Procédures d'imputation jointe pour des variables catégorielles - Une application à l'enquête Patrimoine 2010

Hélène CHAPUT[1], Laurianne SALEMBIER[2], Julie SOLARD[3], Guillaume CHAUVET[4], David HAZIZA[5]

L'imputation simple consiste à remplacer une valeur manquante par une valeur plausible. L'objectif principal de l'imputation est de réduire le biais de non-réponse, qui peut être important lorsque répondants et non-répondants diffèrent par rapport aux variables étudiées. Pour réduire le biais de non-réponse, il est nécessaire de disposer de variables auxiliaires bien explicatives, et disponibles pour toutes les unités de l'échantillon.

Dans les enquêtes auprès des ménages et dans les enquêtes sociales, on recueille généralement des variables catégorielles. Afin d'éviter d'imputer des valeurs impossibles dans le fichier de données, on utilise généralement une forme d'imputation par donneur telle que la méthode du plus proche voisin ou le hot-deck aléatoire.
Ce type d'imputation consiste à sélectionner un répondant (donneur) dans l'ensemble des répondants, en utilisant des caractéristiques du non-répondant (receveur). Nous considérons ici une version aléatoire pondérée du hot-deck, où les donneurs sont sélectionnés au hasard avec une probabilité proportionnelle au poids d'échantillonnage. En pratique, le hot-deck aléatoire pondéré est généralement appliqué dans des classes d'imputation, qui sont formées sur la base des informations auxiliaires connues pour toutes les unités de l'échantillon.

On s'intéresse souvent à l'estimation de paramètres simples, tels que des totaux ou des moyennes. Dans ce cas l'imputation marginale, qui consiste à imputer les variables séparément, conduit à des estimateurs asymptotiquement sans biais si le modèle utilisé est correctement spécifié (Haziza, 2009).
On peut par exemple utiliser un hot-deck aléatoire pour chaque variable à imputer. Cependant, l'imputation marginale tend à distordre les relations entre variables. En particulier, les estimateurs de paramètres bivariés, tels que les coefficients de régression ou de corrélation, peuvent être sévèrement biaisés surtout si les taux de non-réponse sont importants. Il est donc souhaitable d'utiliser des méthodes d'imputation qui permettent de préserver les relations entre variables. Pour des paramètres bivariés impliquant des variables continues, Shao et Wang (2002) ont proposé une méthode d'imputation jointe permettant une estimation asymptotiquement sans biais d'un coefficient de corrélation. Chauvet et Haziza (2011) ont proposé une version équilibrée de la méthode de Shao et Wang, permettant d'éliminer (ou de réduire fortement) la variance d'imputation.

[1] (helene.chaput@insee.fr
[2] (laurianne.salembier@insee.fr, Division Revenus et patrimoine des ménages, DSDS, Insee

[3] (julie.solard@insee.fr

[4] (chauvet@ensai.fr), Crest (Ensai).

[5] haziza@DMS.UMontreal.CA, Université de Montréal

Nous étudions ici une procédure d'imputation jointe proche de la méthode du hot-deck aléatoire. Nous montrons que la méthode proposée permet de préserver le coefficient de corrélation entre deux variables catégorielles. Nous illustrons les méthodes considérées dans le cadre de l'Enquête Patrimoine 2010, réalisée par l'Insee. Dans le cadre de cette enquête, les non réponses partielles aux variables centrales ont été imputées par hot deck aléatoire équilibré, et par la méthode proposée le cas échéant.

Bibliographie

Chauvet, G., and Haziza, D. (2011). Fully efficient estimation of coefficients of correlation in the presence of imputed data. A paraître dans Canadian Journal of Statistics.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference, Editors: C.R. Rao and D. Pfeffermann, 215-246.

Shao, J. and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. Journal of the American Statistical Association, 97, 544-552.