Estimation par calage en présence de non-réponse: un survol

David Haziza [1]

[1] *Département de mathématiques et de statistique, Université de Montréal, Montréal, Canada. email: david.haziza@umontreal.ca*

Résumé. Dans la grande majorité des enquêtes, les taux de réponse sont à la baisse, ce qui accroît le risque de biais de non-réponse. Il existe essentiellement deux approches de pondération permettant de corriger la non-réponse: (i) la pondération en deux étapes qui est l'approche traditionnelle dans les agences statistiques et (ii) la pondération en une étape dont l'utilisation a crû depuis une quinzaine d'années et qui fait l'objet d'un ouvrage (Särndal et Lundström, 2005). Dans cette présentation, nous présenterons les deux approches et leurs propriétés en termes de biais et de variance. Dans le contexte de la pondération en une étape, nous discuterons choix de la fonction de calage, qui est un aspect important, un choix inapproprié pouvant conduire à des estimateurs biaisés. Nous présenterons brièvement le calage généralisé et ses propriétés. En particulier, nous discuterons des problèmes de l'amplification du biais et de l'amplification de la variance.

Mots-clés. Calage, Fonction de calage, Non-réponse totale, Pondération en une étape, Pondération en deux étapes, Probabilité de réponse.

1 Pondération dans les enquêtes

La pondération occupe une place prépondérante dans les enquêtes menées au sein d'agences statistiques. Les erreurs non dues à l'échantillonnage (par exemple, les erreurs dues à la non-réponse et les erreurs de couverture) sont habituellement traitées au moyen de procédures d'ajustement des poids; voir Kalton and Flores-Cervantes (2003) pour une excellente discussion de la pondération dans les enquêtes. Dans la grande majorité des enquêtes, les taux de réponse sont à la baisse, ce qui accroît le risque de biais de non-réponse. Il existe essentiellement deux approches de pondération permettant de corriger la non-réponse: (i) la pondération en deux étapes qui est l'approche traditionnelle dans les agences statistiques et (ii) la pondération en une étape dont l'utilisation a crû depuis une quinzaine d'années et qui fait l'objet d'un ouvrage (Särndal et Lundström, 2005).

Soit d_k le poids initial de sondage de l'unité k découlant du plan de sondage et défini comme l'inverse de la probabilité d'inclusion dans l'échantillon. Le système de pondération basique est donné par $\{d_k; k \in s\}$, où s désigne l'échantillon tiré de la population selon un certain plan de sondage. Bien qu'il garantisse une absence de biais dans des conditions

idéales, ce système n'est généralement pas approprié dans un contexte de non-réponse totale. De plus, il ne garantit pas la cohérence entre les estimations produites par l'enquête et les totaux connus au niveau de la population.

En l'absence d'erreurs non dues à l'échantillonnage, le calage consiste à déterminer un système de pondération calé, $\{\widetilde{w}_k; k \in s\}$, aussi proche que possible du système de pondération basique, $\{d_k; k \in s\}$, et tel que les contraintes de calage, définies par les utilisateurs, sont satisfaites. Un poids de calage s'exprime comme le produit du poids de base et d'un facteur de calage, ce dernier dépendant de la fonction de calage utilisée. On a donc $\widetilde{w}_k = d_k \times F_k$, où F_k est un facteur de calage associé à l'unité k. La fonction linéaire, la fonction exponentielle, la fonction linéaire tronquée et la fonction logit sont des fonctions de calage communément utilisées en pratique. Deville and Särndal (1992) ont montré que les estimateurs par calage sont asymptotiquement convergents par rapport au plan de sondage et que toutes les fonctions de calage sont asymptotiquement équivalentes au sens où elles conduisent toutes à l'estimateur par calage obtenu au moyen de la fonction linéaire lorsque la taille de l'échantillon et celle de la population sont "suffisamment" grandes. La fonction de calage est généralement choisie de manière à ce que la distribution des facteurs de calage soit "acceptable". Par exemple, lorsque le nombre de contraintes de calage est important et/ou que la taille de l'échantillon est modérée, il n'est pas rare d'aboutir à des poids négatifs dans le cas de la fonction linéaire, ce qui n'est pas souhaitable du point de vue des micro données. La fonction exponentielle, quant à elle, garantit des poids de calage positifs mais potentiellement extrêmes, pouvant alors conduire à des estimateurs instables. Les fonctions linéaire tronquée et logit permettent d'obtenir des facteurs de calage compris entre une borne inférieure et une borne supérieure fixées par l'utilisateur, ce qui permet d'éviter le problème des poids négatifs ou extrêmes.

En présence de non-réponse totale, la pondération en deux étapes consiste à ajuster les poids de sondage en deux étapes distinctes: dans un premier temps, le poids de base d_k des répondants est multiplié par un facteur d'ajustement, défini comme l'inverse la probabilité de réponse estimée. Soit $\hat{p}_k$ la probabilité de réponse associé à l'unité k et $w_k^* = d_k \times \frac{1}{\hat{p}_k}$, son poids ajusté pour la non-réponse. Le système de pondération ajusté pour la non-réponse est donc donné par $\{w_k^*; k \in s_r\}$, où s_r désigne l'ensemble des répondants à l'enquête. Dans un deuxième temps, les poids ajustés w_k^* sont de nouveau modifiés au moyen d'un calage. Soit $\widetilde{w}_k = w_k^* \times F_k$ le poids (final) associé à l'unité k après calage. Le système de pondération final est donné par $\{\widetilde{w}_k; k \in s_r\}$.

À la première étape, l'objectif est de réduire le biais de non-réponse au moyen de l'information auxiliaire disponible pour les répondants et les non-répondants. Il est bien connu que les biais de non-réponse tendent à être importants lorsque les taux de réponse sont faibles et que le comportement des répondants est différent de celui des non-répondants par rapport aux caractéristiques mesurées par l'enquête. Réduire le biais

de non-réponse passe par la bonne spécification du modèle de non-réponse, permettant d'obtenir les probabilités de réponse estimées. Il s'agit de déterminer les variables auxiliaires qui influent à la fois sur le fait de répondre ou non à l'enquête mais également sur les variables d'intérêt. Les probabilités de réponse peuvent être obtenues au moyen d'un modèle paramétrique ou non-paramétrique. La régression logistique et la régression probit sont deux exemples de modèles paramétriques; voir Kim et Kim (2007). En pratique, les méthodes paramétriques sont rarement utilisées car les estimateurs résultants sont vulnérables à la mauvaise spécification de la fonction de lien. Une approche non-paramétrique fréquemment utilisée en pratique est la méthode des scores qui consiste à obtenir des estimations des probabilités de réponse préliminaires au moyen d'un modèle paramétrique et à diviser l'échantillon en classes homogènes en fonction de ces probabilités préliminaires. Le poids initial d_k associé à l'unité k dans une certaine classe est alors ajusté par l'inverse du taux de réponse observé dans la même classe; voir, par exemple, Little (1986), Eltinge et Yansaneh (1997) et Haziza et Beaumont (2007). Les méthodes par noyau (Giommi, 1987 et Da Silva et Opsomer, 2006), les méthodes par polynômes locaux (Da Silva et Opsomer, 2009) et les arbres de régression (Phipps et Toth, 2012) sont d'autres méthodes non-paramétriques récemment étudiées dans la littérature. Si le modèle de non-réponse est correctement spécifié, les estimateurs ajustés pour la non-réponse sont asymptotiquement sans biais et ce, quelle que soit la variable d'intérêt pour laquelle on désire une estimation. À la deuxième étape, une procédure de calage (par exemple, une post-stratification ou un calage sur marges) permet de construire des poids finaux garantissant la cohérence entre les estimations et les vrais totaux de la population pour un certain nombre de variables de calage. Il convient de noter que la fonction de calage dans un contexte de pondération en deux étapes est choisie selon des critères similaires à ceux utilisés en l'absence de non-réponse. De même, les estimateurs calés (ou finaux) tendent à être plus efficaces que les estimateurs non calés si les variables de calage sont fortement liées aux variables d'intérêt.

La pondération en une étape consiste à effectuer un calage avec trois objectifs simultanés en tête: (i) réduire le biais de non-réponse, (ii) garantir la cohérence entre les estimations et les totaux connus au niveau de la population et (iii) si possible, réduire la variance des estimateurs. Contrairement à l'approche en deux étapes, une estimation explicite des probabilités de réponse n'est pas requise. En revanche, le choix de la fonction de calage devient important, différentes fonctions pouvant potentiellement conduire à des estimateurs exhibant des propriétés très différentes en termes de biais et de variance. Par conséquent, le choix de la fonction de calage ne doit plus reposer sur des considérations cosmétiques uniquement (l'allure de la distribution des facteurs de calage F_k) mais également sur des considérations statistiques. Un choix inapproprié de la fonction de calage peut conduire à des estimateurs biaisés dont le biais peut être supérieur à celui des estimateurs non ajustés dans certaines situations. Autrement dit, bien que la pondération en une étape n'utilise pas explicitement les probabilités de réponse estimées dans la construction des

estimateurs, un exercice de modélisation est généralement inévitable afin d'assurer un bon choix de la fonction de calage. Durant la présentation, nous discuterons du choix de la fonction de calage dans un contexte de pondération en une étape et illustreront les résultats au moyen d'études par simulation. Pour une discussion de la pondération en une étape, voir Haziza et Lesage (2014) et Kott et Liao (2012).

Finalement, nous présenterons brièvement la méthode du calage généralisé (appelé également calage avec variables instrumentales) dans un contexte de non-réponse totale. La pondération en une étape classique repose sur la disponibilité de variables de calage qui sont minimalement observées pour toutes les unités échantillonnées ou encore qui ne sont observées que sur les répondants mais dont le total au niveau de la population est connu. Contrairement à la pondération en une étape classique, le calage généralisé permet d'incorporer des variables observées sur les répondants seulement; voir, par exemple, Deville (2002) et Kott (2006). Nous discuterons des propriétés des estimateurs de calage résultant sen termes de biais et de variance. En particulier, nous montrerons que les estimateurs de calage peuvent souffrir d'un phénomène d'amplification du biais et/ou d'amplification de la variance; voir Lesage, Haziza et D'Haultfœuille (2014). Ce problème d'amplification du biais a aussi été discuté dans la littérature épidémiologique; voir, par exemple, Pearl (2010, 2012).

Bibliographie

[1] Da Silva, D.N. et J.D. Opsomer (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics*, 34, 563–579.

[2] Da Silva, D.N. et J.D. Opsomer (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35, 165–176.

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

[4] Deville, J.-C. et Särndal, C.-E. (1992). Calibration Estimators In Survey Sampling. *Journal of the American Statistical Association*, 87, 376–382.

[5] Eltinge, J.L. et 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.

[6] Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology*, 13, 127–134.

[7] Haziza, D. et Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75, 25–43.

[8] Haziza, D. et Lesage, E. (2014). A discussion of weighting procedures for unit nonresponse. En révision pour *Journal of Official Statistics*.

[9] Kalton, G. et Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19, 81–97.
[10] Kim, J.K. et Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501–514.
[11] Kott, P. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. *Survey Methodology*, 32, 133–142.
[12] Kott, P.S. et Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods*, 6, 105–111.
[13] Lesage, E., Haziza, D. et D'Haultfœuille, X. (2014). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. En révision pour *Journal of Survey Statistics and Methodology*.
[14] Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, 54, 139–157.
[15] 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.
[16] Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. arXiv preprint arXiv:1203.3503.
[17] Phipps, P. et Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6, 772–794.
[18] Särndal, C.E. et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley and Sons.