Handling missing data with superpopulation models, design based approach and machine learning

Brigitte Gelein

Mots clés : apprentissage, calage, imputation, MIVQUE, non réponse, pondération, sondage.

En statistique d'enquête, on distingue non réponse totale et non réponse partielle. La première a lieu lorsqu'aucune information n'est utilisable pour une unité de l'échantillon, la seconde correspond au cas où seules quelques variables d'intérêt sont renseignées.

Après une introduction générale, nous nous plaçons dans un contexte de non réponse partielle. Nous proposons une méthode d'imputation en deux étapes et préservant la corrélation entre les variables d'intérêt (Gelein et al., 2014). La première étape consiste à obtenir des valeurs imputées initiales par la méthode de Shao et Wang (2002). Ces valeurs sont ensuite modifiées de façon à respecter des contraintes de calages basées des estimateurs MIVQUE des paramètres de modèle (Causeur, 2006).

Nous traitons également l'imputation de variables présentant un grand nombre de valeurs nulles. Nous proposons deux procédures d'imputation basées sur des modèles de mélange qui préservent la fonction de répartition. Les résultats d'une étude par simulation illustrent les bonnes performances de ces procédures.

Enfin, nous étudions l'estimation de probabilités de réponse dans un contexte de pondération pour correction de la non réponse totale. Nous comparons un grand nombre de méthodes d'estimation par apprentissage supervisé. Nous couvrons un large champ de méthodes paramétriques ou non, avec des règles de décisions simples ou agrégées telles que Bagging, Random Forests (Breiman, 1996), Boosting (Freund et Shapire, 1996, Friedman et al. 2000), Gradient boosting and Stochastic Gradient Boosting (Friedman 2002, Culp et al. 2006). Pour chaque méthode, les performances de l'estimateur par expansion et de l'estimateur de Hajek d'un total sont évaluées.