CIBLER DES VARIABLES IMPORTANTES DURANT LE TRAITEMENT DE LA NON-RÉPONSE TOTALE DANS LES ENQUÊTES

David HAZIZA (*), Sixia CHEN (**), Yimeng GAO (*)

(*) Département de mathématiques et de statistique, Université de Montréal
(**) Department of biostatistics and epidemiology, University of Oklahoma

david.haziza@umontreal.ca

Mots-clés : Calage, Non-réponse totale, Probabilité de réponse.

Résumé

La non-réponse totale dans les enquêtes est caractérisée par une absence totale d'information sur les unités échantillonnées. Les estimateurs non-ajustés (par exemple, la moyenne des répondants) sont généralement biaisés car les non-répondants tendent à exhiber des caractéristiques différentes de celles observées chez les répondants.

Il existe un certain nombre de méthodes permettant de réduire le biais causé par la non-réponse totale. La plus courante consiste à éliminer les non-répondants du fichier et de modifier le poids de base des unités répondantes afin de compenser pour l'élimination des unités non-répondantes. Généralement, le poids de base des unités est ajuté en le multipliant par l'inverse de la probabilité de réponse estimée.

La modélisation des probabilités de réponse comporte une étape importante : la sélection de variables explicatives de la non-réponse. Cependant, l'inclusion dans le modèle de variables fortement liées à la probabilité de réponse tend à générer des probabilités estimées très petites, ce qui conduit à des poids ajustés très dispersés et à des estimateurs ajustés pour la non-réponse instables. Idéalement, le vecteur des variables explicatives ne devrait inclure que celles expliquant à la fois la probabilité de réponse et les variables d'intérêt de l'enquêtes. En effet, si certaines des variables choisies ne sont pas liées aux variables d'intérêt, elles ne contribueront pas à réduire le biais des estimateurs mais contribueront à faire augmenter la variance des estimateurs. En pratique, compte tenu du grand nombre de variables d'intérêt collectées, les méthodologues se contentent généralement de choisir les variables explicatives de la non-réponse sans prendre en compte les variables d'intérêt de l'enquête. Pour des variables jugées importantes, il est donc souhaitable de développer une procédure de repondération qui permettra d'améliorer la qualité des estimations.

Considérons un sous-ensemble de p variables d'intérêt jugées importantes. La procédure proposée peut être décrite comme suit : pour chacune des p variables, on construira un modèle au moyen de l'information disponible pour les répondants et les non-répondants ce qui conduira à un ensemble de p vecteurs de valeurs prédites, un pour chacune des variables. Ensuite, au moyen d'une méthode par calage, on obtiendra un système de pondération qui satisfera p+2 contraintes de calage : p contraintes correspondant aux variables jugées importantes, une contrainte sur la taille de la population ainsi qu'une contrainte sur les probabilités de réponse. Le système de pondération ainsi obtenu conduira à des estimateurs convergents pour toutes les variables d'intérêt de l'enquête, pourvu que le modèle de non-réponse est valide. De plus, pour les p variables jugées importantes, les estimateurs résultants auront la propriété de double robustesse. Nous présenterons les résultats d'une étude par simulation qui comparera plusieurs estimateurs ajustés pour la non-réponse en termes de biais et d'efficacité.