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

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

[1] *Laboratoire de Statistique d'Enquête, CREST-ENSAI, Campus de Ker Lann, 35170 Bruz, 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.*

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 et de variance de l'estimateur par calage généralisé en présence de non-réponse. Ce type de phénomène a été étudié notamment en épidémiologie ; Pearl (2010) et Myers et al. (2011). Une étude pas simulation illustre nos résultats.

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

1 Introduction

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 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.

L'objectif de cette présentation est de mettre en évidence les risques d'amplification du biais de l'estimateur par calage généralisé en présence de non-réponse. On montre que les variables de calage $\boldsymbol{x}$ sont des **instruments économétriques** pour le modèle de non-réponse et qu'un mauvais choix de ces $\boldsymbol{x}$ peut conduire au problème des instruments faibles bien connu en économétrie. Nous montrons également que, même s'il n'y a pas de biais, la variance peut être amplifiée lorsque les variables de calage sont faiblement corrélées aux instruments. On trouve des résultats préliminaires dans Lesage (2012) et Osier (2012).

Considérons une population finie U de taille N. Notre 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 dans U selon un plan de sondage $p(s)$. L'esti-

mateur par expansion est un estimateur de t_y construit à partir de données complètes

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

où $d_k = 1/\pi_k$ est le poids d'échantillonnage associé à l'unité k et $\pi_k = P(k \in s)$ est sa probabilité d'inclusion à l'ordre un. En présence de non-réponse, seul un sous-ensemble s_r de s est observé, ce qui rend impossible le calcul de $\hat{t}_\pi$.

Afin de définir une estimateur de t_y, ajusté pour la non-réponse, on suppose qu'on dispose d'un vecteur de variables auxiliaires $\mathbf{x}$ pour $k \in s_r$ et du vecteur des totaux sur la population, $\mathbf{t_x} = \sum_{k \in U} \mathbf{x}_k$. En pratique, le vecteur $\mathbf{x}$ est souvent définis par le commanditaire de l'enquête, qui veut assurer la cohérence entre les estimations et les vrais totaux de certaines variables (par exemple : age et sexe). En outre, on suppose qu'un vecteur d'**instruments de calage z**, de la même dimension que $\mathbf{x}$, est disponible pour $k \in s_r$. Il n'est pas nécessaire de connaître le vecteur des totaux sur la population $\mathbf{t_z} = \sum_{k \in U} \mathbf{z}_k$. Les instruments de calage sont supposés être liés à la probabilité de répondre des unités.

Soit R_k la variable indicatrice de réponse associée à l'uinité k telle que $R_k = 1$ si l'unité k est répondante et $R_k = 0$, sinon. Nous considérons un estimateur par calage généralisé 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 continument dérivable. Le vecteur de paramètres $\hat{\boldsymbol{\lambda}}_r$ est défini comme solution des équations de calage

$$\sum_{k \in s} d_k \; R_k \; F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{3}$$

Puisqu'on s'intéresse avant tout à l'erreur due à la non-réponse et non à l'erreur d'échantillonnage, on choisit pour la suite de la présentation de se placer dans le cas particulier d'un recensement $s = U$.

Quelque soit la fonction de calage $F(.)$, l'estimateur par calage $\hat{t}_C$ estime exactement le vrai total t_y si la variable d'intérêt y est parfaitement expliquée

par le vecteur $\mathbf{x}$, i.e., $y_k = \mathbf{x}_k^\top \boldsymbol{\beta}$ pour un vecteur $\boldsymbol{\beta}$ donné. De ce fait, on s'attend à ce que $\hat{t}_C$ ait un faible biais lorsque la variable y et le vecteur $\mathbf{x}$ sont liés linéairement et que le terme d'erreur $\left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta}\right)$ n'est pas lié au vecteur des instruments $\mathbf{z}_k$.
Toutefois, dans des enquêtes multi-sujets, le nombre de variables d'intérêt est grand et un certain nombre d'entre elles sont des variables catégorielles et non des variables continues. Il est donc peu réaliste, dans la plupart des situations pratiques, de considérer que la variable y est liée linéairement au vecteur $\mathbf{x}$.
En revanche, si $F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k)$ est un bon estimateur de l'inverse de la probabilité de réponse de l'unité k, p_k^{-1}, alors $\hat{t}_C$ est asymptotiquement sans biais pour t_y quelque soit la variable y considérée. Il reste que, en pratique, la sélection du vecteur $\mathbf{z}$ n'est pas claire dans la mesure où ces variables ne sont observées que pour les répondants. En outre, à supposer que le vecteur $\mathbf{z}$ soit correctement choisi, comment faut-il procéder pour valider la forme fonctionnelle dans le modèle de non-réponse et partant la fonction de calage ? Bien que ces questions soient importantes, elles sortent du périmètre de notre présentation. Les lecteurs intéressés peuvent consulter Haziza et Lesage (2013) à ce sujet.

2 Convergence de l'estimateur par calage généralisé

Par mesure de simplicité, nous considérons que $\mathbf{x}_k^\top = (1, x_k)$ et $\mathbf{z}_k^\top = (1, z_k)$. Soient $\{(x_k, y_k, z_k, r_k)^\top, k \in U\}$ les réalisations des vecteurs aléatoires *iid* $\{(X_k, Y_k, Z_k, R_k)^\top, k \in U\}$. On suppose que la corrélation entre x et z, $\mathbb{C}orr\,(X_k,\ Z_k)$, est non nulle.
La probabilité de réponse, p_k, associée à l'unité k est définie par :

$$p_k = \mathbb{E}\left\{R_k \mid (Y_k, Z_k) = (y_k, z_k)\right\} \tag{4}$$

Conditionnellement à la population finie, le mécanisme de non-réponse est modélisé par une loi de Bernoulli $R_k \sim \mathcal{B}(1, p_k)$.

Dans cette section, on suppose qu'on a **les relations d'exclusion** suivantes

$$R_k \perp X_k \mid Z_k \tag{5}$$
$$R_k \perp Y_k \mid Z_k. \tag{6}$$

Le graphe de la Figure 1 illustre ces relations d'exclusion. On peut noter que la relation d'exclusion (6) ne correspond pas à une situation *Missing*

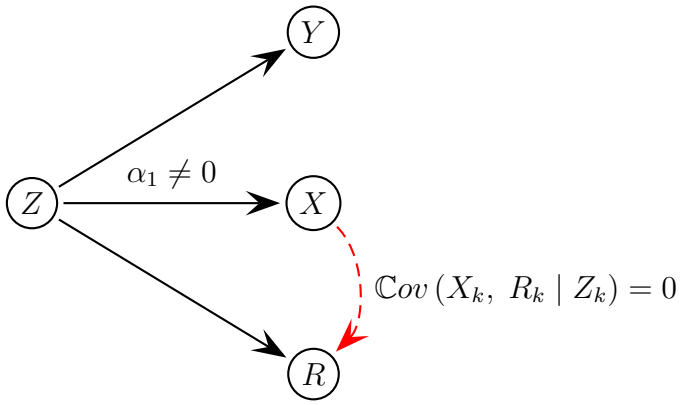

FIGURE 1 – Relation entre les variables Y, Z, X and R

At Random (Rubin, 1976) parce que la variable z n'est observée que sur le répondants.
De (4) et (6), on déduit que la probabilité de réponse n'est fonction que de la variable z et on suppose en outre que la relation peut s'écrire :

$$p_k = h\left(\boldsymbol{\lambda}_0^\top \mathbf{z}_k\right), \tag{7}$$

où $h(\cdot)$ est une fonction a valeurs strictement positives et inférieures à 1.

Soit $\boldsymbol{\lambda}_N$ le vecteur solution des équations estimantes moyennes :

$$\sum_{k\in U} \left\{1 - p_k F\left(\boldsymbol{\lambda}_N^\top \mathbf{z}_k\right)\right\} \mathbf{x}_k = \mathbf{0}. \tag{8}$$

On peut montrer que le biais approché de $\hat{t}_C$ s'écrit

$$ABias_q(\hat{t}_C) \quad = \quad -\sum_{k\in U}(1 - p_k F_k)\left(y_k - \mathbf{z}_k^\top \boldsymbol{\beta}\right) + \frac{\beta_1}{\alpha_1}\sum_{k\in U}(1 - p_k F_k)\left(x_k - \mathbf{z}_k^\top \boldsymbol{\alpha}\right).$$

où $F_k \equiv F\left(\boldsymbol{\lambda}_N^\top \mathbf{z}_k\right)$, et $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$ et $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ sont des vecteurs quelconques.

Si $h^{-1}(\cdot)$ correspond à la fonction de calage $F(\cdot)$ alors, sous certaines hypothèses techniques et sous les hypothèses (5), (6) et (7), **l'estimateur par calage généralisé est convergent**.

Si la forme fonctionnelle $h(\cdot)$ du modèle de non-réponse est inconnue mais que les conditions d'exclusion (5) et (6) sont toujours vérifiées, l'estimateur

par calage généralisé peut encore être convergent à la condition qu'on ait des modèles de régression linéaire entre y et $\boldsymbol{z}$ et entre x et $\boldsymbol{z}$:

$$\mathbb{E}(X_k \mid Z_k) = \boldsymbol{Z}_k^\top \boldsymbol{\alpha} \tag{9}$$
$$\mathbb{E}(Y_k \mid Z_k) = \boldsymbol{Z}_k^\top \boldsymbol{\beta}. \tag{10}$$

3 Amplification du biais de l'estimateur par calage généralisé

On a montré que l'estimateur par calage $\hat{t}_C$ pouvait être convergent (approximativement sans biais) sous les conditions d'exclusion (5) et (6). On examine à présent la situation où la condition d'exclusion (5) est violée : $\mathbb{C}ov(R_k, X_k \mid Z_k) \neq 0$.
On met en évidence que dans ce contexte l'estimateur par calage $\hat{t}_C$ n'est pas convergent et que son bais approché est une fonction proportionnelle à la corrélation entre y et z, à $\mathbb{C}ov(R_k, X_k \mid Z_k)$ et à l'inverse de la corrélation entre x et z. On constate alors que le bias est amplifié si la corrélation entre x et z est proche de zéro. Il en ressort qu'il est important de veiller à ce que les variables de calage soient fortement liées aux instruments de calage avant de procéder à un calage généralisé. A défaut d'éliminer le biais, cela permet au moins de se prémunir contre une amplification de ce dernier.

References

Chang, T. and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. Biometrika, 95, 557–571.

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.

Haziza, D. and Lesage, E. (2013). A discussion of weighting procedures in the presence of unit nonresponse. Submitted for publication.

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

Kott, P.S. (2009). Calibration weighting : Combining probability samples and linear prediction models : Handbook of Statistics 29B, Sample Surveys : Inference and Analysis. In Pfeffermann, D. & Rao, C.R. (Eds.), Oxford, UK : Elsevier.

Kott, P.S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. Journal of the American Statistical Association, 105, 1265–1275.

Kott, P.S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. Survey Research Methods, 6, 105–111.

Lesage, E. (2012). Correction de la non-réponse non ignorable par une approche modèle. Actes des Journées de Méthodologie Statistique de l'Insee, Paris, France.

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 and K. J., Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. American Journal of Epidemiology, 174, 1213–1222.

Osier, G. (2012). Traitement de la non-réponse non-ignorable par calage généralisé : une simulation à partir de l'enquête Budget des Ménages au Luxembourg. Actes des Journées de Méthodologie Statistique de l'Insee, Paris, France.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In : Grünwald P, Spirtes P, editors. Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence Corvallis, 425–432.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–590.

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.

Wooldridge, J. (2009). Should Instrumental Variables Be Used As Matching Variables ? East Lansing, MI : Michigan State University. Available at www.msu.edu/ ec/faculty/wooldridge/current%20research/treat1r6.pdf.