# Consistance sous un modèle de réponse de la fonction de répartition estimée en présence de données manquantes

*Hélène BOISTARD[1], Guillaume CHAUVET[2], David HAZIZA[3]*

On rencontre des problèmes de données manquantes dans les enquêtes quand certaines des unités refusent de répondre, ou quand il est impossible de les contacter.
Dans le contexte des Essais Cliniques, on peut également obtenir des données manquantes pour les sujets qui abandonnent ou sont perdus de vue pendant le traitement. Les estimateurs non ajustés pour la non-réponse peuvent être fortement biaisés si les répondants diffèrent des non-répondants au regard des variables étudiées.
Il est donc souhaitable de développer des procédures d'estimation conduisant à des biais aussi faibles que possible. Les procédures d'estimation doublement robustes ont été largement étudiées dans le cadre de la Statistique classique, voir par exemple Robins et al. (1994), Scharfstein et al. (1999), Tan (2006, 2007, 2010), Kang et Schafer (2007), Cao et al. (2009), parmi d'autres. Dans le cadre d'un sondage en population finie, l'estimation doublement robuste a été étudiée par Kott (1994), Kim et Park (2006), et Haziza et Rao (2006), parmi d'autres.

Nous nous intéressons ici au cas où une méthode d'imputation simple est utilisée pour remplacer une valeur manquante par une valeur artificielle.
L'objectif principal de l'imputation est de réduire le biais d'estimation, ce qui nécessite de disposer de variables auxiliaires bien explicatives et disponibles pour toutes les unités de l'échantillon. Les méthodes d'imputation simples sont souvent utilisées dans les agences de Statistique pour le traitement de la non-réponse partielle.
Afin d'étudier les propriétés des estimateurs, nous considérons deux approches : (i) l'approche par le modèle de non-réponse (NM), et (ii) l'approche par le modèle d'imputation (IM) qui nécessite la spécification d'un modèle décrivant la distribution de la variable étudiée. Un estimateur est dit doublement robuste s'il reste asymptotiquement sans biais et consistant quand l'un des deux modèles spécifiés (NM ou IM) est correct. Les procédures doublement robustes offrent, dans une certaine mesure, une protection contre une mauvaise spécification de l'un des deux modèles.
Cependant, la littérature mentionnée ci-dessus traite seulement le cas où l'on estime la valeur moyenne de le variable d'intérêt. A notre connaissance, le cas d'un estimateur doublement robuste de la fonction de répartition est peu (ou pas) abordé dans la littérature.

Dans ce travail, nous considérons le cas d'un mécanisme d'imputation des valeurs manquantes par la régression aléatoire pondérée (motivé par le modèle IM). Plus précisément, pour chaque non-répondant, un résidu observé pour un répondant est sélectionné au hasard avec une probabilité proportionnelle à un poids d'imputation, et utilisé dans l'imputation de la valeur

[1] helene@boistard.fr, Université Toulouse 1
[2] chauvet@ensai.fr, Crest (Ensai)

[3] (haziza@DMS.UMontreal.CA), Université de Montréal.

manquante. Nous étudions les propriétés de double robustesse de la fonction de répartition estimée, en fonction des poids d'imputation utilisés. Les propriétés de la fonction de répartition estimée sont évaluées théoriquement, et à l'aide d'une étude par simulations.

Bibliographie

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96, 723-734.

Haziza, D. and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. Survey Methodology 32, 53-64.

Kang, D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). Statistical Science 22, 523-539.

Kim, J.K. and Park, H.A. (2006). Imputation using response probability. Canadian Journal Statistics 34, 171–182.

Kott, P.S. (1994). A note on handling nonresponse in sample surveys. Journal of the American Statistical Association 89, 693–696.

Robins J, Rotnitzky A, Zhao LP. 1994. Estimation of regression coefficients when some of the regressors are not always observed. Journal of the American Statistical Association 89, 846-866.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion and rejoinder). Journal of the American Statistical Association 94, 1096-1146.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. Journal of the American Statistical Association 101, 1619-1637.

Tan, Z. (2007). Understanding OR, PS and DR. Statistical Science 22, 560-568.

Tan, Z. (2010). Nonparametric likelihood and and doubly robust estimating equations for marginal and nested structural models. Canadian Journal of Statistics 38, 609-632.