USING ML TECHNIQUES FOR ESTIMATION WITH NON-PROBABILISTIC SURVEY DATA

Jorge Rueda[1], Maria del Mar Rueda [1], Ramón Ferri[1] and Beatriz Cobo[2]

[1] Department of Statistics and O.R., University of Granada, (e-mail: jorgerueda279@correo.ugr.es, mrueda@ugr.es, rferri@ugr.es)

[2] Department of Quantitative Methods for Economics and Business, University of Granada, (beacr@ugr.es)

ABSTRACT: Online surveys, despite their cost and effort advantages, are particularly prone to selection bias due to the differences between target population and potentially covered population. Some techniques have arisen in the last years regarding this issue. Propensity Score Adjustment, kernel weighting, Statistical Matching (or mass imputation), double robust estimation and superpopulation modeling are relevant techniques to mitigate selection bias. These techniques use the sample to train a model capturing the behaviour of a target variable which is to be estimated, or the propensity of the units to participate in the volunteer sample. The modeling step has been usually done with linear regression, but machine learning (ML) algorithms have been pointed out as promising alternatives. In this study we examine the use of these algorithms in the nonprobability survey context, in order to evaluate and compare their performance and adequacy to the problem.

KEYWORDS: survey sampling, non-probability samples, propensity score adjustment, machine learning.

1 Estimation in non probability surveys

The use of web surveys and big data sources for population inference is an active research field in social science and survey research. Such data sources allow to produce statistics cheaper, faster, and on a higher level of detail. However, these data most often lacks a sampling design, population coverage is incomplete and the data-generating mechanism is unknown. No valid inferences can be drawn and new methodologies are needed to evaluate the potential biases and make accurate estimates of the population parameters.

Different inference procedures are proposed in the literature to correct for selection bias induced by non-random selection mechanisms. There are three important approaches: the pseudo-design based inference (or pseudo-randomization),

statistical matching and predictive inference.

Pseudo-randomization and Statistical Matching require, apart from the nonprobability sample, a probability sample to do the adjustments. Propensity score adjustment (PSA) originally developed for balancing groups in nonrandomized clinical trials (Rosenbaum & Rubin, 1983) is the most used method for removing bias in nonprobability surveys (Lee & Valliant, 2009). Statistical Matching was firstly proposed in Rivers, 2007. The difference between both methods is the sample used in the estimators: PSA estimates the propensity of each individual of the nonprobability sample to participate in the survey and then this propensity is used to construct the weights of the estimators, while Statistical Matching adjusts a prediction model using data from the nonprobability sample, applies it in the probability sample to predict their values for the target variable y and uses them in the parametric estimators.

Superpopulation modelling requires data from the complete census of the target population for the covariates used in the adjustment, which is assumed to be a realization (sample) of a superpopulation where the (unknown) target values follow a model. The main idea is to fit a regression model on the target variable with data from the nonprobability sample, and use the model to predict the values of the target variable for each individual in the population. The prediction can be used for estimation using a model-based approach or some alternative versions such as model-assisted and model-calibrated.

Usually the linear regression model is considered for estimation, $E_m(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\beta$, and the predicted values of y_i in the probability sample (in the nonsampled individuals) are used for making estimators in the statistical matching inference (in the predictive inference). Logistic regression is usually used in PSA to predict the propensity (probability of the i-th individual of being included in the sample), $\pi_{vi} = P(I_{vi} = 1|\mathbf{x}_i)$.

Alternatively to the linear regression models, Machine Learning (ML) methods have been proposed for the estimation of the propensities and the nonsampled population values. In situations where additivity and/or linearity do not hold, ML algorithms are more suitable for regression and classification. Some of these algorithms, such as decision trees and related (Random Forests, Gradient Boosting Machines) can also take interactions into account without the need of specifying the terms. The use of some ML algorithms for non probability samples has been studied in the last few years (e.g. Buelens *et al.*, n.d.,

Ferri-García *et al.*, 2021, Castro-Martín *et al.*, 2021. In this work we consider some of the most important ML algorithms that can be used to define different estimators for a non-probability sample.

References

BUELENS, BART, BURGER, JOEP, & VAN DEN BRAKEL, JAN A. Comparing Inference Methods for Non-probability Samples. *International Statistical Review*, **86**(2), 322–343.

CASTRO-MARTÍN, LUIS, RUEDA, MARÍA DEL MAR, FERRI-GARCÍA, RAMÓN, & HERNANDO-TAMAYO, CÉSAR. 2021. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *Mathematics*, **9**(23).

FERRI-GARCÍA, RAMÓN, CASTRO-MARTÍN, LUIS, & DEL MAR RUEDA, MARÍA. 2021. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling. *Mathematics and Computers in Simulation*, **186**, 19–28. MATCOM Special Issue MACMAS 2019: First International Conference on Mathematical and Computational Modelling, Approximation and Simulation.

LEE, SUNGHEE, & VALLIANT, RICHARD. 2009. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research*, **37**(3), 319–343.

RIVERS, DOUGLAS. 2007. "Sampling for Web Surveys.".

ROSENBAUM, PAUL, & RUBIN, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.