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