NONPROBSVY – AN R PACKAGE FOR NON-PROBABILITY SAMPLES

Łukasz Chrostowski[1] and Maciej Beręsewicz[2]

[1] Poznan University of Economics and Business, Poland

[2] Poznan University of Economics and Business, Poland
e-mail: maciej.beresewicz@ue.poznan.pl

Abstract

The aim of the nonprobsvy R package is to perform statistical inference on non-probability survey samples (including big data) when auxiliary information from external sources, such as probability samples or population totals or means, is available.

It should be noted that there are several packages that allow correcting for selection bias in nonprobability samples, such as GJRM (Marra et al. 2017), NonProbEst (Rueda et al. 2020), or even sampling (Tillé and Matei 2021). However, these packages do not implement state-of-the-art approaches recently proposed in the literature: Chen et al. (2020), Yang et al. (2020), Wu (2022), nor do they use the survey package (Lumley 2004) for inference.

We have implemented propensity score weighting (e.g. with calibration constraints), mass imputation (e.g. predictive mean matching) and doubly robust estimators that take into account minimisation of the asymptotic bias of the population mean estimators, variable selection or overlap between random and non-random samples. The package uses the functionality of the survey package when a probability sample is available. During the presentation, the functionality of the package and examples will be presented.

The package is under development and can be found on https://github.com/ncn-foreigners/nonprobsvy/

Keywords: Data integration, Doubly robust estimation, Propensity score estimation.

References

Marra et al. (2017). A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. JASA, 112(518), 484-496.

Rueda et al. (2020). The R package NonProbEst for estimation in non-probability surveys. The R J, 12(1), 406-418.

Tillé and Matei (2021) sampling: Survey Sampling

Lumley (2004) Analysis of complex survey samples. JSS 9(1): 1-19

Chen et al. (2020). Doubly robust inference with nonprobability survey samples. JASA, 115(532), 2011-2021.

Yang et al. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. JRSS B, 82(2), 445.

Wu (2022). Statistical inference with non-probability survey samples. SM 48(2), 283-311.