

IMPUTATIONS DE DONNÉES DE REVENU À L'AIDE DE CALAGE GÉNÉRALISÉ ET DE LOIS GB2 : ILLUSTRATION SUR LES DONNÉES SILC SUISSES 2009

Éric Graf¹ & Yves Tillé¹

¹ *Institut de Statistique, Université de Neuchâtel, Pierre-à-Mazel 7, CH - 2000 Neuchâtel, eric.graf@unine.ch, yves.tille@unine.ch*

Dans les enquêtes par échantillonnage auprès de ménages et des personnes, les questions sur le revenu sont souvent sensibles et donc sujettes à un taux de non-réponse plus élevé ; la non réponse partielle s'ajoutant à la non-réponse totale.

Malgré cela, le revenu du ménage ou de la personne cible sont des variables centrales de l'enquête. La distribution des revenus ainsi récoltés n'est pas normale, ni log-normale. Les hypothèses des modèles de régressions classiques pour expliquer les revenus (ou leur logarithme) ne sont donc pas vérifiées. Les imputations s'appuyant sur de tels modèles altèrent la distribution originale (et "naturelle" !) des données, ce qui n'est pas souhaitable et peut conduire l'utilisateur à interpréter faussement les résultats calculés à partir d'un jeu de données ainsi imputé (par exemple les indicateurs de Læken). En fait, on risque d'introduire un biais supplémentaire.

La loi bêta généralisée de seconde espèce (GB2) est une distribution à quatre paramètres. Elle a été développée par McDonald (1984). Des études empiriques sur le revenu montrent que la GB2 s'ajuste bien à de telles données et qu'elle est souvent plus adaptée que d'autres distributions à quatre paramètres. Les résultats du projet européen AMELI (2011) confirment pour des données de revenus issues de l'enquête EU-SILC.

L'avantage d'une estimation paramétrique d'une distribution de revenu est qu'il existe des formules explicites pour les mesures d'inégalité comme des fonctions des paramètres de la loi ajustée aux données.

Deux grandes philosophies d'imputations sont envisageables pour résoudre le problème : imputations non paramétriques par donneurs (par ex. plus proches voisins) et les imputations paramétriques modélisant les données. On présentera une méthode du deuxième type reposant sur l'ajustement d'une loi GB2 en s'appuyant sur des poids obtenus par calage généralisé - voir par ex. Deville *et al.* (1993) - et corrigeant mieux pour la non réponse non ignorable dont est affectée la variable d'intérêt.

On illustrera la méthode avec des données réelles de l'enquête SILC suisse 2009.

Bibliographie

Deville, J-C., et Särndal, C-E., et Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Graf, M., et Nedyalkova, D., et Münnich, R., et Seger, J., et Zins, S. (2011). Parametric Estimation of Income Distributions, and Indicators of Poverty and Social Exclusion. *Projet européen AMELI, Deliverable 2.1.*

McDonald, J. (1984). Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647-663.