z\pi})'(\hat{\boldsymbol{T}}^*)^{-1}\hat{\boldsymbol{T}}^* \\ &= \sum_{k\in U}\boldsymbol{z}'_k. \qquad (4.2)\end{aligned}$$

(4.1) and (4.2) reveal the weighting and calibration properties of $\hat{\bar{t}}^*_{yr}$ respectively, which are highly desirable properties in survey sampling.

Theorem 6. *If Conditions 1 - 4 are satisfied and $(n\lambda_N^4)^{-1} = o(1)$, then the improved estimator $\hat{\bar{t}}^*_{yr}$, defined in (3.6), is asymptotically design unbiased and design consistent. The design mean squared error of $\hat{\bar{t}}^*_{yr}$ is*

$$MSE_p\left(\hat{\bar{t}}^*_{yr}\right) = E_p\left(\hat{\bar{t}}^*_{yr} - \bar{t}_y\right)^2 = \frac{1}{N^2}\sum_{k,l\in U}\frac{\Delta^*_{kl}}{\pi^*_k\pi^*_l}E_kE_l + O\left(n^{-3/2}\lambda_N^{-4}\right).$$

Moreover, if $(n\lambda_N^8)^{-1} = o(1)$ and $(n^\kappa\lambda_N^2\lambda_N^*)^{-1} = o(1)$, then an asymptotically design unbiased and design consistent estimator of the design mean

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

squared error is given by

$$\widehat{AMSE}_p\left(\hat{\check{t}}^*_{yr}\right) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\check{\Delta}^*_{kl}}{\pi^*_k\pi^*_l}e^*_k e^*_l.$$

It is clear that Theorem 6 is similar to Result 6.6.1 of Särndal et al. (1992) who discussed the design properties of $\hat{\check{t}}_{yr}$.

The following theorem theoretically compares the efficiency of the two estimators $\hat{\check{t}}^*_{yr}$ and $\hat{\check{t}}_{yr}$.

Theorem 7. *If Conditions 1 - 4 hold and* $\left(n\lambda_N^8\right)^{-1} = o(1)$, *then*

$$MSE_p(\hat{\check{t}}^*_{yr}) \leq MSE_p(\hat{\check{t}}_{yr}) + o\left(n^{-1}\right).$$

Theorem 7 implies that the MSE$_p$ of the improved estimator $\hat{\check{t}}^*_{yr}$ is asymptotically not larger than that of the traditional estimator $\hat{\check{t}}_{yr}$. It is worth pointing out that, when the linear superpopulation model has heteroskedastic errors, i.e., $V_\xi(\epsilon_k) = \sigma^2_k$, our modification idea is equally applicable to the weighted least-squares model-assisted estimator given by the equation (6.4.13) of Särndal et al. (1992).

5. Improved Nonparametric Model-Assisted Estimation

When the variable of interest has a complex non-linear relationship with an auxiliary variable, Breidt and Opsomer (2000) proposed a nonparametric model-assisted estimator based on local polynomial regression. In this section, we use the hard-threshold method for the first-order inclusion probabilities to improve their nonparametric model-assisted estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Assume that there is only one auxiliary variable x, and the value of the auxiliary variable for the i^{th} population unit is denoted as x_i. Our target remains to estimate the population mean, $\bar{t}_y = N^{-1}\sum_U y_i$, assuming that we have observed (y_i, x_i) for $i \in s$, and x_i with $i \in U - s$ are known.

We suppose that $\{(y_i, x_i), i \in U\}$ are from the following superpopulation model $\tilde{\xi}$,

$$y_i = m(x_i) + \tilde{\epsilon}_i, \tag{5.1}$$

where $m(x)$ is a smooth function of x, $\tilde{\epsilon}_i$ $(i \in U)$ are independent random variables with $E_{\tilde{\xi}}(\tilde{\epsilon}_i) = 0$ and $V_{\tilde{\xi}}(\tilde{\epsilon}_i) = v(x_i)$, and $v(x)$ is smooth and strictly positive.

Let $\mathcal{K}(\cdot)$ be a continuous kernel function and h_N be the bandwidth. To obtain the local polynomial estimator of degree q for the regression function $m(x)$ based on the entire finite population, we denote

$$\boldsymbol{X}_{Ui} = \begin{bmatrix} 1 & x_1 - x_i & \dots & (x_1 - x_i)^q \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_N - x_i & \dots & (x_N - x_i)^q \end{bmatrix} \triangleq \begin{bmatrix} 1 & x_j - x_i & \dots & (x_j - x_i)^q \end{bmatrix}_{j \in U},$$

$\boldsymbol{y}_U = (y_1, \dots, y_N)' \triangleq [y_i]_{i \in U}$ and $W_{Ui} = \operatorname{diag}\left\{\frac{1}{h_N}\mathcal{K}\left(\frac{x_j - x_i}{h_N}\right)\right\}_{j \in U}$. Thus, the local polynomial estimator of degree q for the regression function $m(x)$ at x_i is given by

$$m_i = \boldsymbol{e}_1'(\boldsymbol{X}_{Ui}'W_{Ui}\boldsymbol{X}_{Ui})^{-1}\boldsymbol{X}_{Ui}'W_{Ui}\boldsymbol{y}_U \triangleq \boldsymbol{w}_{Ui}'\boldsymbol{y}_U,$$

where $\boldsymbol{e}_1$ is a vector of length $q + 1$ with one in the first position and zeros elsewhere. Note that m_i cannot be calculated in the present context,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

because only the y_k in s are known. A design-based estimator of m_i is given by

$$\hat{m}_i^o = \boldsymbol{e}_1'(\boldsymbol{X}_{si}'W_{si}\boldsymbol{X}_{si})^{-1}\boldsymbol{X}_{si}'W_{si}\boldsymbol{y}_s \triangleq \boldsymbol{w}_{si}^{o\prime}\boldsymbol{y}_s, \tag{5.2}$$

where $\boldsymbol{X}_{si} = \left[1, x_j - x_i, \ldots, (x_j - x_i)^q\right]_{j\in s}$, $W_{si} = \mathrm{diag}\left\{\frac{1}{\pi_j h_N}\mathcal{K}\left(\frac{x_j - x_i}{h_N}\right)\right\}_{j\in s}$ and $\boldsymbol{y}_s = \left[y_i\right]_{i\in s}$. Further, from the perspective of model-assisted estimation, we have the following local polynomial regression estimator for the population mean

$$\tilde{t}_{yr}^o = \frac{1}{N}\left(\sum_{i\in s}\frac{y_i - \hat{m}_i^o}{\pi_i} + \sum_{i\in U}\hat{m}_i^o\right). \tag{5.3}$$

In order to make $\boldsymbol{X}_{si}'W_{si}\boldsymbol{X}_{si}$ invertible for all x_i, Breidt and Opsomer (2000) proposed an adjusted sample estimator for m_i,

$$\hat{m}_i = \boldsymbol{e}_1'\left(\boldsymbol{X}_{si}'W_{si}\boldsymbol{X}_{si} + \mathrm{diag}\left\{\frac{\delta}{N^2}\right\}_{j=1}^{q+1}\right)^{-1}\boldsymbol{X}_{si}'W_{si}\boldsymbol{y}_s \triangleq \boldsymbol{w}_{si}'\boldsymbol{y}_s,$$

for some small $\delta > 0$. The nonparametric model-assisted estimator of $\bar{t}_y$ is then constructed by replacing $\hat{m}_i^o$ in (5.3) by $\hat{m}_i$,

$$\tilde{t}_{yr} = \frac{1}{N}\left(\sum_{i\in s}\frac{y_i - \hat{m}_i}{\pi_i} + \sum_{i\in U}\hat{m}_i\right). \tag{5.4}$$

From Theorem 2 of Breidt and Opsomer (2000), the design mean squared error of the nonparametric model-assisted estimator $\tilde{t}_{yr}$ is

$$\mathrm{MSE}_p(\tilde{t}_{yr}) = \frac{1}{N^2}\sum_{k,l\in U}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}(y_k - m_k)(y_l - m_l) + o(n^{-1}).$$

It is clear that, when the first-order inclusion probabilities of some units are relatively small, $\mathrm{MSE}_p(\tilde{t}_{yr})$ may become large due to inverse probability weighting.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

To overcome this shortcoming, we use the modified first-order inclusion probabilities $\{\pi_k^*\}_{k=1}^N$ defined by Definition 1 to improve the nonparametric model-assisted estimator. An improved design-based estimator of m_i is given by

$$\hat{m}_i^{o*} = \boldsymbol{e}_1'(\boldsymbol{X}_{si}'W_{si}^*\boldsymbol{X}_{si})^{-1}\boldsymbol{X}_{si}'W_{si}^*\boldsymbol{y}_s \triangleq \boldsymbol{w}_{si}^{o*'}\boldsymbol{y}_s,$$

where $W_{si}^* = \operatorname{diag}\left\{\frac{1}{\pi_j^* h_N}\mathcal{K}\left(\frac{x_j - x_i}{h_N}\right)\right\}_{j\in s}$. Then the improved nonparametric model-assisted (INMA) estimator for the population mean is

$$\tilde{t}_{yr}^{o*} = \frac{1}{N}\left(\sum_{i\in s}\frac{y_i - \hat{m}_i^{o*}}{\pi_i^*} + \sum_{i\in U}\hat{m}_i^{o*}\right). \tag{5.5}$$

Further, to make $\boldsymbol{X}_{si}'W_{si}^*\boldsymbol{X}_{si}$ invertible for all x_i, we adjust $\hat{m}_i^{o*}$ to

$$\hat{m}_i^* = \boldsymbol{e}_1'\left(\boldsymbol{X}_{si}'W_{si}^*\boldsymbol{X}_{si} + \operatorname{diag}\left\{\frac{\delta}{N^2}\right\}_{j=1}^{q+1}\right)^{-1}\boldsymbol{X}_{si}'W_{si}^*\boldsymbol{y}_s \triangleq \boldsymbol{w}_{si}^{*'}\boldsymbol{y}_s.$$

Finally, the INMA estimator for the population mean is given by

$$\tilde{t}_{yr}^* = \frac{1}{N}\left(\sum_{i\in s}\frac{y_i - \hat{m}_i^*}{\pi_i^*} + \sum_{i\in U}\hat{m}_i^*\right). \tag{5.6}$$

In the next section, we will discuss some properties of the new nonparametric model-assisted estimator, and theoretically show its effectiveness compared to the existing nonparametric model-assisted estimator.

6. Properties of INMA Estimation

In this section, we investigate the properties of the improved nonparametric model-assisted estimator. The following regularity conditions are required.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Condition 5. The errors $\tilde{\epsilon}_i$ $(i \in U)$ have compact support uniformly for all N.

Condition 6. The x_i $(i \in U)$ are independent and identically distributed as $F(x) = \int_{-\infty}^{x} f(t)dt$, where $f(\cdot)$ is a density with compact support $[a_x, b_x]$.

Condition 7. The mean function $m(\cdot)$ has the $(q+1)$th continuous derivative, and the variance function $v(\cdot)$ is bounded and strictly greater than 0.

Condition 8. The kernel $\mathcal{K}(\cdot)$ has compact support $[-1, 1]$, is symmetric and continuous, and satisfies $\int_{-1}^{1} \mathcal{K}(u)du = 1$.

Condition 9. As $N \to \infty$, the sampling fraction $f = n/N \to \pi \in$ (0,1], $h_N \to 0$ and $Nh_N^2/(\log\log N) \to \infty$.

Condition 10. Let $I_{i,k}(h_N) = I_{\{|x_k - x_i| \leq h_N\}}$ be an indicator function. Then as $n \to \infty$, $\sum_{k \in s} \frac{I_{i,k}(h_N)}{Nh_N\pi_k}$ is uniformly bounded in i and s.

Conditions 5 - 9 are common assumptions, which are used by Breidt and Opsomer (2000) and Breidt et al. (2007). Breidt and Opsomer (2000) discussed the rationality of these conditions in detail. Condition 10 is to ensure that the nonparametric model-assisted estimators are uniformly bounded in i and s. In particular, for the simple random sampling without replacement, Condition 10 is obtained from Lemma 1 of Breidt and Opsomer (2000).

We first study the design properties of the improved nonparametric

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

model-assisted estimator. Note from (5.5) that

$$\begin{aligned}
\tilde{t}_{yr}^{o*} &= \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i^*} + \frac{1}{N}\sum_{j\in U}\left(1-\frac{I_j}{\pi_j^*}\right)\boldsymbol{w}_{sj}^{o*'}\boldsymbol{y}_s \\
&= \frac{1}{N}\sum_{i\in s}\left\{\frac{1}{\pi_i^*}+\sum_{j\in U}\left(1-\frac{I_j}{\pi_j^*}\right)\boldsymbol{w}_{sj}^{o*'}\boldsymbol{e}_i\right\}y_i \\
&= \frac{1}{N}\sum_{i\in s}\omega_{is}^* y_i. \qquad (6.1)
\end{aligned}$$

Thus $\tilde{t}_{yr}^{o*}$ is a linear combination of the sample y_k's with the weights only depending on the auxiliary information. In addition, when the weights are applied to the auxiliary variables $\{1, x_i, \ldots, x_i^q\}$, we have

$$\begin{aligned}
\sum_s \omega_{is}^* x_i^l &= \sum_{i\in s}\left\{\frac{1}{\pi_i^*}+\sum_{j\in U}\left(1-\frac{I_j}{\pi_j^*}\right)\boldsymbol{w}_{sj}^{o*'}\boldsymbol{e}_i\right\}x_i^l \\
&= \sum_{i\in s}\frac{x_i^l}{\pi_i^*}+\sum_{j\in U}\left(1-\frac{I_j}{\pi_j^*}\right)x_j^l \\
&= \sum_{i\in U}x_i^l \qquad (6.2)
\end{aligned}$$

for $l = 0, \ldots, q$, where $\boldsymbol{e}_i$ is a vector with one on component i and zero elsewhere. (6.1) and (6.2) show the weighting and calibration properties of $\tilde{t}_{yr}^{o*}$ respectively. As in the case of linear model-assisted estimation, these are highly desirable properties in survey sampling.

Theorem 8. *If Conditions 2 and 4 - 10 are satisfied, and* $(nh_N^2\lambda_N^4)^{-1} = o(1)$*, then the improved estimator* $\tilde{t}_{yr}^*$*, defined in (5.6), is asymptotically design unbiased and design consistent. The design mean squared error of* $\tilde{t}_{yr}^*$ *is*

$$MSE_p(\tilde{t}_{yr}^*) = \frac{1}{N^2}\sum_{k,l\in U}\frac{\Delta_{kl}^*}{\pi_k^*\pi_l^*}(y_k-m_k)(y_l-m_l)+O\left(n^{-3/2}h_N^{-1}\lambda_N^{-4}\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Moreover, if $(nh_N^2\lambda_N^8)^{-1} = o(1)$ *and* $(n^{\kappa}\lambda_N^2\lambda_N^*)^{-1} = o(1)$*, then an asymptotically design unbiased and design consistent estimator of the design mean squared error is given by*

$$\widehat{AMSE}_p(\tilde{t}_{yr}^*) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\check{\Delta}_{kl}^*}{\pi_k^*\pi_l^*}(y_k - \hat{m}_k^*)(y_l - \hat{m}_l^*).$$

It is clear that Theorem 8 is similar to Theorems 1 - 3 of Breidt and Opsomer (2000) who discussed the design properties of $\tilde{t}_{yr}$.

The following theorem theoretically compares the efficiency of the two estimators $\tilde{t}_{yr}^*$ and $\tilde{t}_{yr}$.

Theorem 9. *If Conditions 2 and 4 - 10 are satisfied, and* $(nh_N^2\lambda_N^8)^{-1} = o(1)$*, then*

$$MSE_p(\tilde{t}_{yr}^*) \leq MSE_p(\tilde{t}_{yr}) + o\left(n^{-1}\right).$$

Theorem 9 implies that the MSE_p of the improved estimator $\tilde{t}_{yr}^*$ is asymptotically not larger than that of the traditional estimator $\tilde{t}_{yr}$.

7. Numerical Studies

In this section, we assess the empirical performances of our proposed estimators based on two simulation examples and one real data analysis. For each example, $D = 2000$ replicated samples are selected from a finite population by the unequal probability sampling, and then the squared-bias (Bias^2), variance (Var) and mean squared error (MSE) of each estimator are computed empirically.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

7.1 Linear model-assisted estimation

We generate a finite population U of size $N = 1000$ as follows:

$$y_k = \sqrt{12}\rho_1 \cdot x_{1k} + \sqrt{12}\rho_2 \cdot x_{2k} + \sqrt{12}\rho_3 \cdot x_{3k} + \sqrt{1-\rho_1^2-\rho_2^2-\rho_3^2} \cdot e_k, \quad k \in U,$$

where $x_{1k} \overset{\text{iid}}{\sim} U(0,1)$, x_{2k} and $x_{3k} \overset{\text{iid}}{\sim} U(1,2)$, $e_k \overset{\text{iid}}{\sim} N(0,1)$, and are all mutually independent. The correlation between y_k and x_{ik} can be controlled by ρ_i, and denote $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$. To realize πps sampling, we set the first-order inclusion probabilities $\pi_k \propto x_{1k}$ for all $k \in U$. Let the sampling fraction f vary at $\{0.02,\ 0.04,\ 0.06,\ 0.08,\ 0.10,\ 0.12\}$. We consider the following estimators of the finite population mean $\bar{t}_y$:

(a) The HT estimator given by (2.1).

(b) A trimmed Horvitz-Thompson (HT-Beta) estimator in which weights from upper tail of the weight distribution, say $1-F_\omega(\cdot) < 0.01$, are trimmed. Based on Beta distribution, the parameters in $F_\omega(\cdot)$ are estimated by the method-of-moment (Potter, 1990).

(c) A trimmed Horvitz-Thompson (HT-MSE) estimator in which threshold is based on an unbiased estimator of mean squared error (Potter, 1988).

(d) The robust Horvitz-Thompson (HT-Rob) estimator proposed by Beaumont et al. (2013).

(e) The IHT estimator given by (2.3).

(f) The Hájek (HA) estimator defined as the HT estimator with $\hat{N} = \sum_{k\in s} \pi_k^{-1}$ replacing N (Hájek, 1971).

(g) The trimmed Hájek (HA-Beta) estimator based on the weight distribution, analogous to (b).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

(h) The trimmed Hájek (HA-MSE) estimator based on an unbiased estimator of mean squared error, analogous to (c).

(i) The robust Hájek (HA-Rob) estimator, analogous to (d).

(j) The improved Hájek (IHA) estimator defined as the IHT estimator with $\hat{N}^* = \sum_{k\in s} \pi_k^{*-1}$ replacing N.

(k) The LMA estimator given by (3.3).

(l) The trimmed linear model-assisted (LMA-Trim) estimator with the calibration weights being trimmed, analogous to (b).

(m) The robust linear model-assisted (LMA-Rob) estimator proposed by Beaumont et al. (2013).

(n) The ILMA estimator given by (3.6).

Table 1 shows the empirical MSEs of the above estimators under different f and $\boldsymbol{\rho}$. We also set up a scenario where the first-order inclusion probabilities are equal inside classes defined by the quantiles of the variable x_{1k}. Tables 2 and 3 show the simulation results for the cases of 50 classes and 100 classes respectively. Under each scenario, the minimum MSE of all estimators is displayed in bold. It is clear that the accuracy of all estimators increases as the sample size n increases. The LMA-type estimators [(k) - (n)] are superior to the HA-type estimators [(f) - (j)], while the HA-type estimators are better than the HT-type estimators [(a) - (e)]. It is also observed that our proposed ILMA estimator has the best performance, followed by LMA-Rob estimator. This may be because the two methods not only make use of additional auxiliary information (x_{2k} and x_{3k}), but

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

also modify the weights based on the model unlike the LMA-Trim estimator which trims calibration weights by an estimated weight distribution. We mark the minimum MSE of HT-type estimators with a symbol ($*$), and it is found that the performance of the IHT estimator is the best. Similarly, for the HA-type estimators, the IHA estimator performs the best except for the case of low correlation (ρ_1=0.1).

In the supplementary material S2.1, we also report the biases and variances of all estimators, and the results are presented in Tables S1 - S6. It is obvious that the squared biases of all estimators are negligible compared to their variances. For each estimation type, the untrimmed estimators often have lower biases than the trimmed estimators.

Tables S15 - S17 show the empirical MSE of each estimator under small sample. In this case, our proposed estimators (IHT, IHA and ILMA) still have the best performance in their respective estimation types, but the LMA-type estimators are no longer the best compared to the HT-type and HA-type estimators. Based on the estimated conditional bias, the robust trimmed estimators (HT-Rob, HA-Rob and LMA-Rob) perform poorly when the sample size is small. There are little differences in MSEs between the trimmed estimators (HT-Beta, HT-mse or HA-Beta, HA-mse) and the untrimmed estimators (HT or HA), this may be because the thresholds obtained by these methods only slightly modify the first-order inclusion probabilities in the case of small sample.

Tables S18 - S20 show the empirical MSE of each estimator under the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 1: Empirical MSE of each estimator under different f and $\boldsymbol{\rho}$.

f	**0.02**	**0.04**	**0.06**	**0.08**	**0.10**	**0.12**
	$\boldsymbol{\rho}$=(0.1, 0.3, 0.8)					
HT	3.5529	1.3939	0.8923	0.6165	1.2633	0.5387
HT-Beta	1.1211	0.5925	0.4846	0.4187	0.3762	0.3428
HT-MSE	3.5992	1.3898	0.9216	0.6223	1.2820	0.5686
HT-Rob	1.7290	0.7976	0.5672	0.4173	0.5530	0.3299
IHT	0.7562*	0.4643*	0.3686*	0.2950*	0.2516*	0.2071*
HA	0.0750	0.0418	0.0298	0.0229	0.0192	0.0163
HA-Beta	0.0638	0.0308	0.0216	0.0159*	0.0115*	0.0098*
HA-MSE	0.0537*	0.0287 *	0.0211*	0.0162	0.0123	0.0107
HA-Rob	0.0641	0.0346	0.0246	0.0189	0.0149	0.0128
IHA	0.0613	0.0337	0.0244	0.0190	0.0145	0.0125
LMA	0.0200	0.0103	0.0071	0.0052	0.0041	0.0034
LMA-Trim	0.0847	0.0354	0.0252	0.0177	0.0131	0.0110
LMA-Rob	0.0196	0.0098	0.0066	**0.0049**	**0.0038**	**0.0031**
ILMA	**0.0195**	**0.0097**	**0.0066**	0.0049	0.0038	0.0031
	$\boldsymbol{\rho}$=(0.5, 0.4, 0.3)					
HT	2.7486	0.7330	0.4736	0.5075	0.2898	0.2587
HT-Beta	0.5178	0.2849	0.2355	0.1977	0.1826	0.1683
HT-MSE	3.2123	0.9176	0.6222	0.5845	0.3574	0.3117
HT-Rob	1.0575	0.3962	0.2748	0.2553	0.1777	0.1520
IHT	0.3368*	0.2178*	0.1744*	0.1370*	0.1139*	0.0958*
HA	0.0952	0.0534	0.0409	0.0331	0.0263	0.0210
HA-Beta	0.0812	0.0425	0.0320	0.0259	0.0219	0.0180
HA-MSE	0.0802	0.0436	0.0330	0.0261	0.0210	0.0165
HA-Rob	0.0813	0.0443	0.0333	0.0268	0.0214	0.0169
IHA	0.0756*	0.0407*	0.0300*	0.0236*	0.0187*	0.0146*
LMA	0.0485	0.0230	0.0162	0.0115	0.0098	0.0076
LMA-Trim	0.0911	0.0411	0.0296	0.0231	0.0202	0.0163
LMA-Rob	**0.0475**	**0.0219**	0.0153	0.0106	0.0092	0.0070
ILMA	0.0477	0.0221	**0.0152**	**0.0104**	**0.0091**	**0.0068**
	$\boldsymbol{\rho}$=(0.8, 0.3, 0.1)					
HT	0.2833	0.1488	0.0968	0.0805	0.0626	0.0469
HT-Beta	0.1850	0.0957	0.0696	0.0585	0.0511	0.0474
HT-MSE	0.5569	0.2588	0.1625	0.1348	0.1003	0.0704
HT-Rob	0.1837	0.1035	0.0707	0.0573	0.0453	0.0378
IHT	0.1194*	0.0731*	0.0528*	0.0415*	0.0342*	0.0295*
HA	0.1123	0.0616	0.0438	0.0350	0.0277	0.0237
HA-Beta	0.0986	0.0540	0.0390	0.0318	0.0263	0.0245
HA-MSE	0.1096	0.0615	0.0414	0.0317	0.0252	0.0224
HA-Rob	0.0999	0.0555	0.0377	0.0297	0.0237	0.0209
IHA	0.0852*	0.0463*	0.0322*	0.0246*	0.0199*	0.0178*
LMA	0.0250	0.0118	0.0082	0.0057	0.0047	0.0039
LMA-Trim	0.0491	0.0259	0.0207	0.0182	0.0161	0.0148
LMA-Rob	0.0248	0.0114	**0.0078**	**0.0053**	**0.0043**	**0.0036**
ILMA	**0.0243**	**0.0113**	0.0078	0.0053	0.0044	0.0037

[1] Under each scenario, the minimum MSE of all estimators is displayed in bold.

[2] The minimum for each estimation type, except for the bold values, is marked with a symbol (∗).

[3] Since four decimal places are reserved, some values in the table have the same display.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 2: Empirical MSE of each estimator in the case of 50 classes.

f	**0.02**	**0.04**	**0.06**	**0.08**	**0.10**	**0.12**
			$\boldsymbol{\rho}$=(0.1, 0.3, 0.8)			
HT	2.2006	0.8781	0.6466	0.4426	0.3678	0.2974
HT-Beta	1.1506	0.5299	0.4195	0.3324	0.2933	0.2585
HT-MSE	2.1459	0.8739	0.6444	0.4355	0.3604	0.2945
HT-Rob	1.3062	0.6047	0.4686	0.3439	0.2877	0.2342
IHT	0.6857*	0.4302*	0.3506*	0.2494*	0.2392*	0.1875*
HA	0.0877	0.0461	0.0333	0.0232	0.0208	0.0161
HA-Beta	0.0720	0.0355	0.0232	0.0155*	0.0138*	0.0103*
HA-MSE	0.0607*	0.0332*	0.0225*	0.0157	0.0144	0.0110
HA-Rob	0.0738	0.0388	0.0270	0.0189	0.0171	0.0130
IHA	0.0672	0.0379	0.0263	0.0180	0.0171	0.0128
LMA	0.0222	0.0111	0.0077	0.0053	0.0045	0.0038
LMA-Trim	0.0856	0.0371	0.0235	0.0155	0.0136	0.0102
LMA-Rob	0.0218	**0.0107**	**0.0074**	0.0050	**0.0042**	0.0035
ILMA	**0.0217**	0.0107	0.0074	**0.0050**	0.0042	**0.0034**
			$\boldsymbol{\rho}$=(0.5, 0.4, 0.3)			
HT	1.1110	0.5450	0.3520	0.2498	0.1880	0.1626
HT-Beta	0.5512	0.2805	0.1980	0.1729	0.1512	0.1357
HT-MSE	1.4694	0.6952	0.4255	0.2961	0.2203	0.1845
HT-Rob	0.6424	0.3447	0.2394	0.1786	0.1443	0.1222
IHT	0.3340*	0.2120*	0.1592*	0.1359*	0.1087*	0.0915*
HA	0.0953	0.0562	0.0378	0.0297	0.0243	0.0199
HA-Beta	0.0750	0.0405	0.0273	0.0208*	0.0176	0.0157
HA-MSE	0.0735	0.0423	0.0287	0.0219	0.0183	0.0156
HA-Rob	0.0777	0.0446	0.0303	0.0231	0.0192	0.0161
IHA	0.0692*	0.0389*	0.0269*	0.0212	0.0175*	0.0147*
LMA	0.0428	0.0207	0.0141	0.0104	0.0085	0.0076
LMA-Trim	0.0761	0.0374	0.0242	0.0184	0.0147	0.0138
LMA-Rob	0.0423	0.0198	0.0134	0.0096	0.0078	0.0069
ILMA	**0.0418**	**0.0194**	**0.0131**	**0.0096**	**0.0077**	**0.0068**
			$\boldsymbol{\rho}$=(0.8, 0.3, 0.1)			
HT	0.2104	0.1073	0.0697	0.0576	0.0432	0.0346
HT-Beta	0.1600	0.0862	0.0602	0.0477	0.0420	0.0385
HT-MSE	0.4259	0.1874	0.1110	0.0913	0.0630	0.0510
HT-Rob	0.1498	0.0853	0.0575	0.0456	0.0363	0.0304
IHT	0.1045*	0.0706*	0.0486*	0.0357*	0.0321*	0.0264*
HA	0.1140	0.0668	0.0447	0.0334	0.0280	0.0241
HA-Beta	0.0943	0.0541	0.0379	0.0297	0.0270	0.0229
HA-MSE	0.1071	0.0608	0.0407	0.0304	0.0263	0.0210
HA-Rob	0.0969	0.0565	0.0378	0.0283	0.0247	0.0202
IHA	0.0812*	0.0475*	0.0315*	0.0233*	0.0216*	0.0170*
LMA	0.0220	0.0113	0.0077	0.0057	0.0045	0.0042
LMA-Trim	0.0503	0.0285	0.0203	0.0182	0.0159	0.0142
LMA-Rob	0.0216	0.0107	0.0071	0.0052	0.0041	0.0037
ILMA	**0.0214**	**0.0105**	**0.0069**	**0.0050**	**0.0040**	**0.0035**

[1] Under each scenario, the minimum MSE of all estimators is displayed in bold.

[2] The minimum for each estimation type, except for the bold values, is marked with a symbol (∗).

[3] Since four decimal places are reserved, some values in the table have the same display.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 3: Empirical MSE of each estimator in the case of 100 classes.

f	**0.02**	**0.04**	**0.06**	**0.08**	**0.10**	**0.12**
	$\boldsymbol{\rho}$=(0.1, 0.3, 0.8)					
HT	2.8816	1.3489	0.8871	0.6931	0.5217	0.4772
HT-Beta	1.1556	0.6403	0.4566	0.3998	0.3585	0.3382
HT-MSE	2.8965	1.3218	0.8664	0.6821	0.5188	0.4677
HT-Rob	1.5082	0.8107	0.5551	0.4385	0.3504	0.3140
IHT	0.7501*	0.4976*	0.3508*	0.2773*	0.2416*	0.1991*
HA	0.0789	0.0451	0.0331	0.0258	0.0197	0.0189
HA-Beta	0.0666	0.0342	0.0219	0.0162*	0.0121*	0.0106*
HA-MSE	0.0591*	0.0325*	0.0217*	0.0168	0.0127	0.0116
HA-Rob	0.0686	0.0378	0.0259	0.0201	0.0150	0.0140
IHA	0.0653	0.0359	0.0241	0.0186	0.0142	0.0126
LMA	0.0241	0.0109	0.0071	0.0057	0.0042	0.0035
LMA-Trim	0.0899	0.0407	0.0238	0.0169	0.0132	0.0109
LMA-Rob	0.0240	0.0107	0.0069	0.0055	0.0040	0.0033
ILMA	**0.0236**	**0.0107**	**0.0068**	**0.0054**	**0.0040**	**0.0032**
	$\boldsymbol{\rho}$=(0.5, 0.4, 0.3)					
HT	1.0895	0.6551	0.4665	0.4585	0.3573	0.2343
HT-Beta	0.5378	0.2745	0.2271	0.1942	0.1696	0.1543
HT-MSE	1.3852	0.8410	0.5160	0.4882	0.3811	0.2640
HT-Rob	0.6196	0.3652	0.2727	0.2551	0.1899	0.1416
IHT	0.3511*	0.2046*	0.1717*	0.1358*	0.1131*	0.0909*
HA	0.0878	0.0463	0.0352	0.0265	0.0229	0.0186
HA-Beta	0.0762	0.0361	0.0267	0.0213	0.0169	0.0141
HA-MSE	0.0751	0.0369	0.0279	0.0214	0.0167	0.0139
HA-Rob	0.0746	0.0376	0.0288	0.0221	0.0179	0.0147
IHA	0.0703*	0.0346*	0.0259*	0.0197*	0.0158*	0.0132*
LMA	0.0431	0.0195	0.0154	0.0108	0.0092	0.0074
LMA-Trim	0.0770	0.0310	0.0225	0.0177	0.0142	0.0116
LMA-Rob	**0.0422**	**0.0187**	**0.0143**	0.0099	0.0084	0.0067
ILMA	0.0424	0.0188	0.0145	**0.0098**	**0.0083**	**0.0066**
	$\boldsymbol{\rho}$=(0.8, 0.3, 0.1)					
HT	0.4907	0.2211	0.1332	0.1044	0.0819	0.0672
HT-Beta	0.1848	0.0946	0.0766	0.0668	0.0583	0.0519
HT-MSE	1.2444	0.4291	0.2405	0.1853	0.1388	0.1108
HT-Rob	0.2428	0.1247	0.0832	0.0647	0.0535	0.0440
IHT	0.1228*	0.0717*	0.0570*	0.0452*	0.0367*	0.0319*
HA	0.1181	0.0655	0.0502	0.0360	0.0308	0.0249
HA-Beta	0.0958	0.0523	0.0437	0.0330	0.0308	0.0251
HA-MSE	0.1064	0.0585	0.0461	0.0326	0.0289	0.0223
HA-Rob	0.0975	0.0543	0.0422	0.0297	0.0266	0.0208
IHA	0.0805*	0.0451*	0.0363*	0.0251*	0.0223*	0.0176*
LMA	0.0252	0.0138	0.0093	0.0071	0.0055	0.0047
LMA-Trim	0.0490	0.0270	0.0232	0.0191	0.0183	0.0167
LMA-Rob	0.0248	0.0130	0.0085	0.0064	**0.0050**	**0.0042**
ILMA	**0.0242**	**0.0128**	**0.0084**	**0.0064**	0.0051	0.0043

[1] Under each scenario, the minimum MSE of all estimators is displayed in bold.

[2] The minimum for each estimation type, except for the bold values, is marked with a symbol (∗).

[3] Since four decimal places are reserved, some values in the table have the same display.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

misspecified model without considering the covariate $\{x_{3k}\}$. Compared to Tables 1 - 3, the efficiency of LMA-type estimators decreases. Specially, when the misspecified degree of model is high ($\rho_3 = 0.8$), the performances of LMA-type estimators are worse than those of HA-type estimators. It is observed that our proposed estimators (IHT, IHA and ILMA) perform the best overall in their respective estimation types.

Additionally, we compare the empirical performances of the LMA and ILMA estimators in terms of their MSEs, biases, variances and coverage rates. From Tables S21 - S23, the improved estimator has smaller MSE than the traditional estimator, and the threshold K^* decreases as the sample size increases. Under the same sample size, the MSEs in the case $\boldsymbol{\rho} = (0.5, 0.4, 0.3)$ are higher than those in other cases, $\boldsymbol{\rho} = (0.1, 0.3, 0.8)$ and $\boldsymbol{\rho} = (0.8, 0.3, 0.1)$. This may be due to their different SNRs (Signal Noise Ratios). It is found that the coverage rates CR1 and CR2 are roughly the same under various scenarios with the latter corresponding to a shorter interval length, and both grow when n increases.

7.2 Nonparametric model-assisted estimation

Similar to the settings of Breidt and Opsomer (2000), we consider the following mean functions:

$$\begin{aligned} &\text{Linear:}\quad m_1(x) = 1 + 2(x - 0.5), && \text{Cycle1:}\ m_4(x) = 2 + \sin(2\pi x),\\ &\text{Quadratic:}\ m_2(x) = 1 + 2(x - 0.5)^2, && \text{Cycle4:}\ m_5(x) = 2 + \sin(8\pi x),\\ &\text{Exponential:}\ m_3(x) = \exp\{-8x\}, && \text{CDF:}\quad m_6(x) = \Phi\left(\tfrac{1.5-2x}{0.4}\right), \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

where Φ is the standard normal cdf. For each mean function, we generate a finite population U of size $N = 1000$ based on the superpopulation model (5.1) with $x_i \overset{\text{iid}}{\sim} U(0,1)$ and $\tilde{\epsilon}_i \overset{\text{iid}}{\sim} N(0, 0.1^2)$. In order to realize πps sampling, we set the first-order inclusion probabilities $\pi_k \propto x_k$ for $k \in U$, and the sampling fraction $f = 0.06$, 0.10. In this subsection, we compare the performances of ten estimators:

(i) The HT-type estimators, (a) - (e) in Subsection 7.1.

(ii) The LMA-type estimators with $\boldsymbol{x}_k = (1, x_k)$, (k) - (n) in Subsection 7.1.

(iii) The nonparametric model-assisted (NMA) estimator given by (5.4).

(iv) The INMA estimator given by (5.6).

The Epanechnikov kernel $\mathcal{K}(t) = 3/4(1 - t^2)I_{\{|t|\leq 1\}}$ is used for all nonparametric model-assisted estimators. We set the degree $q = 1$ and two different bandwidths h_N =0.1, 0.25. The empirical MSEs of all estimators under various superpopulation models are provided in Table 4. In addition, Tables 5 and 6 report the simulation results where the inclusion probabilities are equal inside classes.

Similar to the simulation results in Subsection 7.1, from Tables 4 - 6, we see that the IHT and ILMA estimators perform the best overall in HT-type and LMA-type estimators respectively, and the model-assisted approaches are generally better than the design-based approaches. Note that the modified HT-type estimators perform well under the linear superpopulation model. This may be because the sample is drawn by πps sampling with the first-order inclusion probability proportional to x. In addition, our proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 4: Empirical MSE of each estimator under different models.

f	**0.06**	**0.10**	**0.06**	**0.10**	**0.06**	**0.10**
	Linear		Quadratic		Exponential	
HT	0.00058	0.00037	0.10803	0.33274	0.06180	0.02202
HT-Beta	0.00031	0.00020	0.02255	0.01805	0.00453	0.00384
HT-mse	0.02885	0.05023	0.07371	0.19313	0.00659	0.00479
HT-Rob	0.00031	0.00019	0.04420	0.09744	0.01916	0.00823
IHT	**0.00028**	**0.00015**	0.01752*	0.01247*	0.00387*	0.00289*
LMA	0.00034	0.00019	0.00221	0.00134	0.00273	0.00194
LMA-Trim	0.00569	0.00614	0.00216	0.00141	0.00316	0.00288
LMA-Rob	0.00033	0.00018	0.00226	0.00135	0.00260	0.00180
ILMA	0.00033*	0.00018*	**0.00198**	0.00115*	**0.00222**	**0.00152**
NMA$_1$	0.03146	0.00339	0.02867	0.02003	0.03844	0.01338
INMA$_1$	0.03146	0.00339	0.02866	0.02003	0.03844	0.01315
NMA$_2$	0.01320	0.00065	0.01420	0.00062	0.00441	0.00173
INMA$_2$	0.01320*	0.00065*	0.01420*	**0.00062**	0.00440*	0.00173*
	Cycle1		Cycle4		CDF	
HT	0.17389	0.17535	0.25265	0.19873	1.03E-04	5.46E-05
HT-Beta	0.10474	0.07165	0.07353	0.05697	1.03E-04*	5.45E-05*
HT-mse	0.17735	0.17287	0.21400	0.16959	1.03E-04	5.46E-05
HT-Rob	0.12114	0.09103	0.11934	0.08852	1.06E-04	5.55E-05
IHT	0.08461*	0.05351*	0.06166*	0.04423*	1.03E-04	5.46E-05
LMA	0.00862	0.00657	0.01481	0.00881	8.05E-05	4.62E-05
LMA-Trim	0.01155	0.00841	**0.01341**	**0.00827**	4.19E-04	3.49E-04
LMA-Rob	0.00796	0.00604	0.01524	0.00888	7.99E-05	4.47E-05
ILMA	0.00723*	0.00530*	0.01423	0.00862	7.69E-05*	4.33E-05*
NMA$_1$	0.02129	0.02372	0.04034	0.03526	3.04E-07	1.56E-07
INMA$_1$	0.02117	0.02358	0.04029*	0.03524	**3.04E-07**	**1.56E-07**
NMA$_2$	0.00662	0.00162	0.06441	0.02920	9.69E-06	4.85E-06
INMA$_2$	**0.00660**	**0.00161**	0.06421	0.02911*	9.61E-06	4.81E-06

[1] NMA$_1$ and INMA$_1$ are nonparametric model-assisted estimators with $h_N = 0.1$.

[2] NMA$_2$ and INMA$_2$ are nonparametric model-assisted estimators with $h_N = 0.25$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 5: Empirical MSE of each estimator in the case of 50 classes.

f	**0.06**	**0.10**	**0.06**	**0.10**	**0.06**	**0.10**
	Linear		Quadratic		Exponential	
HT	0.00034	0.00020	0.02011	0.01233	0.00294	0.00173
HT-Beta	0.00029	0.00019	0.01708	0.01183	0.00221	0.00144
HT-mse	0.01079	0.00795	0.01782	0.01139	0.00240	0.00156
HT-Rob	**0.00024**	**0.00014**	0.01712	0.01113	0.00226	0.00150
IHT	0.00026	0.00015	0.01301*	0.00941*	0.00180*	0.00126*
LMA	0.00028	0.00017	0.00140	0.00078	0.00096	0.00062
LMA-Trim	0.00638	0.00661	0.00135	0.00085	0.00096	0.00073
LMA-Rob	0.00027	0.00016	0.00141	0.00081	0.00088	0.00059
ILMA	0.00027*	0.00016*	0.00124*	0.00073*	0.00080*	0.00054*
NMA$_1$	0.00053	0.00042	0.00250	0.00076	0.00074	0.00034
INMA$_1$	0.00053*	0.00042	0.00250	0.00076	0.00074	0.00034
NMA$_2$	0.00094	0.00036	0.00049	0.00021	0.00060	0.00024
INMA$_2$	0.00094	0.00036*	**0.00049**	**0.00021**	**0.00060**	**0.00024**
	Cycle1		Cycle4		CDF	
HT	0.07327	0.04546	0.03884	0.02230	1.25E-04	7.15E-05
HT-Beta	0.08094	0.05147	0.03689	0.02208	1.25E-04*	7.12E-05*
HT-mse	0.07654	0.04677	0.03496	0.02039*	1.25E-04	7.15E-05
HT-Rob	0.07499	0.04602	0.03464	0.02087	1.30E-04	7.30E-05
IHT	0.06772*	0.04176*	0.03197*	0.02067	1.25E-04	7.15E-05
LMA	0.00360	0.00213	0.01206	0.00649	8.75E-05*	4.91E-05*
LMA-Trim	0.00528	0.00324	0.01142*	0.00616*	1.36E-04	8.08E-05
LMA-Rob	0.00354	0.00212	0.01283	0.00678	9.09E-05	5.01E-05
ILMA	0.00343*	0.00202*	0.01186	0.00641	8.77E-05	4.91E-05
NMA$_1$	0.00367	0.00060	0.00814	0.00347	3.72E-07	1.99E-07
INMA$_1$	0.00367	0.00060	**0.00813**	**0.00346**	**3.72E-07**	**1.99E-07**
NMA$_2$	0.00112	0.00050	0.02011	0.01019	1.03E-05	5.60E-06
INMA$_2$	**0.00111**	**0.00049**	0.01993	0.01012	1.02E-05	5.59E-06

[1] NMA$_1$ and INMA$_1$ are nonparametric model-assisted estimators with $h_N = 0.1$.

[2] NMA$_2$ and INMA$_2$ are nonparametric model-assisted estimators with $h_N = 0.25$.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 6: Empirical MSE of each estimator in the case of 100 classes.

f	**0.06**	**0.10**	**0.06**	**0.10**	**0.06**	**0.10**
	Linear		Quadratic		Exponential	
HT	0.00048	0.00024	0.02895	0.01947	0.00552	0.00344
HT-Beta	0.00035	0.00023	0.02102	0.01590	0.00320	0.00235
HT-mse	0.03140	0.01564	0.02493	0.01700	0.00386	0.00271
HT-Rob	0.00028	**0.00015**	0.02230	0.01542	0.00373	0.00261
IHT	**0.00028**	0.00015	0.01553*	0.01093*	0.00271*	0.00200*
LMA	0.00031	0.00017	0.00173	0.00102	0.00155	0.00099
LMA-Trim	0.00769	0.00762	0.00171	0.00108	0.00184	0.00153
LMA-Rob	0.00030	0.00016*	0.00175	0.00101	0.00144	0.00094
ILMA	0.00029*	0.00016	0.00156*	0.00088*	0.00128*	0.00083*
NMA$_1$	0.00073	0.00063	0.00531	0.00166	0.00145	0.00073
INMA$_1$	0.00073*	0.00063	0.00531	0.00166	0.00145	0.00073
NMA$_2$	0.00117	0.00043	0.00102	0.00035	0.00092	0.00038
INMA$_2$	0.00117	0.00043*	**0.00102**	**0.00035**	**0.00092**	**0.00038**
	Cycle1		Cycle4		CDF	
HT	0.09935	0.05699	0.07824	0.04604	1.10E-04	6.06E-05
HT-Beta	0.09763	0.06322	0.05850	0.04023	1.10E-04*	6.04E-05*
HT-mse	0.10184	0.05860	0.06227	0.03891	1.10E-04	6.06E-05
HT-Rob	0.09379	0.05573	0.06176	0.03997	1.14E-04	6.19E-05
IHT	0.07837*	0.04777*	0.04721*	0.03244*	1.10E-04	6.06E-05
LMA	0.00494	0.00334	0.01286	0.00832	7.85E-05	4.27E-05
LMA-Trim	0.00819	0.00568	**0.01136**	**0.00774**	2.52E-04	1.81E-04
LMA-Rob	0.00473	0.00324	0.01329	0.00855	8.06E-05	4.30E-05
ILMA	0.00439*	0.00299*	0.01223	0.00799	7.79E-05*	4.22E-05*
NMA$_1$	0.00924	0.00212	0.01890	0.00792	3.41E-07	1.69E-07
INMA$_1$	0.00924	0.00212	0.01889*	0.00792*	**3.41E-07**	**1.69E-07**
NMA$_2$	0.00190	0.00078	0.03874	0.01871	1.02E-05	5.03E-06
INMA$_2$	**0.00188**	**0.00078**	0.03861	0.01863	1.01E-05	5.00E-06

[1] NMA$_1$ and INMA$_1$ are nonparametric model-assisted estimators with $h_N = 0.1$.

[2] NMA$_2$ and INMA$_2$ are nonparametric model-assisted estimators with $h_N = 0.25$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

INMA estimator outperforms the traditional NMA estimator, while their accuracy is affected by the bandwidths. Under the complex model, the NMA estimator is better than the LMA estimator when an appropriate bandwidth is selected. Tables S7 - S12 in the supplementary material S2.1 show the biases and variances corresponding to Tables 4 - 6. Similar to Tables S1 - S6, the squared biases of all estimators are negligible and the untrimmed estimators often have lower biases than the trimmed estimators.

7.3 Empirical example

We use a real data set called “BigLucy” in R package “TeachingSampling” to compare the estimators given in Subsection 7.1. This data set is a full business population database, and includes some financial variables of 85,396 industrial companies of a city in a particular fiscal year. $D = 2000$ replicated samples are selected from the data set by Poisson sampling, and the first-order inclusion probability is proportional to the total amount of a company’s earnings (Income, denoted by x). We report some numerical characteristics and histograms of the inclusion probabilities in the supplementary material S2.5. Let n_0 be the sum of all first-order inclusion probabilities. For the LMA-type estimators, the auxiliary variable is set to be $\boldsymbol{x} = (1, x)$. We focus on estimating the mean number of employees in the company.

Table 7 reports the empirical MSE of each estimator for the “BigLucy” data set. Similar to the conclusions in Subsection 7.1, the accuracy of all

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Table 7: Empirical MSE of each estimator for the "BigLucy" data set.

n_0	**256**	**512**	**853**	**1279**	**1706**	**2132**
HT	28.6199	12.1969	8.6787	6.3637	4.2780	3.6931
HT-Beta	20.1850	11.5708	7.3177	5.1427	4.2289	3.8235
HT-mse	37.3624	14.3698	9.5720	7.3897	5.0499	4.1914
HT-Rob	22.6932	11.4337	7.0996	4.7703	3.4765	3.0529
IHT	19.8424*	10.9296*	6.5290*	4.1379*	3.1888*	2.7307*
HA	5.9226	2.9018	1.8254	1.3513	0.9601	0.8111
HA-Beta	5.2591	2.6245	1.7547	1.1579	0.8529	0.8150
HA-mse	5.2330	2.5328	1.6285	1.0518*	0.7318	0.6494
HA-Rob	5.5099	2.6745	1.7026	1.1371	0.7906	0.7302
IHA	5.1643*	2.5235*	1.6140*	1.0592	0.7317*	0.6426*
LMA	3.9425	2.0699	1.2487	0.9047	0.6417	0.5581
LMA-Trim	4.1171	2.2671	1.4558	1.1946	0.8941	0.7629
LMA-Rob	3.9132	2.0377	1.1750	0.8178	0.5696	0.4978
ILMA	**3.7602**	**1.9636**	**1.1248**	**0.7802**	**0.5262**	**0.4722**

estimators increases as n_0 increases. It is observed that the LMA-type estimators are the best, and the HT-type estimators are the worst. The improved estimators obtained by probability thresholding perform the best in their respective estimation types, and the threshold K^* (= 1800, 1080, 756, 576, 504, 451) decreases as n_0 increases. Tables S13 - S14 in the supplementary material S2.1 show the biases and variances corresponding to Table 7. It can be found that the squared biases of all estimators are negligible, and the untrimmed estimators have lower biases than the trimmed estimators.

8. Concluding Remarks

In this paper, using the modified first-order inclusion probabilities given by Zong et al. (2019), we have developed an improved model-assisted estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

tion approach via probability thresholding. Like classical model-assisted estimators, we have established the design properties of the improved model-assisted estimators including calibration, design consistency and asymptotic design unbiasedness. The design mean squared errors and their estimators for the proposed estimators have also been derived. Moreover, we have theoretically compared the accuracy of the classical model-assisted estimators and the improved model-assisted estimators. Two simulation experiments and a real data analysis illustrate the promising of our method.

Our theoretical development on the model-assisted estimation could be extended to the situations of some other superpopulation models, such as semiparametric model-assisted estimation (Breidt et al., 2007), single-index model-assisted estimation (Wang, 2009) and nonparametric additive model-assisted estimation (Wang and Wang, 2011), and these warrant our further researches. When the second-order inclusion probabilities of some units are relatively small, the variance estimators will be very unstable. So it is worth studying how to modify the first-order and second-order inclusion probabilities by the hard-threshold method in order for improving the variance estimators. Additionally, the hard-threshold method is also useful to some other sampling designs such as multi-stage sampling, systematic sampling with unequal probabilities and adaptive cluster sampling, and other statistical problems like the treatment of missing data where inverse probability weighting is often used.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

Supplementary Material

The supplementary material contains the proofs of theoretical results and additional numerical results.

Acknowledgments

The authors thank the co-editors, an associate editor and three referees for their constructive comments and suggestions which have substantially improved the original manuscript. Zong's work was partially supported by the National Natural Science Foundation of China (Grant Nos.12201018 and 71973116). Tso's work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region of China (Grant No.11507817). Zou's work was partially supported by the Beijing Outstanding Young Scientist Program (Grant No. JWZQ20240101027) and the National Natural Science Foundation of China (Grant Nos.12031016 and 11971323).

References

Beaumont, J. F., D. Haziza, and A. Ruiz-Gazen (2013). A unified approach to robust estimation in finite population sampling. *Biometrika 100*, 555–569.

Benrud, C. H. (1978). Final report on national assessment of educational progress: sampling and weighting activities for assessment year 11. Research Triangle Park, North Carolina: National Assessment of Education Progress.

Boistard, H., H. P. Lopuhaä, and A. Ruiz-Gazen (2012). Approximation of rejective sampling

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics 6*, 1967–1983.

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.

Breidt, F. J., J. D. Opsomer, A. A. Johnson, and M. G. Ranalli (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology 33*, 35–44.

Cardot, H., M. Chaouch, C. Goga, and C. Labruère (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference 140*, 75–91.

Cardot, H. and E. Josserand (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika 98*, 107–118.

Chen, Q., M. R. Elliott, D. Haziza, Y. Yang, M. Ghosh, R. J. A. Little, J. Sedransk, and M. Thompson (2017). Approaches to improving survey-weighted estimates. *Statistical Science 32*, 227–248.

Delevoye, A. and F. Sävje (2020). Consistency of the Horvitz-Thompson estimator under general sampling and experimental designs. *Journal of Statistical Planning and Inference 207*, 190–197.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Favre-Martinoz, C., D. Haziza, and J. F. Beaumont (2015). A method of determining the winsorization threshold, with an application to domain estimation. *Survey Methodology 41*, 55–77.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

finite population. *The Annals of Mathematical Statistics 35*, 1491–1523.

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston, 236.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Kokic, P. N. and P. A. Bell (1994). Optimal winsorising cutoffs for a stratified finite population estimator. *Journal of Official Statistics 10*, 419–435.

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association 100*, 1429–1442.

Potter, F. (1988). Survey of procedures to control extreme sampling weights. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 453–458. American Statistical Association, Alexandria, VA.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 225–230. American Statistical Association, Alexandria, VA.

Rivest, L. P. and D. Hurtubise (1995). On Searls' winsorized means for skewed populations. *Survey Methodology 21*, 107–116.

Robinson, P. M. and C. E. Särndal (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: Indian Journal of Statistics, Series B 45*, 240–248.

Rosenbaum, P. R. (2002). *Observational Studies.* New York: Springer.

Särndal, C. E., B. Swensson, and J. H. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0295

REFERENCES

Wang, H., R. Zhu, and P. Ma (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association 113*, 829–844.

Wang, L. (2009). Single-index model-assisted estimation in survey sampling. *Journal of Nonparametric Statistics 21*, 487–504.

Wang, L. and S. Wang (2011). Nonparametric additive model-assisted estimation for survey data. *Journal of Multivariate Analysis 102*, 1126–1140.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.

Zong, X., R. Zhu, and G. Zou (2019). Improved Horvitz-Thompson estimator in survey sampling. *Survey Methodology 45*, 165–184.

School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing 100124, China

E-mail: xpzong@bjut.edu.cn

Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

E-mail: msgtso@cityu.edu.hk

School of Mathematical Sciences, Capital Normal University, Beijing 100048, China

E-mail: ghzou@amss.ac.cn   Tel.: 86-15001122035

Statistica Sinica