Risque d'amplification du biais de l'estimateur par calage généralisé enprésence de non-réponse

Éric Lesage [1] & David Haziza [2] & Xavier D'Haultfoeuille [3]

1 INSEE, 18 boulevard Adolphe Pinard, 92240 Malakoff, France. eric.lesage@insee.fr
2 Département de mathématiques et de statistique, Université de Montréal,
Québec, H3C 3J7, Canada. david.haziza@umontreal.ca.
3 CREST, 15 boulevard Gabriel Péri, 92240 Malakoff, France. xavier.dhaultfoeuille@ensae.fr.

Les procédures de repondération sont des pratiques courantes en méthodologie d'enquête. Les instituts de statistique utilisent généralement une procédure à deux étapes : dans une première étape les poids sont modifiés pour corriger la non-réponse totale, puis dans une seconde étape, les poids sont de nouveau ajustés afin que les estimations de l'enquête coïncident avec les totaux connus de la population. A la première étape, le statisticien d'enquête a pour objectif de réduire le biais de non-réponse qui peut être important lorsque les caractéristiques des non-répondants sont différentes de celle des répondants. La réduction efficace du biais de non-réponse repose sur la disponibilité d'une information auxiliaire explicative de la non-réponse qui consiste en un vecteur de variables auxiliaires disponible pour les répondants et les non-répondants. A cette étape le poids d'échantillonnage d'une unité est divisé par sa probabilité de répondre estimée à l'aide d'un modèle de réponse paramétrique ou non-paramétrique. Une méthode couramment utilisée consiste à répartir les répondants et les non-répondants dans des classes de pondération et d'ajuster les poids d'échantillonnage des répondants par l'inverse des taux de réponse dans chaque classe ; voir par exemple Eltinge et Yansaneh (1997), et Little (1986). A la seconde étape, un calage (par exemple une prost-stratification) est mis en œuvre afin d'assurer la cohérence entre les estimations de l'enquête et les totaux connus sur la population entière. Le calage nécessite l'existence de variables auxiliaires disponibles pour les répondants et dont les totaux sur la population sont également disponibles. En outre, si la variable d'intérêt est liée aux variables auxiliaires alors l'estimateur calé sera plus efficace que l'estimateur non-calé.

Une méthode de repondération alternative a reçu beaucoup d'attention ces dernières année : il s'agit d'une approche en une étape qui utilise un estimateur par calage qui vise 3 objectifs simultanés : réduire le biais de non-réponse, assurer la cohérence entre les estimations de l'enquête et les totaux connus sur la population et, si possible, réduire la variance. A la différence de l'approche en deux étapes, il n'est pas nécessaire ici de spécifier un modèle de non-réponse ; voir par exemple Deville (2000), Sautory (2003), Särndal et Lundström (2005) et Kott (2006).

Nous nous consacrons dans notre présentation à l'approche en une étape et nous mettons en évidence les risques d'amplification du biais et de la variance de l'estimateur par calage généralisé. Si on note **X** le vecteur des variables de calage et **Z** le vecteur des instruments de calage, on montre que les variables **Z** ont vocation à être les variables explicatives de la non-réponse et que les variables **X** sont des **instruments économétriques** (ou variables proxy de **Z**) pour le modèle de non-réponse. A ce titre, il faut choisir des variables **X** qui vérifient la relation d'exclusion $\mathbf{X} \perp R \mid \mathbf{Z}$, où R est la variable indicatrice de réponse. Si cette condition n'est pas vérifiée exactement, l'estimateur par calage généralisé présente un biais. Et ce biais est d'autant plus amplifié que la corrélation entre **X** et **Z** est faible. Ce phénomène est similaire au problème des instruments faibles rencontré par les économètres dans le traitement des problèmes d'endogénéité (Baker, 1995).