Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

IMPUTATION MULTIPLE DE DONNÉES MANQUANTES SUR LE REVENU AUX NUCEAUX INDUVUDUEL ET FAMILIAL PAR RÉGRESSION SÉQUENTIELLE: APPLICATION À LA NATIONA HEALTH INTERVIEW SURVEY

Trivellore E. Raghunathan, Nathaniel Schender, Pei-Lu Chiu et Diane Makuc[1]

RÉSUMÉ

Les analystes de données sur la santé cherchent souvent à étudier les liens entre la revenue t la santé. La National Health Interview Survey, réalisée par le National Center for Health Statistics des Centers for Disease Control and Prevention des États-Unis, constitue une riche source de données pour l'étude de ce genre de liens. Cependant, les taux de non-réponse à deux questions essentielles sur le revenu, à savoir le revenu individuel et le revenu familial total, sont supérieurs à 20%. En outre, ces taux de non-réponse semblent augmenter avec le temps. Un projet en cours vise à proécer à une imputation multiple du revenu individuel et du revenu familial, ainsi que des valeurs de certaines autres covariables pour les cycles de la National Health Interview Survey de 1997 et des années subséquentes.

La mise au point de méthodes d'imputation multiple pour des enquêtes à aussi grande échelle pose de nombreux défis. En premier lieu, il existe un grand nombre de variables de type différent pour lesquelles les sauts de questions et les relations logiques diffèrent. En deuxième lieu, on ignore quelles associations seront étudiées par las analystes des données résultant d'imputations multiples. Enfin, les données sur certaines variables, comme le revenu familial, sont recueillies au niveau de la famille et d'autres, comme celles sur le revenu tiré d'un travail, sont recueillies au niveau individuel. Afin que les imputations pour les variables de niveau familial et individuel soient subordonnées à un aussi grand nombre que possible de prédicateurs, et pour simplifier la modélisation, nous utilisons une version modifiée de la méthode d'imputation par régression séquentielle décrite dans Raghunathan et coll. (2001, *Techniques d'enquête*). Nous créons des imputations, à deux degrés pour tenir compte des deux niveaux distincts d'unités analytiques, l'individu et la famille. À la première étape, nous imputons les valeurs des covariables de niveau individuel, puis nous créons des valeurs sommaires au niveau familial. Ensuite, nous sommons ces variables pour créer des valeurs au niveau familial et nous imposons la valeur des ces sommes comme conditions. Pour créer une interdépendance entre les variables de niveau familial et de niveau individuel, nous procédons à une itération entre les deux étages en utilisant une variable dont la valeur a été imputée au niveau familial (individuel) lors de l'itération précédente comme prédicateur dans l'itération courante de l'imputation des valeurs des variables au niveau individuel (familial).

Outre les problèmes liés à la nature hiérarchique des imputations que nous venons de décrire, d'autres questions méthodologiques méritent d'être examinées, comme l'utilisation de transformations des variables de revenu, l'imposition de restrictions sur les valeurs des variables, la validité générale de l'imputation par régression séquentielle et, de façon encore plus générale, la validité des inférences basées sur une imputation multiple dans le cas d'enquêtes à plan d'échantillonnage complexe.

La présente communication décrira l'imputation multiple de données sur le revenu dans le cas de la National Health Interview Survey, ainsi que les problèmes méthodologiques qui se posent. En outre, nous présenterons des résumés empiriques des imputations, ainsi que les résultats d'une évaluation par la méthode de Monte Carlo des inférences basées sur des données sur le revenu résultant d'une imputation multiple.

[1] University of Michigan et NCHA, É.-U., teraghu@umich.edu