Inférence robuste multiple en présence de données d'enquête manquantes

David Haziza[1] & Sixia Chen[2]

1 Département de mathématiques et de statistique, Université de Montréal,Québec, H3C 3J7, Canada. david.haziza@umontreal.ca.
2WESTAT, 1600 Research Blvd, Rockville MD 20850, USA. chensixiaster@gmail.com.

Dans les enquêtes, il faut se résigner au fait qu'une partie des données sera manquante. On distingue la non-réponse totale de la non-réponse partielle. Dans cette présentation, l'accent sera mis sur la non-réponse partielle qui est habituellement traitée par imputation. Les valeurs imputées sont habituellement construites au moyen du modèle (paramétrique ou non-paramétrique) décrivant le lien entre la variable à imputer et un ensemble de variables auxiliaires observées pour les répondants et les non-répondants. Ce modèle est souvent appelé modèle d'imputation. Il est également possible de décrire la relation entre la probabilité de réponse pour la variable à imputer et un ensemble de variables auxiliaires observées pour les répondants et les non-répondants. Ce modèle est souvent appelé modèle de non-réponse. Lorsque les deux modèles sont utilisés simultanément dans la construction des variables imputées, cela conduit à des estimateurs doublement robustes ; voir, par exemple, Haziza et Rao (2006), Cao, Kang et Schaffer (2008), Cao, Tsiatis et Davidian (2009), Kim et Haziza (2014). Un estimateur est dit doublement robuste s'il est asymptotiquement sans biais et convergent pour le vrai paramètre lorsque le modèle d'imputation et/ou le modèle de non-réponse est correctement spécifié. Autrement dit, la propriété de double robustesse est intéressante car elle fournit une certaine protection lorsque l'un des deux modèles est correctement spécifié.

Récemment, une nouvelle famille de méthode a été étudiée dans la littérature dans un contexte de populations infinies : celle des méthodesrobustes multiples ; voir Han et Wang (2013). En pratique, il est possible d'ajuster plusieurs modèles d'imputation et plusieurs modèles de non-réponse, chacun comprenant potentiellement différentes variable auxiliaires et fonctions de liens. Autrement dit, on distingue de classes de modèle : l'une comprenant plusieurs modèle d'imputation et l'autre comprenant plusieurs modèles de non-réponse. Un estimateur est dit robuste multiple s'il est asymptotiquement sans biais et convergent pour le vrai paramètre lorsque l'un des modèles dans l'une des deux classes est correctement spécifié. Le concept de robustesse peut donc être vu comme une extension de celui de double robustesse. Dans cette présentation, nous adapterons et généraliserons cette famille de méthode au cas de population finies et présenterons les propriétés théoriques des estimateurs résultants. Habituellement, les agences statistiques fournissent des fichiers complétés aux utilisateurs. Nous montrerons comment implanter un estimateur robuste multiple au moyen d'une imputation calée. Finalement, nous présenterons des estimateurs de variance au moyen de méthodes de ré-échantillonnage, qui sont également robustes multiples, pourvu que la fraction de sondage soit petite. Finalement, les résultats d'une étude par simulation seront présentés. Ces résultats confirmeront les bonnes propriétés des estimateurs ponctuels et des estimateurs de variance.

Références

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

Han, P. et Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100, 417-430.

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

Kang, J.D.Y. etSchafer, J.L. (2008). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523-539.

Kim, J.K. et Haziza, D. (2014). Doubly robust inference with missing survey data. *StatisticaSinica*, 24, 375--394.