L'estimateur par calage généralisé en présence de non-réponse

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*

En présence de la non-réponse totale, on fait recours souvent aux méthodes de repondération pour la corriger. Pour cela, deux procédures de repondération sont utilisées : la procédure à deux étapes où dans une première étape les poids sont modifiés pour corriger la non-réponse, puis dans une seconde étape, les poids sont de nouveau ajustés de manière à refléter les totaux connus dans la population d'un ensemble de variables auxiliaires. La deuxième procédure de repondération est celle à une étape qui utilise un estimateur par calage généralisé qui vise à la fois la réduction du biais de non-réponse, l'estimation exacte des totaux connus dans la population des variables auxiliaires et éventuellement la réduction de la variance. Chacune des deux procédures a des avantages et des inconvénients. Dans cette présentation, nous proposons un critère permettant d'orienter le choix entre ces deux procédures.

Mots-clés. Calage généralisé, non-réponse, modèle de superpopulation.

1 Introduction

En présence de non-réponse totale, les méthodes de repondération sont les plus utilisées afin d'en tenir compte au moment de l'estimation des paramètres de la population. Ainsi, deux procédures de repondération sont utilisées pour corriger cette non-réponse. La procédure à deux étapes où 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 de manière à refléter les totaux connus dans la population d'un ensemble de variables auxiliaires. La première étape a pour objectif la réduction du biais dû à la non-réponse en se basant sur la connaissance d'une information auxiliaire représentée par des variables auxiliaires disponibles pour les répondants et les non-répondants. Pour cela, le poids d'échantillonnage de chaque unité est multiplié par l'inverse de sa probabilité de réponse estimée au moyen d'un modèle paramétrique ou non-paramétrique. Ensuite, la deuxième

étape a pour objectif d'assurer la cohérence entre les estimations obtenues à partir des répondants et l'information auxiliaire disponible sur la population en réalisant par exemple un calage sur les totaux connus sur la population. Cette approche de repondération nécessite donc la spécification d'un modèle de non-réponse à sa première étape.

La deuxième procédure de repondération est celle à une étape qui utilise un estimateur par calage généralisé qui vise à la fois la réduction du biais de non-réponse, l'estimation exacte des totaux connus dans la population des variables auxiliaires et éventuellement la réduction de la variance; voir par exemple Deville (2002), Sarndäl et Lundström (2005). Contrairement à la procédure à deux étapes, il n'est pas nécessaire de spécifier un modèle de non-réponse. Cependant, la procedure à une étape peut générer un problème d'amplification de biais; voir par exemple Lesage (2013).

Comme chacune des deux procédures a des avantages et des inconvénients, nous proposons dans cette présentation un critère permettant d'orienter le choix entre ces deux procédures.

2 L'estimateur par calage en présence de non-réponse

Soit $U = \{1, \ldots, N\}$ une population de taille N à partir de laquelle un échantillon s de taille n est sélectionné selon un plan de sondage $p(\cdot)$ dont les probabilités d'inclusion d'ordre 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 :

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

Un estimateur de t_y obtenu à partir de données complètes est le π-estimateur

$$\widehat{t}_{y\pi} = \sum_{k \in S} d_k y_k$$

où $d_k = 1/\pi_k$ est le poids d'échantillonnage de l'unité k. En présence de non-réponse, on n'observe qu'un sous-échantillon s_r de s, ce qui rend impossible le calcul de $\widehat{t}_{y\pi}$.

Pour définir un estimateur de t_y tout en tenant compte de non-réponse, on suppose qu'on dispose de p variables auxilaires $X_1, \ldots, X_p$ dont les valeurs peuvent être représentées par les vecteurs $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kp}\right)^{'}$ disponibles pour $k \in s_r$ où le vecteur des totaux

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

est supposé connu.

Par ailleurs, on suppose qu'un vecteur de variables instrumentales $\mathbf{z}_k$, de même dimension que $\mathbf{x}_k$ est disponible pour $k \in s_r$. Les variables instrumentales sont supposées expliquer les probabilités de répondre à l'enquête des unités.

Un estimateur du total t_y en présence de non-réponse est donné par

$$\widehat{t}_{yC} = \sum_{k\in S} w_k R_k y_k \quad (1)$$

où R_k est l'indicatrice de réponse de l'unité k telle que $R_k = 1$ si k répond à l'enquête et $R_k = 0$, sinon. Les poids w_k dans (1) sont censés tenir compte de la non-réponse.

De plus, dans le cas de l'estimateur par calage, les poids (3) sont construits de manière à ce que les équations de calage

$$\sum_{k\in S_r} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k \quad (2)$$

soient satisfaites.

Pour la procédure à deux étapes les poids de $\widehat{t}_{yC}$ peuvent s'écrire $w_k = w_k^\star/\widehat{p}_k$ où $\widehat{p}_k$ est la probabilité de réponse de l'unité $k \in s$ estimée à l'aide d'un modèle paramètrique où non paramètrique que constitue la première étape de cette approche. A la deuxième étape, les poids $w_k^\star$ sont construits à l'aide par exemple d'un calage de telle sorte à assurer la cohérence entre les estimations calculées à partir des répondants et les totaux connus sur la population entière, c'est-à-dire la vérification des équations de calage (2).

Pour la procédure à une étape utilisant le calage généralisé (Deville, 2002), la forme générale des poids de l'estimateur $\widehat{t}_{yC}$ est donnée par

$$w_k = d_k F(\widehat{\boldsymbol{\lambda}}\mathbf{z}_k) \quad (3)$$

avec $F(\cdot)$ est une fonction monotone deux fois différentiable. Il s'agit du produit du poids d'échantillonnage d_k et du facteur d'ajustement $F(\widehat{\boldsymbol{\lambda}}\mathbf{z}_k)$ qui joue le rôle de l'estimateur de l'inverse de la probabilité de réponse de l'unité k.

Pour pouvoir comparer les différentes procédures de repondération par calage, utilisées en présence de non réponse, nous considérons l'approche modèle qui consiste à supposer que les valeurs de la variable d'intérêt $\mathbf{y}$ sont les réalisations d'un modèle de superpopulation ξ défini par

$$y_k = \mathbf{x}_k^{'}\boldsymbol{\beta} + \epsilon_k \quad (4)$$

avec

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{'},$$

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

Les v_k^2 sont supposé connus et E_ξ, var_ξ et cov_ξ représentent respectivement l'espérance, la variance et la covariance sous le modèle (4).

Un critère de mesure de la précision de l'estimateur $\widehat{t}_{yC}$ peut être contruit en considérant l'Ecart Quadratique Moyen sous le plan de sondage, le mécanise de non-réponse et le modèle de superpopulation (4) :

$$EQM_{p_r\xi}(\widehat{t}_{yw}) = E_{p_r}E_\xi(\widehat{t}_{yw} - t_y)^2$$

où $E_{p_r} = E_p E_r(\cdot \mid s)$ représente l'espérance sous les mécanismes d'échantillionnage et de non-réponse.

En effet, on peut montrer que

$$EQM_{p_r\xi}(\widehat{t}_{yw}) \;=\; \sigma^2 \sum_{k\in U} v_k^2 \left[var_{p_r}(w_k R_k I_k) + [E_{p_r}(w_k R_k I_k) - 1]^2\right]$$

et une approximation de $EQM_{p_r\xi}(\widehat{t}_{yw})$ peut être donnée par

$$EQM_{p_r\xi}(\widehat{t}_{yw}) \;\approx\; \sigma^2 \sum_{k\in U} v_k^2 \left[(d_k - 1)\frac{(p_k w_k)^2}{d_k^2} + \left[\frac{p_k w_k}{d_k} - 1\right]^2\right] \tag{5}$$

Ainsi, cette approximation peut être utilisée comme un critère de comparaison entre les différentes procédures de calage, utilisées en présence de non-réponse. Des exemples d'utilisation de ce critère seront donnés dans cette présentation.

Bibliographie

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

[2] Lesage, E. (2013), *Utilisation d'information auxiliaire en théorie des sondages à l'étape de l'échantillonnage et à l'étape de l'estimation*, thèse de doctorat, Rennes.

[3] Sarndäl, C.-E. et Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, John Wiley and Sons, New York.