ROC2025 - International Workshop on ROC Analysis and Related Topics
Vigo, January 23 and 24, 2025

Dealing with the estimation of the ROC curve and the area under the curve in the presence of complex sampling design data

Amaia Iparragirre[1], Irantzu Barrio[1,2] and Inmaculada Arostegi[1,2]

[1]Department of Mathematics, University of the Basque Country UPV/EHU
[2]BCAM - Basque Center for Applied Mathematics

Abstract

Complex survey data is becoming increasingly popular in a variety of fields, including health and social sciences. This type of data is collected from a finite population, concerned to be studied, following some complex sampling design, such as stratification or clustering. Due to the particularities of the data collection process, each sampled unit is assigned a sampling weight indicating the number of units it represents in the finite population. Thus, the straightforward application of the most commonly applied statistical techniques, which are typically designed to be applied to simple random samples, is usually not suitable in this context of complex survey data.

In particular, in this work, we focus on the evaluation of the discrimination ability of logistic regression models by means of the receiver operating characteristic (ROC) curve and the area under it (AUC). We propose an estimator for the estimation of the ROC curve ($\widehat{ROC}_w$) and AUC ($\widehat{AUC}_w$) that account for complex sampling designs. In addition, it is well known that when the same data is used to fit the model and estimate its AUC, this estimate overestimates the real discrimination ability of the fitted model. Thus, we propose to correct for the optimism of the $\widehat{AUC}_w$ by means of replicate weights methods such as the design-based cross-validation and the rescaling bootstrap in the context of complex survey data. The proposed methods have been validated by means of several simulation studies.

All the methods proposed to estimate the ROC curve and the AUC, as well as, the AUC correction methods are available in the R-package svyROC in CRAN (https://cran.r-project.org/web/packages/svyROC/index.html).

References

Iparragirre, A., Barrio, I., Arostegui, I. (2023). Estimation of the ROC curve and the area under it with complex survey data. *Stat*, 12(1). https://doi.org/10.1002/sta4.635

Iparragirre, A., Barrio, I. (2024). Optimism Correction of the AUC with Complex Survey Data. In: Einbeck, J., Maeng, H., Ogundimu, E., Perrakis, K. (eds) *Developments in Statistical Modelling. IWSM 2024. Contributions to Statistics.* Springer, Cham. https://doi.org/10.1007/978-3-031-65723-8_7