Journées de méthodologie statistique de l'Insee 2022

Treatment of unit nonresponse in surveys through machine learning methods : an empirical comparison - Working version

Khaled Larbi (*), David Haziza (**), Mehdi Dagdoug (***)

(*) Insee / Ensae
(**) University of Ottawa
(***) Université de Bourgogne France-Comté

khaled.larbi@insee.fr/khaled.larbi@ensae.fr
dhaziza@uottawa.ca
mohamed_mehdi.dagdoug@univ-fcomte.fr

Mots-clés. Non réponse totale, machine learning, calage, simulation, échantillonnage.

Domaines. Non réponse, Machine Learning / apprentissage automatique, classification.

Résumé

Cet article propose une étude empirique sur l'effet de différentes méthodes d'apprentissage automatique (machine learning) dans un contexte de repondération de la non-réponse totale pour des données d'enquête. Soit $\mathscr{U}$ une population de taille N et $\mathscr{S}$ un échantillon probabiliste de taille n tiré dans $\mathscr{U}$ selon un plan de sondage aléatoire donné. Dans un contexte de non-réponse totale, seul un sous-ensemble $\mathscr{S}_r$ de l'échantillon $\mathscr{S}$ répond à l'enquête. Le but est d'estimer le total d'une variable d'intérêt y donné par $t_y = \sum_{i \in \mathscr{U}} y_i$. Soit $(p_k)_{k \in \mathscr{S}}$ le vecteur des probabilités de réponse. Si les probabilités de réponse étaient connues, un estimateur sans biais du total serait l'estimateur par double dilatation, $\hat{t}_{y,exp} = \sum_{i \in \mathscr{S}_r} \frac{y_i}{p_i \pi_i}$ où π_i désigne les probabilités d'inclusion d'ordre un de l'invididu i associées au plan de sondage. Cependant, les probabilités de réponse $(p_k)_{k \in \mathscr{S}_r}$ n'étant pas connues en pratique, elles sont estimées au moyen d'un modèle de non-réponse. À partir de ces estimations des probabilités de réponse, deux estimateurs du total sont considérés : l'estimateur ajusté par les scores de propension PSA, $\hat{t}_{y,PSA} = \sum_{i \in \mathscr{S}_r} \frac{y_i}{\hat{p}_i \pi_i}$, et l'estimateur de Hàjek $\hat{t}_{y,Hajek} = \frac{N}{\hat{N}} \hat{t}_{y,PSA}$ où $\hat{N}$ est l'estimateur PSA du nombre d'individus dans la population.

Nous présenterons les résultats d'une étude empirique dont le but est de comparer la performance de ces deux estimateurs en termes de biais et d'efficacité. Dans un contexte de plan de sondage stratifié à probabilités inégales, nous avons utilisé plusieurs méthodes afin d'estimer les probabilités de réponse.

Trois autres méthodes dérivées des estimations obtenues par apprentissage automatique sont proposées :

- après avoir fourni un jeu de probabilités estimées (par exemple, par régression logistique), il est possible de calculer d'autres estimations des probabilités de réponse en utilisant la méthode des scores (ou groupe homogène de réponse) [HB07]. Les probabilités estimées servent à créer des groupes d'individus homogènes par rapport à la probabilité estimée (en triant l'échantillon selon la probabilité estimée et en découpant en K groupes). Pour chaque groupe homogène, la probabilité de réponse estimée correspond au taux de réponse observé dans ce groupe. Cette approche permet d'être plus robuste à des problèmes de mauvaise spécification du modèle de non-réponse.
- il est également possible de combiner plusieurs estimations des probabilités de réponse provenant de différentes méthodes (CART, SVM et BART par exemple) afin de créer un jeu de probabilités estimées. Nous considérons deux manières de combiner les estimations dans le but d'accroître la robustesse des estimateurs :
  - estimation robuste basée sur le calage de plusieurs jeux de probabilités estimées.
  - estimation robuste basée sur une méthode ensembliste appliquées aux probabilités estimées.

Les performances de ces estimateurs seront estimées à l'aide du biais relatif Monte-Carlo et du risque quadratique moyen Monte-Carlo.

Bibliography

[BFOS83] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. 1983.

[BGC18] David Haziza Brigitte Gelein and David Causeur. Pondération pour correction de la non-réponse totale et machine learning. *Acte des JMS 2018*, 1, 2018.

[BGV92] Bernhard E. Boser, Isabelle Guyon, and Vladimir Naumovich Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92*, 1992.

[Bre04] Leo Breiman. Random forests. *Machine Learning*, 45 :5–32, 2004.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[CGM10] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart : Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), Mar 2010.

[DGH20] Mehdi Dagdoug, Camelia Goga, and David Haziza. Imputation procedures in surveys using nonparametric and machine learning methods : an empirical comparison, 2020.

[DS92] Jean-Claude Deville and Carl-Erik Sarndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382, 1992.

[HB07] David Haziza and Jean-Francois Beaumont. On the construction of imputation classes in surveys. *International Statistical Review*, 75 :25–43, 2007.

[HT52] D.G. Horvitz and DJ Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260) :663–685, 1952.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[LD04] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects : a comparative study. *Statistics in medicine*, 23 19 :2937–60, 2004.

[Nar51] R. D. Narain. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3 :169–175, 1951.

[Pla99] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[Qui92] J. Ross Quinlan. Learning with continuous classes. 1992.

[Qui93] J. Ross Quinlan. Combining instance-based and model-based learning. In *ICML*, 1993.

[RBK05] Susana Rubin-Bleuer and Ioana Schiopu Kratina. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6) :2789–2810, 2005.

[RUB76] DONALD B. RUBIN. Inference and missing data. *Biometrika*, 63(3) :581–592, 12 1976.

[SS97] P.L.D. Nascimento Silva and C.J. Skinner. Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1) :23–32, 1997.

[ZH07] Achim Zeileis and Kurt Hornik. Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61 :488–508, 2007.

[ZHH08] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17 :492 – 514, 2008.