Dois-t-on toujous utiliser la pondration par calage ?

Mohammed El Haj Tirari[1] & Boutaina Hdioud[2]

[1]*INSEA, Madinat al Irfane, Rabat-Instituts, B.P. 6217, Rabat, Maroc,*
mtirari@hotmail.fr
[2]*ENSIAS, Avenue Mohamed Ben Abdellah Regragui, Rabat, Maroc,*
hdioud.boutaina@hotmail.fr

Résumé. En présence d'information auxiliaire, la technique de calage est la plus utilisée pour en tenir compte dans le but d'améliorer la précision des estimations produites. Cependant, les pondérations par calage peuvent ne pas convenir à toutes les variables d'intérêt de l'enquête, en particulier celles qui ne sont pas liées aux variables auxiliaires utilisées dans le calage. Nous proposons une *mesure* permettant d'évaluer pour chaque variable d'intérêt, l'apport de la pondération par calage à la précision de l'estimateur de son total. Cette *mesure* peut être utilisée pour déterminer les variables d'intérêt pour lesquelles il convient d'utiliser la pondération par calage.

Mots-clés. Calage, modèle de superpopulation, Effet de plan.

1 Introduction

Lors de l'estimation des paramètres de la population, on fait souvent recours aux techniques de redressement pour réduire la variance ou corriger la non-réponse. En présence d'information auxiliaire, le calage est la technique de redressement la plus utilisée en pratique. Les poids de l'estimateur par calage permettent de redresser l'échantillon de manière à refléter les totaux connus dans la population d'un ensemble de variables auxiliaires (Deville et Särndal 1992). L'amélioration en termes de précision apportée par l'estimateur par calage dépend des variables auxiliaires utilisées dans le calage. En effet, le biais et la variance de l'estimateur par calage sont faibles quand les variables de calage sont fortement liées à la variable d'intérêt.

En pratique, une fois les poids de calage sont calculés, ces derniers remplacent les poids de sondage et ils sont utilisés pour produire les estimations des paramètres de toutes les variables d'intérêt de l'enquête. Cependant, l'utilisation de la pondération par calage peut engendrer une augmentation de l'erreur quadratique moyenne (EQM) de certaines variables d'intérêt, en particulier celles qui ne sont pas liées aux variables auxiliaires utilisées dans le calage. Il est donc nécessaire d'élaborer une *mesure* permettant d'évaluer pour chaque variable d'intérêt, l'impact de l'utilisation de la pondération par calage en termes de précisions des estimations produites.

Pour élaborer ce type de *mesure*, on peut se baser sur l'effet de plan (Deff) mesurant l'augmentation ou la diminution relative de la variance d'un estimateur par rapport à celle correspondante au cas d'un sondage aléatoire simple. Kish (1965) propose une version adaptée de l'effet de plan pour mesurer l'impact d'utiliser des poids inégaux mais sans tenir compte de l'apport de ces derniers en termes d'efficacité statistique. Cependant, souvent l'utilisation de poids inégaux peut améliorer la précision des estimations puisqu'ils permettent de corriger les erreurs de couverture ou de non-réponse (Särndal et Lundström 2005; Kott 2009).

Henry et Valliant (2015) ont proposés sous l'approche assistée par un modèle une *mesure* de l'effet de plan qui traduit les effets conjoints d'un plan de sondage à probabilités inégales et de l'ajustement des poids de sondage par ceux de calage. En suivant la même démarche de Henry et Valliant mais en considérant l'approche basée sur le plan et le modèle, nous proposons dans ce travail une nouvelle *mesure* de l'effet de la pondération par calage. La *mesure* proposée tient compte de degré du lien existant entre la variable d'intérêt et les variables de calage. De plus, elle est facile à calculer pour chaque variable d'intérêt afin de savoir s'il convient d'utiliser les poids de calage pour estimer ses paramètres.

2 Notations et définitions

Soit $U = \{1, \ldots, N\}$ une population de taille N à partir de laquelle on sélectionne un échantillon s de taille n selon un plan de sondage $p(\cdot)$ dont les probabilités d'inclusion d'ordres un et deux sont données respectivement par π_k et π_{kl}. On s'intéresse à une variable d'intérêt $\mathbf{y} = (y_1, \ldots, y_N)^{'}$ en ayant pour objectif l'estimation de son total :

$$t_y = \sum_{k \in U} y_k$$

On dispose de p variables auxilaires $X_1, \ldots, X_p$ dont les valeurs peuvent être représentées par les vecteurs $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kp}\right)^{'}$ pour tout $k \in U$. On suppose que le vecteur des totaux

$$t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$$

des p variables auxiliaires $(X_1, \ldots, X_p)$ est connu.

Sous l'approche basée sur le modèle, on suppose que les valeurs de la variable d'intérêt $\mathbf{y}$ sont les réalisations d'un modèle de superpopulation ξ définit par

$$y_k = \mathbf{x}_k^{'}\boldsymbol{\beta} + \epsilon_k \tag{1}$$

avec

$$E_\xi(\epsilon_k) = 0, \;\; var_\xi(\epsilon_k) = \sigma_u^2 v_k^2 \;\text{ et }\; cov_\xi(\epsilon_k, \epsilon_l) = 0.$$

Les v_k^2 sont supposés connus et par souci de simplification, nous supposons dans ce qui suit que $v_k^2 = 1$. E_ξ, var_ξ et cov_ξ représentent respectivement l'espérance, la variance et la covariance sous le modèle.

Pour estimer le total t_y de la variable d'intérêt $\mathbf{y}$, on considère la classe des estimateurs par calage qui sont définis par

$$\widehat{t}_{yCal} = \sum_{k \in S} w_{kS} y_k$$

où w_{kS} sont les poids de calage qui vérifient les équations de calage données par

$$\sum_{k \in S} w_{kS} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$$

Dans ce qui suit, nous allons considérer l'approche basée sur le plan et le modèle. Sous cette approche, le critère utilisé pour mesurer la précision d'un estimateur par calage est

$$EQM_{p\xi}(\widehat{t}_{yCal}) = E_p E_\xi (\widehat{t}_{yCal} - t_y)^2$$

qui représente l'Ecart Quadratique Moyen sous le plan et le modèle, appelé aussi l'EQM *anticipé*.

3 La *mesure* proposée pour l'effet de la pondération par calage

La *mesure* de l'effet de la pondération par calage que nous proposons est une extension de l'effet du plan utilisé pour comparer la variance sous un plan de sondage $p(s)$ d'un estimateur $\widehat{t}_y$ du total d'une variable d'intérêt Y, par rapport à la variance de l'estimateur du même total dans le cas où l'échantillon s a été sélectionné selon un sondage aléatoire simple (SAS). Il est défini par :

$$Deff(\widehat{t}_y) = \frac{Var_p(\widehat{t}_y)}{Var_{SAS}(\widehat{t}_{ySAS})}$$

Il permet de mesurer l'effet d'utiliser une stratégie d'estimation $(p(s), \widehat{t}_y)$ autre que celle correspondante au sondage aléatoire simple.

Ainsi, pour pouvoir élaborer une *mesure* de l'effet de la pondération par calage sous l'approche basée sur le plan et le modèle, nous considérons l'adaptation suivante de l'effet du plan :

$$Deff_\xi(\widehat{t}_{yCal}) = \frac{EQM_{p\xi}(\widehat{t}_{yCal})}{Var_{SAS}(\widehat{t}_{ySAS})}$$

Notons qu'on peut montrer qu'une approximation de l'EQM anticipé de l'estimateur par calage $\widehat{t}_{yCal}$ est donnée par (Tirari, 2012) :

$$EQM_{p\xi}(\widehat{t}_{yw}) \approx \sigma_u^2 \sum_{k\in U} \left[R_{w_k}^2(d_k - 1) + (R_{w_k} - 1)^2\right] \quad (2)$$

avec $R_{w_k} = \frac{w_k}{d_k}$ est le rapport des poids de calage w_k sur les poids de sondage $d_k = \frac{1}{\pi_k}$ et σ_u^2 est la variance des résidus du modèle (1).

Par conséquent, l'effet de la pondération par calage sous l'approche basée sur le plan et le modèle peut être approximée par :

$$Deff_\xi(\widehat{t}_{yCal}) \approx \frac{\sigma_u^2 \sum_{k\in U} \left[R_{w_k}^2(d_k - 1) + (R_{w_k} - 1)^2\right]}{N^2(1 - \frac{n}{N})\frac{S_y^2}{n}} \quad (3)$$

$$\approx \frac{n}{N^2(1 - \frac{n}{N})}\frac{\sigma_u^2}{\sigma_y^2} \sum_{k\in U} \left[R_{w_k}^2(d_k - 1) + (R_{w_k} - 1)^2\right] \quad (4)$$

où σ_y^2 est la variance de la variable d'intérêt $\mathbf{y}$. L'approximation (4) de $Deff_\xi$ a l'avantage de tenir compte des effets conjoints du plan de sondage à probabilités inégales (d_k), des poids de calage (w_k) et de la force du lien entre la variable d'intérêt et les variables de calage (σ_u^2). De plus, on peut estimer l'approximation (4) par

$$\widehat{Deff}_\xi(\widehat{t}_{yCal}) \approx \frac{n}{N^2(1 - \frac{n}{N})}\frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_y^2} \sum_{k\in s} d_k \left[R_{w_k}^2(d_k - 1) + (R_{w_k} - 1)^2\right] \quad (5)$$

Ainsi, l'expression (5) peut être utilisée comme une *mesure* de l'effet de la pondération par calage. En effet, pour une variable d'intérêt $\mathbf{y}$, les poids de calage doivent être utilisés lorsque la valeur de (5) est inférieur à 1. Notons que la valeur de (5) est faible quand le lien entre $\mathbf{y}$ est les variables de calage est fort ($\widehat{\sigma}_u^2 \ll \widehat{\sigma}_y^2$). Dans le cas contraire, nous avons

$$\widehat{u}_k = y_k - \mathbf{x}_k^{'}\widehat{\boldsymbol{\beta}} \approx y_k \quad \text{et} \quad \widehat{\sigma}_u^2 \approx \widehat{\sigma}_y^2$$

et l'expression de (5) devient

$$\widehat{Deff}_\xi(\widehat{t}_{yCal}) = \frac{n}{N^2(1 - \frac{n}{N})} \sum_{k\in s} d_k(d_k - 1) \quad (6)$$

qui est égale à 1 dans le cas d'un sondage aléatoire simple ($d_k = \frac{N}{n}$).

4 Conclusion

Dans ce papier, nous avons proposé une nouvelle *mesure* de l'effet de la pondération par calage qui a l'avantage de tenir compte des effets conjoints du plan de sondage à probabilités inégales, des poids de calage et de la force du lien entre la variable d'intérêt et les variables de calage. La *mesure* proposée peut être calculée pour chaque variable d'intérêt pour savoir si on doit utiliser ou non les poids de calage pour estimer ses paramètres.

Bibliographie

[1] Deville, J.-C. et Särndal, C.-E. (1992), Calibration estimators in survey sampling, Journal of the American Statistical Association, 87(418), 376-382.
[2] El Haj Tirari, M. (2012). Critère du choix des variables auxiliaires à utiliser dans l'estimateur par calage. Septième Colloque Francophone sur les sondages, Rennes, France.
[3] Henry, K. A. et Skinner, R. (2015), A design effect measure for calibration weighting in single-stage samples, Survey Methodology, 41, N 2, 315-331.
[4] Kott, P. (2009). Calibration weighting: Combining probability samples and linear prediction models. Dans *Handbook of Statistics, Sample Surveys: Design, Methods and Application*, (éds., D. Pfeffermann et C.R. Rao), 29B, Amsterdam : Elsevier BV.
[5] Särndal, C.-E. et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley and Sons.