University of California, Irvine
Statistics Seminar

Using Auxiliary Information in Probability Survey Data to Improve Pseudo-Weighting in Non-Probability Samples: A Copula Model Approach

Lan Xue
Professor of Statistics
Oregon State University

4 p.m., Thursday, June 1, 2023
6011 Donald Bren Hall

While probability sampling has been considered the gold standard of survey methods, nonprobability sampling such as online opt-in surveys is increasingly popular due to its convenience and low cost. However, nonprobability samples can lead to biased estimates due to the unknown nature of the underlying selection mechanism. In this work, we propose both parametric and semi-parametric approaches to integrate probability and nonprobability samples using common ancillary variables observed in both samples. In the parametric approach, the joint distribution of ancillary variables is assumed to follow the latent Gaussian copula model, which is flexible to accommodate both discrete and continuous variables. In contrast, the semi-parametric approach requires no assumptions about the distribution of ancillary variables. In addition, logistic regression is used to model the mechanism by which population units enter the nonprobability sample. The unknown parameters in the copula model are estimated through the pseudo maximum likelihood approach. The logistic regression model is estimated by maximizing the sample likelihood constructed from the nonprobability sample. The proposed method is evaluated in the context of estimating the population mean. Our simulation results show that the proposed method is able to correct the selection bias in the nonprobability sample by consistently estimating the underlying inclusion mechanism. By incorporating additional information in the nonprobability sample, the combined method can provide a more efficient estimation of the population mean than using the probability sample alone. A real data application is provided to illustrate the practical use of the proposed method.