Doit-on utiliser toujours la pondération par calage ?

Mohammed El Haj Tirari
Institut National de Statistique et d'Economie Appliquée

Neuvième Colloque Francophone sur les Sondages, Gatineau, Québec, Canada

11-14/10/2016

Introduction

- 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.
- En effet, lors de l'estimation des paramètres de la population, on dispose de deux séries de pondération :
  1. Les poids de sondage $d_k = \frac{1}{\pi_k}$
  2. Les poids de calage w_k

Introduction

- La question qui se pose :

  $\hookrightarrow$ Pour une variable d'intérêt $\mathbf{y}$ donnée, laquelle parmi les deux séries de poids convient mieux d'utiliser pour produire les estimations les plus précises des paramètres de $\mathbf{y}$?

- L'objet de cette présentation est de proposer une réponse à cette question.

Notations

- Soit $U = \{1, \ldots, N\}$ une population de taille N à partir de laquelle on sélectionne un échantillon s de taille n.

- On s'intéresse à une variable d'intérêt $\mathbf{y} = (y_1, \ldots, y_N)^{'}$ en ayant comme objectif l'estimation de son total :

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

- On suppose qu'on dispose de p variables auxilaires $X_1, \ldots, X_p$ dont les totaux

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

sont connus, où $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kp}\right)^{'}$ pour tout $k \in U$.

Approche modèle

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

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

avec

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

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

Les v_k^2 sont supposé connus avec $\sum_{k\in U} v_k = N$

Le π-estimateur et l'estimateur par calage

- Pour estimer le total t_y d'une variable d'intérêt $\mathbf{y}$, on considère la classe des estimateurs linéaires qui peuvent s'écrire

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

où w_{kS} sont des poids qui peuvent dépendre de l'échantillon.

- Dans le cas où les variables auxilaires $X_1, \ldots, X_p$ ne sont pas prises en compte, l'estimateur linéaire utilisé est le π-estimateur donné par :

$$\sum_{k \in S} \frac{y_k}{\pi_k}$$

ce qui correspond à utiliser les poids de sondage $d_k = \frac{1}{\pi_k}$ pour la pondération.

Le π-estimateur et l'estimateur par calage

- Pour tenir compte des variables auxilaires $X_1, \ldots, X_p$, on peut utiliser l'estimateur par calage donné par

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

où les poids w_{kS} satisfient les équations de calage :

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

Les poids w_{kS} dépendent de l'échantillon.

- Le calage vise à réduire la variance des estimateurs.

Critère du choix entre les deux séries de pondération

- Pour mesurer la précision de l'estimateur $\widehat{t}_{yw}$, nous allons considèrer l'approche basée sur le plan et le modèle.
- Sous cette approche, la précision de $\widehat{t}_{yw}$ peut être mesurée par

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

où E_p et E_ξ sont respectivement les espérances sous le plan et le modèle de superpopulation.

Critère du choix entre les deux séries de pondération

L'impact de l'utilisation des poids de calage w_k à la place des poids des sondages d_k peut être mesuré par :

$$Weff = \frac{EQM_{p\xi}(\widehat{t}_{yC})}{EQM_{p\xi}(\widehat{t}_{y\pi})}$$

qui peut être approximé par

$$Weff \approx \frac{\sigma_u^2 \sum_{k\in U} v_k^2 \left[R_{w_k}^2 (d_k - 1) + (R_{w_k} - 1)^2\right]}{\sigma_u^2 \sum_{k\in U} v_k^2 (d_k - 1) + \widehat{V}_\pi}$$

où $R_{w_k} = \frac{w_k}{d_k}$, σ_u^2 est la variance des résidus du modèle et

$$\widehat{V}_\pi = var_p(\widehat{t}_{\widehat{y}\pi}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{\widehat{y}_k}{\pi_k} \frac{\widehat{y}_l}{\pi_l}$$

avec $\widehat{y}_k = \mathbf{x}_k^{'}\boldsymbol{\beta}$ et $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$.

Remarques

(1) Lorsque l'échantillon s est sélectionné selon un plan équilibré sur les variables auxiliaires $X_1, \ldots, X_p$, on a $\widehat{V}_\pi = 0$ et

$$Weff \approx \frac{\sum_{k \in U} v_k^2 \left[R_{w_k}^2 (d_k - 1) + (R_{w_k} - 1)^2 \right]}{\sum_{k \in U} v_k^2 (d_k - 1)} = 1$$

$\hookrightarrow$ C'est la série des poids de sondage qui doit être utilisée pour la pondération.

$\hookrightarrow$ La question de l'utilisation ou non des poids de calage se pose surtout lorsqu'une partie des variables de calage n'a pas été utilisée dans l'équilibrage.

Remarques

(2) Par exemple, lorsque l'échantillon s est sélectionné selon un plan de sondage de taille fixe et $\mathbf{x}_k = \pi_k$, on a

$$w_k = d_k \quad \text{et} \quad \widehat{V}_\pi = 0$$

et on retrouve bien $Weff = 1$:

$$Weff = \frac{\sum_{k\in U} v_k^2(d_k - 1)}{\sum_{k\in U} v_k^2(d_k - 1)} = 1$$

Estimateur composite

- On note que l'approximation de l'effet de pondération ($Weff$) peut nous permettre d'élaborer un nouvel estimateur composite $\widehat{t}_{yw_{Comp}}$ dont les poids sont donnés par :

$$w_k^{Comp} = Weff d_k + (1 - Weff) w_k$$

où d_k et w_k sont respectivement les poids de sondage et de calage.

- Ainsi, l'estimateur composite obtenu est

$$\widehat{t}_{yw_{Comp}} = Weff \widehat{t}_{y\pi} + (1 - Weff) \widehat{t}_{yw_C}$$

Conclusion

- Dans ce travail, nous nous sommes intéressés à la question de l'utilisation ou non de la pondération par calage pour estimer les paramètres d'une variable d'intérêt $\mathbf{y}$.
- Nous avons proposé un critère permettant d'orienter le choix de la série des poids qui convient d'utiliser pour chaque variable d'intérêt de l'enquête.
- Ce critère a l'avantage de tenir compte du plan de sondage mis en oeuvre ainsi que de l'impact des variables auxiliaires utilisées dans le calage.
- Possibilité de construire un estimateur composite à partir de l'estimateur de HT et celui de calage.