Selection of demographic variables in post-stratification

Mingmeng Geng[1] **and Roberto Trotta**[1]

[1] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Italy
e-mail: mgeng@sissa.it, rtrotta@sissa.it

Abstract

Post-stratification has been widely used in survey data and performs well in real data. Although it is implemented in different ways, some demographic variables such as gender, age, region, and education level are usually used for post-stratification weighting. This classic setup has been proven very effective in a variety of situations, and our aim in this work is to find some potentially better choice of demographic variables for post-stratification.

Based on multiple public and private data sets containing more demographic variables, we used individual-based machine learning models to predict people's opinions. Under this framework, the effects of different demographic variables on one person's perspective have been explored. Using different feature selection methods and prediction models, we found some possible better combinations of demographic variables for the individual predictions of different questions in the surveys, not only the ones often used before. As a result, more reasonable options for survey design and post-stratification become possible.

We also investigated the treatment and simplification of continuous variables such as age and income. For example, statistics on income are often noisy and we want to know how it affects the prediction model. In practice, these variables are often divided into groups, which facilitates the calculation of post-stratification but obscures the information for individual predictions. Therefore, we also discussed the trade-off between the two parts, for example, the division of age groups.

Keywords: feature selection, machine learning, survey design, continuous variables