Statistical Inference for a Finite Population Mean with Machine Learning-Based Imputation for Missing Survey Data

David Haziza and Mehdi Dagdoug[1]

Abstract

National statistical offices have increasingly adopted machine learning (ML) for its potential to improve survey estimates. ML techniques offer significant advantages, notably the ability to manage high-dimensional data and to capture complex, nonlinear relationships, thereby enhancing the overall quality of survey statistics. In this article, following the approach of Chernozhukov et al. (2018), we describe a double/debiased machine learning framework that enables valid statistical inference when imputed estimators are derived from ML procedures. Simulation results suggest that the proposed framework performs well in a wide range of scenarios.

Keywords: Cross-fitting; Debiased machine learning; Doubly robust estimation; Item nonresponse; Variance estimation

[1] 2Mehdi Dagdoug,McGill University,Montreal, Quebec; David Haziza, University of Ottawa, Ottawa, Ontario