Risque d'amplification de biais de l'estimateur par calage généralisé en présence de non-réponse

Éric LESAGE [1] & David HAZIZA [2]

[1] *CREST(ENSAI) et IRMAR, Campus de Ker Lann, F-35172 BRUZ, eric.lesage@ensai.fr*
[2] *CREST(ENSAI) et Université de Montréal, Haziza@DMS.UMontreal.CA*

Résumé. Dans cette présentation, il sera question de l'utilisation du calage généralisé comme méthode de pondération en une étape. Le calage poursuit alors simultanément trois objectifs : 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.
Nous examinons les propriétés de l'estimateur par calage généralisé dans le cas où les variables instrumentales (variables explicatives de la probabilité de répondre) ne sont disponibles que pour les répondants à l'enquête. Nous mettons en évidence les risques d'amplification de biais de l'estimateur par calage généralisé en présence de non-réponse. Ce type de phénomène a été étudié en épidémiologie ; Pearl (2010) et Myers et al. (2011).

Mots-clés. Amplification du biais, Calage généralisé, Variable Proxy, non-réponse totale.

Abstract. In this presentation, we discuss the so-called single step approach to weighting in the context of calibration for unit nonresponse. It consists of using calibration with three simultaneous goals in mind: reduce the nonresponse bias, ensure consistency between survey estimates and known population totals and, possibly, contribute to variance reduction.
We examine the properties of instrument vector calibration, where the instrumental variables (related to the response propensity) are available for the respondent units only. More specifically, the problem of bias amplification is illustrated. This phenomena has been discussed in the epidemiological literature; Pearl (2010) and Myers et al. (2011).

Keywords. Bias amplification, Instrument vector calibration, Proxy variable, Unit nonresponse.

1 Présentation

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 puissante 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 oeuvre 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.

Considérons une population U de taille N. L'objectif est d'estimer le total sur la population $t_y = \sum_{k \in U} y_k$, d'une variable d'intérêt y. Un échantillon, s, de taille n, est sélectionné selon un plan de sondage $p(s)$. Un estimateur à partir de données complètes de t_y est l'estimateur par expansion

$$\hat{t}_\pi = \sum_{k \in s} d_k y_k,$$

où $d_k = 1/\pi_k$ est le poids d'échantillonnage de l'unité k et $\pi_k = P(k \in s)$ est la probabilité d'inclusion dans l'échantillon à l'ordre un. En présence de non-réponse on n'observe qu'un sous-ensemble s_r de s, ce qui rend impossible le calcul de $\hat{t}_\pi$.
Afin de définir un estimateur de t_y ajusté pour la non-réponse, on suppose qu'un vecteur de variables auxiliaires $\mathbf{x}$ est disponible pour $k \in s_r$ et que le vecteur de totaux pour la population $\mathbf{t_x} = \sum_{k \in U} \mathbf{x}_k$ est connu. Par ailleurs, on suppose qu'un vecteur de variables

instrumentales $\mathbf{z}$, de même dimension que $\mathbf{x}$, est disponible pour $k \in s_r$. Les variables instrumentales sont supposées expliquer les probabilités de répondre à l'enquête des unités. Soit R_k une indicatrice de réponse de l'unité k telle que $R_k = 1$ si k répond à l'enquête et $R_k = 0$, sinon. On considère un estimateur par calage instrumental de la forme :

$$\hat{t}_C = \sum_{k \in s} w_k R_k y_k, \tag{1}$$

où

$$w_k = d_k F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k), \tag{2}$$

et $F(.)$ est une fonction monotone deux fois différentiable. Le poids de calage w_k dans (2) est le produit du poids d'échantillonnage d_k et du facteur d'ajustement $F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k)$, qui a vocation à être un estimateur de l'inverse de la probabilité de réponse de l'unité k.

Les poids sont construits de manière à ce que les équations de calage

$$\sum_{k \in s_r} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k \tag{3}$$

soient satisfaites.

L'objet de la présentation est d'examiner le problème dit d'amplification de biais dans le contexte du calage généralisé. On mettra en évidence que l'existence et l'intensité d'un biais de l'estimateur par calage généralisé sont liés au choix du vecteur de variables auxiliaires $\boldsymbol{x}$ utilisé dans les équations de calage. On montrera que s'il est difficile de choisir $\boldsymbol{x}$ de manière à annuler le biais, il est par contre possible de guider le choix de $\boldsymbol{x}$ afin d'éviter le risque d'amplification du biais.

2 Références bibliographiques

Deville, J-C. (2002). La correction de la non-réponse par calage généralisé. Actes des Journées de Méthodologie Statistique, Insee.

Eltinge, J. L. and Yansaneh, I. S. (1997). Diagnostics For Formation Of Nonresponse Adjustment Cells, With An Application To Income Nonresponse In The U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33–40.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. *Survey Methodology*, 32, 133–142.

Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, 54, 139–157.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11), 1213-1222.

Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. arXiv preprint arXiv:1203.3503.

Särndal, C.E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley and Sons.

Sautory, O. (2003). Calmar 2: a new version of the Calmar calibration adjustment program. *Proceedings of the Statistics Canada Symposium*, Ottawa, Canada.