Commun. Math. Res.
doi: 10.4208/cmr.2021-0056

Vol. **38**, No. 1, pp. 81-98
February 2022

Model-Assisted Estimators with Auxiliary Functional Data

Chao Liu[1], Huiming Zhang[2,3,*] and Jing Yan[4]

[1] *Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China.*
[2] *Department of Mathematics, Faculty of Science and Technology, University of Macau, Taipa Macau, China.*
[3] *UMacau Zhuhai Research Institute, Zhuhai, China.*
[4] *School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China.*

Received 7 May 2021; Accepted 12 August 2021

Abstract. Few studies focus on the application of functional data to the field of design-based survey sampling. In this paper, the scalar-on-function regression model-assisted method is proposed to estimate the finite population means with auxiliary functional data information. The functional principal component method is used for the estimation of functional linear regression model. Our proposed functional linear regression model-assisted (FLR-assisted) estimator is asymptotically design-unbiased, consistent under mild conditions. Simulation experiments and real data analysis show that the FLR-assisted estimators are more efficient than the Horvitz-Thompson estimators under different sampling designs.

AMS subject classifications: 62K25, 62D05

Key words: Survey sampling, semi-supervised inference, model-assisted estimator, Horvitz-Thompson estimator, functional linear regression.

1 Introduction

In survey sampling, the auxiliary information is often available for all units of the finite population of interest, which can be used to improve the precision of

*Corresponding author. *Email address:* huimingzhang@um.edu.mo (H. Zhang)

estimators. Särndal *et al.* [20] provided a fundamental framework for the estimation of finite population means with the help of auxiliary information, which assumes a superpopulation model to describe the relationship between the auxiliary variable and the study variable. It was called the model-assisted method. While in [19], a linear regression model was assumed to be the superpopulation model, which obtained improved estimators with the aid of auxiliary variables. Following this idea, many researchers use the model-assisted method to construct estimators based on the entire finite population and sampling design under some predefined superpopulation models. For example, Breidt and Opsomer [5] proposed a nonparametric model-assisted estimator based on local polynomial regression. Zhang *et al.* [22] considered a similar problem from the perspective of semi-supervised learning, which is a particular case of Robinson and Särndal [19] when the sampling design was assumed to be simple random sampling. By a geographically weighted regression model-assisted method, Liu *et al.* [16] proposed to estimate the finite population totals using survey data with the aid of a spatially varying coefficient model. To reduce the variance of the estimated treatment effect, Bloniarz *et al.* [3] studied the Lasso-adjusted average treatment effect (ATE) estimate under the Neyman-Rubin model for randomization by adjusting for covariates. Other researches on model-assisted estimators based on nonparametric and semiparametric models can be seen in Breidt and Opsomer [6] and references therein.

All the model-assisted estimators mentioned above are considered with the superpopulation model where auxiliary variable is assumed to be a scalar or a vector. Under the framework of experimental design, the problem of design choice in function-on-scalar regression was studied by Cuevas *et al.* [9] whose consideration is more complicate than in the ordinary finite-dimensional regression. Following this functional framework, Cardot *et al.* [7, 8] developed model-assisted approaches, which enable to use auxiliary vector data. When dealing with the whole functional sample in Big Data, Aaron *et al.* [1] studied how to combine estimators from different subsamples by the popular method of "divide and conquer".

From the perspective of survey sampling, few researches have considered the model-assisted estimation of population totals or means in which the auxiliary variable is functional data through scalar-on-function regression. In fact, recent technology with practical applications can generate an increasing amount of functional data of which each observation represents a curve or a function instead of a scalar or multivariate vector. Functional data analysis (FDA) has gained increasing attention in modern data analysis due to the advances in data recording techniques. FDA is of paramount importance in the field of modern

data analysis, and a lot of monographs emerge, see Ramsay and Silverman [17], Hsing and Eubank [12], Kokoszka and Reimherr [14]. Functional data analysis deals with the analysis and theory of data that are in the form of functions. The atom of functional data is a function, a curve or an image instead of a scalar or multivariate vector. In this paper, the population curve $\{X(t):t\in I\}$ is considered as a square integrable stochastic process on a closed interval I. We observe $X(t_j)$ on a dense and regular grid $\{t_j\in I\}$.

Motivated by a range of applications, functional data becomes more and more common in supervised learning. Researchers are increasingly focusing on relating functional variables to other variables of interest, that is, the regression model. In particular, the functional linear regression model with scalar response in which a functional random variable is used to predict a real random variable has attracted considerable attention. In the problem of functional linear regression (FLR) we observe data

$$\big\{(X_1(t),Y_1),\cdots,(X_N(t),Y_N),\, t\in I\big\},$$

where the regressors $X_i(t)$'s are independent realizations of a random function $X(t)$, and the regression scheme for responses Y_i's are modeled by

$$Y_i=\alpha+\int_I X_i(t)\beta(t)dt+\epsilon_i,\quad 1\le i\le N. \tag{1.1}$$

Here, α is a constant, denoting the intercept in the model, usually assumed or centered to be zero. The $\beta(t)$ is a true slope function on I. The ϵ_i's are independent distributed with zero mean, independent of $X_i(t)$'s.

Several procedures have been proposed to estimate the parameters of the model, the functional principal component regression (FPCR) is currently the most popular method used, see Hall and Hosseini-Nasab [11], Hall and Horowitz [10], Reiss and Ogden [18]. To estimate the slope function, the standard FPCR method enables to regress the response on the principal component scores linked with the largest eigenvalues of the functional predictor covariance operator. For more details of FPCR, see Section 2.

Model-assisted survey sampling in terms of functional data has been scarcely investigated in contrast to finite dimension regression analysis. A known documental record can be found in Cardot *et al.* [7], which studies the mean curve estimation with auxiliary information from the large populations. In this paper, we consider the problem of the estimate of population means of a finite population, which is a realization of an infinite superpopulation defined with the functional linear regression model (1.1). A design-unbiased model-assisted estimator of the population mean is proposed based on the generalized difference estimator,

which is called FLR-assisted estimator. More specifically, for functional auxiliary data $X(t)$, model (1.1) is assumed to relate the auxiliary data $X(t)$ to the variable of interest Y, where $\beta(t)$ is an unknown coefficient function. We derive a generalized regression estimator of $\mu_Y := E(Y)$ based on the usual Horvitz-Thompson estimator and a corrective term that exploits the auxiliary functional data through the functional linear regression model. An estimate $\hat{\beta}(t)$ of the slope function $\beta(t)$ is obtained through functional principal component regression (FPCR) and then the model-assisted estimator is obtained by a plug-in method. The proposed FLR-assisted estimator does not require a strong assumption for the response, which enjoys robust properties in both theoretical and empirical performance.

The remainder of this paper is organized as follows. Section 2 introduces the estimation of the FLR model based on FPCR. The proposed FLR-assisted estimator and its asymptotic properties are given in Section 3. Moreover, the simulation studies and real data analysis are shown in Section 4 and Section 5 respectively, where we make a comparison of the average biase and average mean squared error under different sampling designs for the Horvitz-Thompson estimator and FLR-assisted estimator. Finally, Section 6 concludes.

2 Brief description of FPCR

In this section, we present how to estimate the parameters of the functional linear regression model based on the method of functional principal component regression (FPCR). Assume that random function $\{X(t): t \in I\}$ in model (1.1) has mean function $E[X(t)] = \mu_X(t)$ and covariance function $cov(X(t), X(s)) = \Sigma(t,s), s,t \in I$. Suppose the covariance function is positive definite, in which case it admits essentially a spectral decomposition, that is,

$$\Sigma(t,s) = \sum_{k=1}^{\infty} \lambda_k \psi_k(t) \psi_k(s), \quad s,t \in I, \tag{2.1}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are nonincreasing nonnegative eigenvalues and $\{\psi_k(t)\}_{k=1}^{\infty}$ are the corresponding eigenfunctions. (2.1) is also named as Mercer's theorem, which is analogous to singular-value decomposition of a square matrix.

The eigenfunctions $\{\psi_k(t)\}_{k=1}^{\infty}$ form an orthonormal basis for the space of all square-integrable functions on I, denote this space by $L^2(I)$ (or L^2 for simplicity). The space L^2 is indeed a Hilbert space, see [12, Section 2] and [14, Section 10]. What render the space L^2 so handy in FDA is that the inner product of two functions is defined by

$$\langle f, g \rangle = \int_I f(t) g(t) dt$$

and the corresponding norm is given by

$$\|f\| = \sqrt{\langle f, f \rangle}.$$

Consequently, by the Karhunen-Loève theorem (decomposition), the random function $X(t)$ can be expressed as a linear combination of the eigenfunctions, that is,

$$X(t) = \mu_X(t) + \sum_{k=1}^{\infty} \xi_k \psi_k(t), \tag{2.2}$$

where the functional principal component scores $\xi_k = \langle X - \mu_X, \psi_k \rangle, (k=1,2,\ldots)$ are uncorrelated random variables with mean zero and variance λ_k: i.e.,

$$E\xi_k = 0, \quad E\xi_k^2 = \lambda_k, \quad \text{cov}(\xi_j, \xi_k) = 0 \quad \text{for} \quad j \neq k.$$

Furthermore, we have the accumulated variance over $I : E\|X - \mu_X\|^2 = \sum_{k=1}^{\infty} \lambda_k$, which is viewed as the sum of the variances of random function $X(t)$ in the principal directions ψ_k defined by the Karhunen-Loève decomposition (2.2). Assume the coefficient function $\beta(t) \in L^2$ and can be expressed as

$$\beta(t) = \sum_{k=1}^{\infty} b_k \psi_k(t),$$

where

$$b_k = \langle \beta, \psi_k \rangle \quad \text{for} \quad k = 1,2,\ldots.$$

Under the above representation, the functional linear model (1.1) can be rewritten as

$$Y - \mu_Y \doteq Y^* = \sum_{k=1}^{\infty} b_k \xi_k + \epsilon,$$

where the scalar response Y^* is represented by an infinite linear combination of $\xi_1, \xi_2, \cdots$. As $\beta(t)$ is square integrable, which implies $\sum_{k=1}^{\infty} b_k^2 < \infty$. Then the coefficients b_k can be given by

$$b_k = \frac{Cov(\xi_k, Y^*)}{\lambda_k}.$$

Now we consider the estimation of slope function based on the finite population $\{(X_i(t), Y_i)\}_{i=1}^N$. Empirical versions of covariance function under finite population-level and its spectral decomposition are

$$\tilde{\Sigma}(t,s) = \frac{1}{N} \sum_{i=1}^{N} \{X_i(t) - \tilde{\mu}_X(t)\} \{X_i(s) - \tilde{\mu}_X(s)\} = \sum_{j=1}^{\infty} \tilde{\lambda}_j \tilde{\psi}_j(t) \tilde{\psi}_j(s), \tag{2.3}$$

where $s,t\in I$, $\tilde{\mu}_X(t)=N^{-1}\sum_{i=1}^{N}X_i(t)$. Analogously to $\Sigma(t,s)$, $\tilde{\lambda}_1\geq\tilde{\lambda}_2\geq\cdots$ and $\tilde{\psi}_k(s),k=1,2,\ldots$ are eigenvalues and corresponding eigenfunctions of $\tilde{\Sigma}(s,t)$. Then the estimate of $\beta(t)$ under population-level can be obtained by

$$\tilde{\beta}(t)=\sum_{k=1}^{r_N}\tilde{b}_k\tilde{\psi}_k(t) \tag{2.4}$$

with

$$\tilde{b}_k=\frac{1}{\tilde{\lambda}_k N}\sum_{i=1}^{N}\tilde{\xi}_{ik}(Y_i-\tilde{\mu}_Y),$$

where

$$\tilde{\xi}_{ik}=\int_I\big(X_i(t)-\tilde{\mu}_X(t)\big)\tilde{\psi}(t)dt,\quad \tilde{\mu}_X(t)=N^{-1}\sum_{i=1}^{N}X_i(t),\quad \tilde{\mu}_Y=N^{-1}\sum_{i=1}^{N}Y_i.$$

Here, r_N is the truncation parameter which can be selected by cross-validation or some other criterion such as AIC in [14, Section 6.2].

3 The proposed estimator and its properties

Let $(\Omega,\mathcal{A},P)$ be a probability space where Ω is a sample space, $\mathcal{A}$ is a σ-algebra, and P is a probability measure. For $N\geq 1$, consider a finite population $U_N=\{1,\ldots,N\}$ with N elements as a full sample defined on $(\Omega,\mathcal{A},P)$, and the associated functional auxiliary variables $\{X_i(t),t\in I\}_{i=1}^{N}$ also defined on $(\Omega,\mathcal{A},P)$.

A sample S of size n is drawn from U_N according to sampling design $p_N(S)$, where $p_N(S)$ is the probability of drawing the sample S on the set of 2^N subsets of U_N. We can treat $p_N(S)$ as the probability of selecting a specific sample S. Let $I_k=\mathrm{I}(k\in S)$ be the sample membership indicator which is Bernoulli distributed for $k\in U_N$. The sample membership indicator is the main source of randomness in the derivation of the asymptotical properties, instead of the error terms in the model. The first-order inclusion probability is defined by

$$\pi_k=P\{k\in S\}=EI_k=\sum_{S\in U_N:k\in S}p_N(S).$$

Similarly, denote the second-order inclusion probability by

$$\pi_{kl}=P\{k,l\in S\}=E[I_kI_l]=\sum_{S\in U_N:k,l\in S}p_N(S)$$

for all $i,j\in U_N$. Here π_k,π_{kl} are supposed to be positive.

We know that the whole information of auxiliary functional data $X_i(t)$ can be obtained for each $i \in U_N$. In practice, they may be obtained as discrete realizations. By model-assisted estimators with the aids of functional principal components analysis, Cardot *et al.* [8] considered the estimation of population total curve from samples $\{Y_k(t)\}_{k=1}^N$ for $t \in [0,1]$, they briefly mentioned the following model-assisted estimators which takes the advantages of auxiliary covariate information

$$\hat{Y}_N(t) = \sum_{k \in S} \frac{Y_k(t) - \hat{Y}_k(t)}{\pi_k} + \sum_{k \in U_N} \hat{Y}_k(t), \quad t \in [0,1],$$

where $\hat{Y}_k(t)$ is predicted by some function-on-scalar regressions. The significant difference from Cardot *et al.* [8] is that here we deal with the Y_i defined by the scalar-on-function regression model with the form (1.1).

Our interest is the estimation of the population mean $\bar{Y}_N = \frac{1}{N}\sum_{i \in U_N} Y_i$, where Y_i is the i-th sample drawn from population variable Y. Observe that the information of underlying variable Y_i can be known only for $i \in S$ while the information of auxiliary functional data $X_i(t)$ can be obtained for all $i \in U_N$. If no auxiliary information other than the inclusion probabilities is obtained, a well-known design-unbiased estimator is the Horvitz-Thompson estimator (H-T estimator in short, [21]) via inverse-probability weighting

$$\hat{\bar{Y}}_N^{HT} = \frac{1}{N}\sum_{i \in S} \frac{Y_i}{\pi_i} \tag{3.1}$$

with the design variance

$$Var\big(\hat{\bar{Y}}_N^{HT} | \mathcal{F}_{Y^N}\big) = \frac{1}{N^2} \sum_{i,j \in U_N} (\pi_{ij} - \pi_i \pi_j) \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j},$$

where $\mathcal{F}_{Y^N} := \sigma(\{Y_i\}_{i=1}^N)$.

If π_{kl} are positive for $k,l \in U_N$, then we call the sampling design is measurable, see Särndal *et al.* [20]. Therefore, the unbiased estimator of design variance is given by

$$\hat{Var}\big(\hat{\bar{Y}}_N^{HT} | \mathcal{F}_{Y^N}\big) = \frac{1}{N^2} \sum_{i,j \in U_N} (\pi_{ij} - \pi_i \pi_j) \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \frac{I_i I_j}{\pi_{ij}}, \tag{3.2}$$

where $I_i = \mathrm{I}(i \in S)$ and I_j is similarly defined, see Breidt and Opsomer [6] for more details.

Despite the fact that $\hat{Y}_N^{HT}$ is an appealing and commonly used estimator in survey sampling since it is unbiased, functional covariates $\{X_1(t), \cdots, X_n(t)\}$ as

auxiliary information is available to do adjustments in order to shrink variance. An improved estimator is obtained by the following difference estimator (3.3):

$$\begin{aligned}\tilde{\tilde{Y}}_{diff} &= \frac{1}{N}\sum_{i\in S}\frac{Y_i}{\pi_i} + \int_I \left(\frac{1}{N}\sum_{i\in U_N} X_i(t) - \frac{1}{N}\sum_{i\in S}\frac{X_i(t)}{\pi_i}\right)\beta(t)dt \\ &= \frac{1}{N}\left(\sum_{i\in S}\frac{1}{\pi_i}\left(Y_i - \int_I X_i(t)\beta(t)dt\right) + \sum_{i\in U_N}\left(\int_I X_i(t)\beta(t)dt\right)\right), \end{aligned} \tag{3.3}$$

where $\beta(t)$ is the true slope function in model (1.1). In the first line of Eq. (3.3) the difference

$$\frac{1}{N}\sum_{i\in S}\frac{X_i(t)}{\pi_i} - \frac{1}{N}\sum_{i\in U_N} X_i(t)$$

characterizes the oscillation of the functional covariates in the subsample with regard to the full sample, and the auxiliary slope $\beta(t)$ fits the functional linear relationships between the covariates and responses. Note that

$$\frac{1}{N}\sum_{i\in U_N}\left(\int_I X_i(t)\beta(t)dt\right)$$

is not random given the full sample $\mathcal{F}_N := \sigma(\{Y_i, X_i(t), t\in I\}_{i=1}^N)$. Then, the design variance of the difference estimator is

$$\begin{aligned} Var\left(\tilde{\tilde{Y}}_{diff}|\mathcal{F}_N\right) = \frac{1}{N^2}\sum_{i,j\in U_N}(\pi_{ij}-\pi_i\pi_j)\frac{1}{\pi_i}\left(Y_i - \int_I X_i(t)\beta(t)dt\right) \\ \times\frac{1}{\pi_j}\left(Y_i - \int_I X_i(t)\beta(t)dt\right) \end{aligned}$$

by using expression in the second line of Eq. (3.3) and the variance of Horvitz-Thompson estimator.

While in survey sampling studies, the auxiliary variables can be obtained from the finite population, which is not the same situation for the study variables due to some practical difficulties. Only a sample from the finite population can be obtained under some predefined sampling design. Denote the empirical versions of covariance function under sample-level $\hat{\Sigma}(t,s)$, which is the same as the population-level covariance function $\tilde{\Sigma}(t,s)$, as well as the eigenvalues $\tilde{\lambda}_k$ and eigenfunctions $\tilde{\psi}_k(t)$. Then the estimate of $\beta(t)$ under sample-level can be obtained by

$$\hat{\beta}(t) = \sum_{k=1}^{r_n}\hat{b}_k\tilde{\psi}_k(t) \tag{3.4}$$

with

$$\hat{b}_k=\frac{1}{\tilde{\lambda}_k n}\sum_{i=1}^{n}\tilde{\xi}_{ik}(Y_i-\hat{\mu}_Y),$$

where $\tilde{\xi}_{ik}$,$\tilde{\lambda}_k$ are the same as the estimate under population-level, while $\hat{\mu}_Y=n^{-1}\sum_{i=1}^{n}Y_i$, r_n is the truncation parameter under sample-level, which can be selected by cross-validation or AIC criterion.

After plugging $\hat{\beta}(t)$ into the difference estimator (3.3), then the FLR-assisted estimator for the population means is defined as follow:

$$\tilde{\hat{Y}}_N=\frac{1}{N}\left(\sum_{i\in S}\frac{1}{\pi_i}\left(Y_i-\int_I X_i(t)\hat{\beta}(t)dt\right)+\sum_{i\in U_N}\left(\int_I X_i(t)\hat{\beta}(t)dt\right)\right), \tag{3.5}$$

where $\hat{\beta}(t)$ is computed by (3.4). The corresponding estimate of the variance of the FLR-assisted estimator is given by

$$\begin{aligned}\hat{Var}(\tilde{\hat{Y}}_N|\mathcal{F}_N)=&\frac{1}{N^2}\sum_{i,j\in U_N}(\pi_{ij}-\pi_i\pi_j)\frac{1}{\pi_i}\left(Y_i-\int_I X_i(t)\hat{\beta}(t)dt\right)\\&\times\frac{1}{\pi_j}\left(Y_i-\int_I X_i(t)\hat{\beta}(t)dt\right)\frac{I_iI_j}{\pi_{ij}},\end{aligned}$$

where $I_i=\mathrm{I}(i\in S)$ and I_j is similarly defined.

In the following, we provide theoretical guarantees that our proposed FLR-assisted estimator is asymptotically design unbiased and design consistent, and this inherits the properties of simple mean estimator. In asymptotic analysis, we allow $N\to\infty$ and so we need to define the σ-algebra generated by population and functional auxiliary variables as $\mathcal{F}_N:=\sigma(\{Y_i,X_i(t),t\in I\}_{i=1}^N)$. Then the asymptotical results are conditioning on $\mathcal{F}_N$ as $N\to\infty$ since we prefer statement in terms of random $\{Y_i\}_{i=1}^\infty$. This notation is analogous to the "with ξ-probability one" in Robinson and Särndal [19], and "consistency in conditional probability with probability approaching one" in [2, Theorem 1].

Conditioning on $\mathcal{F}_N$, we propose the required regularity assumptions, which are reasonable in previous references and real data analysis.

(A1) $\limsup_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}Y_i^2<\infty.$

(A2) $\limsup_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}\left(\int_I X_i^2(s)ds\right)\left(\int_I\hat{\beta}^2(s)ds\right)<\infty.$

(A3) $\liminf_{N\to\infty}N\min_{i\in U_N}\pi_i=\infty,\quad\limsup_{N\to\infty}\max_{i,j\in U_N,i\neq j}\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right|=0.$

(A1) is a classical hypothesis that concerns the second-order population moment, and this condition is very mild unless the heavy tail case. (A2) is also a common assumption by our L^2 space framework. (A3) is a regularity condition that is akin to Robinson and Särndal [19] for the parametric regression case. The first condition in (A3) implies that $n\to\infty$ as $N\to\infty$, we do not require that n increase as fast as N. And the second condition in (A3) signifies that we only allow sufficiently small dependence between different sample membership indicators.

Theorem 3.1 (Design consistency)**.** *Under assumptions (A1)-(A3), the FLR-assisted estimator* (3.5) *is asymptotically design unbiased in the sense that $\hat{\bar{Y}}_N$ is consistent to $\bar{Y}_N$ in conditional probability*

$$\lim_{N\to\infty} E\left[\left|\hat{\bar{Y}}_N-\bar{Y}_N\right||\mathcal{F}_N\right]=0 \quad \textit{with probability approaching one,} \tag{3.6}$$

which means that

$$\lim_{N\to\infty} P\left(E\left[\eta\left|\hat{\bar{Y}}_N-\bar{Y}_N\eta\right||\mathcal{F}_N\right]<\varepsilon\right)=1 \quad \textit{for any} \quad \varepsilon>0.$$

Moreover, it is design consistent in the sense that

$$\lim_{N\to\infty} P\left(\left|\hat{\bar{Y}}_N-\bar{Y}_N\right|>\eta|\mathcal{F}_N\right)=0 \quad \textit{with probability approaching one for all } \eta>0.$$

Proof. By Markov's inequality of conditional expectation, we have

$$E\left[I\left(\left|\hat{\bar{Y}}_N-\bar{Y}_N\right|>\eta|\mathcal{F}_N\right)\right]=P\left\{\left|\hat{\bar{Y}}_N-\bar{Y}_N\right|>\eta|\mathcal{F}_N\right\}\le\frac{1}{\eta}E\left[\left|\hat{\bar{Y}}_N-\bar{Y}_N\right|\mathcal{F}_N\right].$$

So it suffices to show that

$$\lim_{N\to\infty} E\left[\left|\hat{\bar{Y}}_N-\bar{Y}_N\right||\mathcal{F}_N\right]=0$$

with probability approaching 1.

Denote

$$a_N=\frac{1}{N}\sum_{i\in U}Y_i\left(\frac{I_i}{\pi_i}-1\right),\quad b_N=\frac{1}{N}\sum_{i\in U}\left(\int_I X_i(t)\hat{\beta}(t)dt\right)\left(\frac{I_i}{\pi_i}-1\right).$$

Then, conditioning on $\mathcal{F}_N$,

$$\begin{aligned}E\left(\left|\hat{\bar{Y}}_N-\bar{Y}_N\right||\mathcal{F}_N\right)&\le E\left(|a_N||\mathcal{F}_N\right)+E\left(|b_N||\mathcal{F}_N\right)\\&\le\left[E\left(a_N^2|\mathcal{F}_N\right)\right]^{\frac{1}{2}}+\left[E\left(a_N^2|\mathcal{F}_N\right)\right]^{\frac{1}{2}},\end{aligned}$$

where the last inequality stems from Jensen's inequality for conditional expectation.

Notice that

$$\begin{aligned}
E\left(a_N^2|\mathcal{F}_N\right) &= \frac{1}{N^2}\sum_{i\in U_N} Y_i^2\left[\frac{1-\pi_i}{\pi_i}\right] + \frac{1}{N^2}\sum_{i,j\in U_N, i\neq j} Y_iY_j\left[\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right] \\
&\leq \frac{1}{N\min\limits_{i\in U_N}\pi_i}\left(\frac{1}{N}\sum_{i\in U_N} Y_i^2\right) + \max_{i,j\in U_N, i\neq j}\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right|\frac{1}{N^2}\sum_{i,j\in U_N, i\neq j} Y_iY_j \\
&\leq \frac{1}{N\min\limits_{i\in U_N}\pi_i}\left(\frac{1}{N}\sum_{i\in U_N} Y_i^2\right) + \max_{i,j\in U_N, i\neq j}\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right|\left(\frac{1}{N}\sum_{i\in U_N}|Y_i|\right)^2 \\
&\leq \frac{1}{N\min\limits_{i\in U_N}\pi_i}\left(\frac{1}{N}\sum_{i\in U_N} Y_i^2\right) + \max_{i,j\in U_N, i\neq j}\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right|\frac{1}{N}\sum_{i\in U_N} Y_i^2,
\end{aligned}$$

and by independence of $\{I_i\}$ and $\mathcal{F}_N$

$$\begin{aligned}
E\left(b_N^2|\mathcal{F}_N\right) &= \frac{1}{N^2}E\left(\left[\sum_{i\in U}\int_I X_i(s)\hat{\beta}(s)ds\left(\frac{I_i}{\pi_i}-1\right)\right]^2\middle|\mathcal{F}_N\right) \\
&\leq \frac{1}{N^2}E\left(\left[\sum_{i\in U}\left(\int_I X_i^2(s)ds\right)^{\frac{1}{2}}\left(\int_I\hat{\beta}^2(s)ds\right)^{\frac{1}{2}}\left(\frac{I_i}{\pi_i}-1\right)\right]^2\middle|\mathcal{F}_N\right) \\
&= \frac{1}{N^2}E\left(\sum_{i\in U_N}\left(\int_I X_i^2(s)ds\right)\left(\int_I\hat{\beta}^2(s)ds\right)\left[\frac{1-\pi_i}{\pi_i}\right]\middle|\mathcal{F}_N\right) \\
&\quad + \frac{1}{N^2}E\left(\frac{1}{N^2}\sum_{i,j\in U_N, i\neq j}\left(\int_I X_i^2(s)ds\right)^{\frac{1}{2}}\left(\int_I X_j^2(s)ds\right)^{\frac{1}{2}}\right. \\
&\qquad \left.\times\left(\int_I\hat{\beta}^2(s)ds\right)\left[\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right]\middle|\mathcal{F}_N\right) \\
&\leq \left(\frac{1}{N\min\limits_{i\in U_N}\pi_i} + \max_{i,j\in U_N, i\neq j}\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right|\right) \\
&\quad \times\left(\frac{1}{N}\sum_{i\in U_N}\left(\int_I X_i^2(s)ds\right)\left(\int_I\hat{\beta}^2(s)ds\right)\right).
\end{aligned}$$

By (A1) and (A3), conditioning on $\mathcal{F}_N$, it gives $(E(a_N^2))^{1/2}\to 0$ as $N\to\infty$. Similarly, applying (A1)-(A3), as $N\to\infty$, we have $(E(b_N^2))^{1/2}\to 0$.

As $N\to\infty$, $\mathcal{F}_N$ is not random by the fact that an infinite population usually has elements that consist of all the outcomes. So that we can draw the conclusion in (3.6) with probability approaching one and complete the proof. □

4 Simulation study

In this section, we compare the efficiency of our proposed FLR-assisted estimators to the H-T estimators without considering the auxiliary information.

Following the simulation setting in Hall and Horowitz [10], we consider the infinite superpopulation defined as the functional linear regression model with the form

$$Y_i=\int_0^1\beta_0(t)X_i(t)dt+\epsilon_i,$$

where

$$\beta_0(t)=\sum_{j=1}^{50}\beta_{0j}\phi_j(t),\quad \beta_{01}=0.3,\quad \beta_{0j}=4(-1)^{j+1}j^{-2},\quad j\geq 2,$$

$$\phi_j(t)=\sqrt{2}\cos(j\pi t),\quad X_i(t)=\sum_{j=1}^{50}\gamma_j Z_{ij}\phi_j(t),$$

$$\gamma_j=(-1)^{j+1}j^{-\frac{a}{2}},\quad a\in\{1.1,2\},\quad Z_{ij}\sim U\big[-\sqrt{3},\sqrt{3}\big].$$

For the error term ϵ_i, we consider four cases $\epsilon\sim N(0,1)$, $N(0,0.25)$, $t(3)$, $t(5)$. $i\in U_N$ with U_N the finite population, a realization of the infinite superpopulation. Here $t\in I=[0,1]$ and $T=100$ equally spaced time points $\{t_{ij}\in[0,1], i\in U_N; j=1,\ldots,100\}$ are used in our simulation study.

The finite population is of size $N=1000$, samples are generated with sample size $n=100,200,400$ under a fixed sampling design. The following two sampling designs are considered:

- Simple Random Sampling (SRS): the inclusion probability of $i\in U_N$ defined by $\pi_i=n/N$, which satisfies $\sum_{i=1}^N\pi_i=n$.
- Unequal Probability Sampling (UPS): the inclusion probability of $i\in U_N$ defined by $\pi_i=n*c_i/\sum_{i=1}^N c_i$, $c_i\sim U[0,1]$ which satisfies $\sum_{i=1}^N\pi_i=n$.

The samples are random arranged in every sampling process. For the four error cases, $n_{sample}=100$ replicated samples are selected from the same population, then the estimators can be calculated, and the design bias and the design mean squared error of the estimators can be computed empirically. The simulations were replicated $n_{simu}=500$ times to obtain the average bias (ABIAS) and average mean squared error (AMSE) defined as follow:

$$ABIAS(\hat{\bar{Y}})=\frac{1}{n_{simu}n_{sample}}\sum_{i=1}^{n_{simu}}\sum_{j=1}^{n_{sample}}\left|\hat{\bar{Y}}^{ij}-\bar{Y}^{ij}\right|,$$

$$AMSE(\hat{\bar{Y}})=\frac{1}{n_{simu}n_{sample}}\sum_{i=1}^{n_{simu}}\sum_{j=1}^{n_{sample}}\left(\hat{\bar{Y}}^{ij}-\bar{Y}^{ij}\right)^2,$$

where $\hat{\bar{Y}}^{ij}$ and $\bar{Y}^{ij}$ are the estimated and true value of the i-th simulation under the jth sampling.

Tables 1 and 2 show the result of the above simulation under simple random sampling and unequal probability sampling, respectively. Obviously, the ABIAS and AMSE in FLR-assisted estimators is substantially smaller than those in H-T estimators under both sampling designs. The results imply that the proposed FLR-assisted estimators have a better performance than the H-T estimators under different error assumptions.

5 The real-world data study

5.1 Gasoline data

The gasoline data was provided by Kalivas [13], which is available in the R package "refund". The data consists of near-infrared reflectance spectra and octane numbers of 60 gasoline samples. It was also analyzed by Reiss and Ogden [18]. The response variable "octane" is a numeric vector of octane numbers for the 60 samples. The predictor variable "NIR" is a 60×401 matrix of NIR spectra, a discrete realization of the functional data. Each NIR spectrum consists of $\log(1/\text{reflectance})$ measurements at 401 wavelengths, in 2-nm intervals from 900 nm to 1700 nm.

For the gasoline data, the size of the finite population U_N is assumed to be $N=60$, and the simulation sample size is set to be $n=20$. Our interest is the population means of octane numbers. The sampling design was repeated 1000 times. Two estimators of the population means are computed under simple random sampling and unequal probability sampling. Fig. 1 presents the boxplots of

Table 1: Comparison of H-T estimators and FLR-assisted estimators under SRS.

n	Error	a	H-T		FLR-assisted	
			ABIAS	AMSE	ABIAS	AMSE
100	N(0,1)	1.1	0.0981	0.0152	0.0843	0.0113
		2	0.0895	0.0126	0.0578	0.0053
	N(0,0.5)	1.1	0.0650	0.0067	0.0649	0.0066
		2	0.0516	0.0042	0.0463	0.0034
	t(3)	1.1	0.1415	0.0318	0.1077	0.0189
		2	0.1392	0.0309	0.0701	0.0079
	t(5)	1.1	0.1155	0.0210	0.0938	0.0141
		2	0.1079	0.0183	0.0592	0.0057
200	N(0,1)	1.1	0.0655	0.0067	0.0551	0.0048
		2	0.0599	0.0056	0.0411	0.0028
	N(0,0.5)	1.1	0.0434	0.0029	0.0463	0.0033
		2	0.0345	0.0019	0.0339	0.0018
	t(3)	1.1	0.0971	0.0151	0.0730	0.0084
		2	0.0897	0.0128	0.0501	0.0040
	t(5)	1.1	0.0766	0.0092	0.0625	0.0062
		2	0.0725	0.0083	0.0449	0.0033
400	N(0,1)	1.1	0.0398	0.0025	0.0443	0.0030
		2	0.0369	0.0021	0.0337	0.0017
	N(0,0.5)	1.1	0.0263	0.0011	0.0338	0.0017
		2	0.0209	0.0007	0.0252	0.0010
	t(3)	1.1	0.0591	0.0055	0.0552	0.0047
		2	0.0570	0.0051	0.0467	0.0034
	t(5)	1.1	0.0467	0.0034	0.0327	0.0014
		2	0.0437	0.0030	0.0397	0.0025

FLR-assisted estimators and H-T estimators of population mean for 1000 replications of sampling process under two sampling designs, which shows the raising efficiency of our proposed FLR-assisted estimators, especially in the unequal probability sampling case.

5.2 Tecator data

Now we consider the tecator data, which comes from a quality control problem in the food industry and can be found at `http://lib.stat.cmu.edu/datasets/`

Table 2: Comparison of H-T estimators and FLR-assisted estimators under UPS.

n	Error	a	H-T		FLR-assisted	
			ABIAS	AMSE	ABIAS	AMSE
100	N(0,1)	1.1	0.1547	0.0519	0.1172	0.0254
		2	0.1408	0.0515	0.1056	0.0343
	N(0,0.5)	1.1	0.1087	0.0789	0.0697	0.0079
		2	0.0808	0.0170	0.0537	0.0052
	t(3)	1.1	0.2212	0.1808	0.1715	0.1072
		2	0.1988	0.0861	0.1502	0.0574
	t(5)	1.1	0.1785	0.0709	0.1393	0.0523
		2	0.1682	0.0678	0.1260	0.0424
200	N(0,1)	1.1	0.1153	0.0341	0.0877	0.0164
		2	0.1051	0.0295	0.0811	0.0170
	N(0,0.5)	1.1	0.0779	0.0141	0.0507	0.0042
		2	0.0620	0.0114	0.0397	0.0033
	t(3)	1.1	0.1557	0.0558	0.1253	0.0344
		2	0.1514	0.0506	0.1186	0.0345
	t(5)	1.1	0.1362	0.0488	0.1063	0.0291
		2	0.1247	0.0335	0.1005	0.0235
400	N(0,1)	1.1	0.0912	0.0675	0.0726	0.0165
		2	0.0800	0.0161	0.0690	0.0107
	N(0,0.5)	1.1	0.0582	0.0080	0.0361	0.0022
		2	0.0467	0.0055	0.0294	0.0016
	t(3)	1.1	0.1270	0.0691	0.1115	0.0552
		2	0.1210	0.0710	0.1098	0.0626
	t(5)	1.1	0.1032	0.0265	0.0905	0.0250
		2	0.0966	0.0237	0.0818	0.0181

`tecator`. It was first studied by Borggaard and Thodberg [4], who used a neural networks approach. This dataset concerns a sample of finely chopped meat, which consists of near-infrared reflectance spectra of 240 samples of ground pork. It is of interest to predict the wet-chemistry measurements using the corresponding NIR spectra on the fat, water, and protein contents. The NIR spectra are recorded on a Tecator Infrared spectrometer that measures the absorbance at 100 wavelengths in the region 850-1050 nm. We use the first 215 samples as suggested by Borggaard and Thodberg [4].

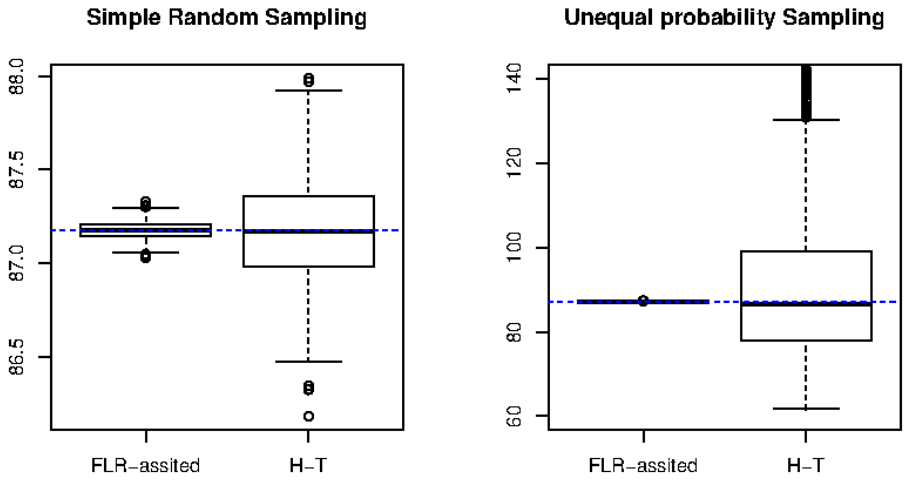

Figure 1: Boxplots of FLR-assisted estimators and H-T estimators for gasoline data: the blue dotted line is the true value of population means.

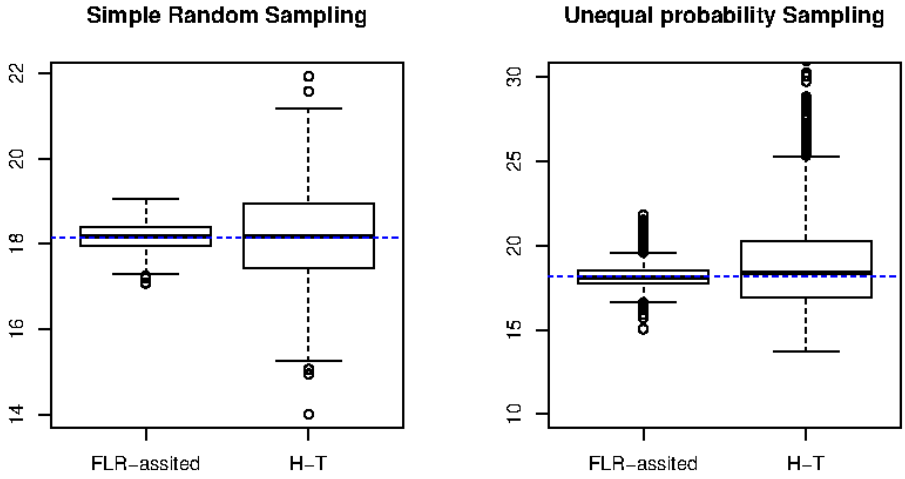

Figure 2: Boxplots of FLR-assisted estimators and H-T estimators for tecator data: the blue dotted line is the true value of population means.

For this data, the size of the finite population is assumed to be $N = 215$, and the simulation sample size is $n = 80$. The sampling design was repeated 1000 times. Our interest is the population means of wet-chemistry measurements. Two estimators of the population means are computed. Fig. 2 shows the boxplots of FLR-assisted estimators and H-T estimators of the population mean for 1000 replications of the sampling process under two sampling designs, implying that our proposed FLR-assisted estimators are more efficient than H-T estimators.

6 Conclusion

This paper investigates the model-assisted estimation of finite population means when the functional linear regression model is assumed to be the infinite super-population model. The main goal of adopting model-assisted approach is to reduce the variance of population means. A FLR-assisted estimator of population means was proposed with some well-defined properties. Simulation results show that our proposed FLR-assisted estimators are more efficient than the traditional H-T estimators, which implies that the auxiliary functional data can indeed help improve the estimation accuracy.

In the future, it is challenging to add conditions that ensure asymptotic distribution of the proposed model-assisted estimator. The obtained asymptotic variance can be applied to construct a conservative confidence interval for the FLR-assisted estimator. It is also interesting to develop other model-assisted estimators from some complex functional data models in the future study, such as functional single-index model, functional additive model and partial linear functional regression model with RKHS framework; see Lei and Zhang [15] and references therein.

Acknowledgments

Chao Liu was supported in part by China Postdoctoral Science Foundation (Grant Nos. 2021M691443, 2021TQ0141) and SUSTC Presidential Postdoctoral Fellowship. Huiming Zhang was supported in part by the University of Macau under UM Macao Talent Programme (UMMTP-2020-01).

References

[1] C. Aaron, A. Cholaquidis, R. Fraiman, and B. Ghattas, *Multivariate and functional robust fusion methods for structured Big Data*, J. Multivar. Anal. 170 (2019), 149–161.

[2] M. Ai, J. Yu, H. Zhang, and H. Wang, *Optimal subsampling algorithms for big data regressions*, Stat. Sin. 31(2) (2021), 749–772.

[3] A. Bloniarz, H. Liu, C. H. Zhang, J. S. Sekhon, and B. Yu, *Lasso adjustments of treatment effect estimates in randomized experiments*, Proc. Natl. Acad. Sci. USA 113(27) (2016), 7383–7390.

[4] C. Borggaard and H. H. Thodberg, *Optimal minimal neural interpretation of spectra*, Anal. Chem. 64(5) (1992), 545–551.

[5] F. J. Breidt and J. D. Opsomer, *Local polynomial regression estimators in survey sampling*, Ann. Stat. 28(4) (2000), 1026–1053.
[6] F. J. Breidt, and J. D. Opsomer, *Model-assisted survey estimation with modern prediction techniques*, Stat. Sci. 32(2) (2017), 190–205.
[7] H. Cardot, M. Chaouch, C. Goga, and C. Labruère, *Properties of design-based functional principal components analysis*, J. Stat. Plan. Inference, 140(1) (2010), 75–91.
[8] H. Cardot, C. Goga, and P. Lardin, *Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data*, Electron. J. Stat. 7 (2013), 562–596.
[9] A. Cuevas, M. Febrero, and R. Fraiman, *Linear functional regression: the case of fixed design and functional response*, Can. J. Stat. 30(2) (2002), 285–300.
[10] P. Hall and J. L. Horowitz, *Methodology and convergence rates for functional linear regression*, Ann. Stat. 35(1) (2007), 70–91.
[11] P. Hall and M. Hosseini-Nasab, *On properties of functional principal components analysis*, J. R. Stat. Soc. Series B Stat. Methodol. 68(1) (2006), 109–126.
[12] T. Hsing and R. Eubank, *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, John Wiley & Sons, 2015.
[13] J. H. Kalivas, *Two data sets of near infrared spectra*, Chemometr. Intell. Lab. Syst. 37(2) (1997), 255–259.
[14] P. Kokoszka and M. Reimherr, *Introduction to Functional Data Analysis*, CRC Press, 2017.
[15] X. Lei and H. Zhang, *Non-asymptotic optimal prediction error for RKHS-based partially functional linear models*, arXiv:2009.04729, 2020.
[16] C. Liu, C. Wei, and Y. Su, *Geographically weighted regression model-assisted estimation in survey sampling*, J. Nonparametr. Stat. 30(4) (2018), 906–925.
[17] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer, 2005.
[18] P. T. Reiss and R. T. Ogden, *Functional principal component regression and functional partial least squares*, J. Am. Stat. Assoc. 102(479) (2007), 984–996.
[19] P. M. Robinson and C. E. Särndal, *Asymptotic properties of the generalized regression estimator in probability sampling*, Sankhyā: The Indian Journal of Statistics, Series B, (1983), 240–248.
[20] C. E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer Science & Business Media, 1992.
[21] M. Thompson, *Theory of Sample Surveys*, CRC Press, 1997.
[22] A. Zhang, L. D. Brown, and T. T. Cai, *Semi-supervised inference: General theory and estimation of means*, Ann. Stat. 47(5) (2019), 2538–2566.