ESTIMATION D'UN TOTAL EN PRESENCE D'INFORMATION AUXILIAIRE

Mohammed EL HAJ TIRARI (*)

(*) ENSAI, Laboratoire de Statistique d'Enquêtes

Introduction

Pour estimer le total d'une population finie, l'estimation par la régression est parmi les méthodes importantes qui s'appuient sur l'information auxiliaire disponible sur la population (ou qui provient d'un échantillon plus grand) pour construire des estimateurs plus efficaces. Les estimateurs par régression entrent dans la catégorie des estimateurs linéaires. Ces estimateurs présentent un avantage particulier dans le cas de l'échantillonnage car, une fois calculé, les coefficients de pondération peuvent s'appliquer à toute variable analysée.
Dans une approche basée sur le plan de sondage, en recherchant le coefficient de régression qui minimise la variance asymptotique de l'estimateur par la régression généralisée *GREG*, Montanari (1987) propose un estimateur *GREG* du total qui est optimal selon ce critère. Malheureusement, en plus de dépendre des probabilités d'inclusion d'ordre deux, cet estimateur n'est optimal que lorsque son coefficient de régression est connu, ce qui n'est pas le cas. De plus, le coefficient de régression, de l'estimateur *GREG* proposé par Montanari, est une fonction des estimateurs de la variance et de la covariance, ce qui le rend vulnérable aux variations de l'échantillonnage. Dans ce travail, nous montrons que, si un plan de sondage équilibré est utilisé, on peut obtenir une approximation de l'estimateur optimal de Montanari, qui est indépendante des probabilités d'inclusion d'ordre deux avec un coefficient de régression plus robuste contre les variations d'échantillonnage.
Nous verrons que l'estimation du total devient très simple à calculer même pour des échantillons de grande taille, puisqu'elle ne dépend plus des probabilités d'inclusion d'ordre deux. L'estimateur proposé dans ce travail peut être exprimé comme un estimateur par la régression classique dans lequel on utilise en même temps, comme variables de régression, les variables de calage et d'équilibrage.

1. Problème et notations

Soit une population finie $U = \{1,\ldots,k,\ldots,N\}$ composée de N unités, et considérons le total

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

d'une variable d'intérêt prenant les valeurs y_k pour $k \in U$. Un échantillon s est sélectionné à partir de U en utilisant un plan de sondage $p(s)$ dont les probabilités d'inclusion sont respectivement notées par p_k et $p_{k\ell}$. Le but est d'estimer Y en utilisant l'information auxiliaire disponible et les valeurs y_k pour $k \in s$.
On considère ici l'approche basée sur le plan. On note que cette approche exige la connaissance des probabilités d'inclusion. Les p_k sont supposées connues, mais par contre les $p_{k\ell}$ sont rarement connues d'une manière exacte pour les plans de sondage complexes. Cependant, il est nécessaire de les calculer pour pouvoir utiliser l'estimateur asymptotiquement optimal de Montanari. Dans ce travail, nous montrons que, si le plan de sondage équilibré est utilisé, il est possible de construire un autre estimateur du total asymptotiquement optimal ayant l'avantage d'être indépendant des $p_{k\ell}$.
On suppose que les valeurs des M variables de calage pour l'unité $k \in U$, données par le vecteur

$$\mathbf{x}_k = (x_{k1},\ldots,x_{kM})',$$

sont connues uniquement pour les unités $k \in s$. Le vecteur des totaux de ces variables $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$ est aussi supposé connu. Les variables de calage peuvent être aussi représentées par la matrice suivante

$$\mathbf{X}_U = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$$

$\mathbf{X}_U$, généralement inconnue, est appelée matrice de calage.

Considérons l'estimateur par la régression généralisée (GREG) proposé par Särndal, Swensson et Wretman (1992) :

$$\hat{Y}_{greg} = \hat{Y} + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \hat{\mathbf{b}}_{y|x},$$

où

$$\hat{\mathbf{b}}_{y|x} = (\tilde{\mathbf{X}}_s' \mathbf{W}_s \tilde{\mathbf{X}}_s)^{-1} \tilde{\mathbf{X}}_s' \mathbf{W}_s \tilde{\mathbf{y}}_s,$$

$\tilde{\mathbf{X}}_s = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k, \ldots, \tilde{\mathbf{x}}_n)'$, $\tilde{\mathbf{y}}_s = (\tilde{y}_1, \ldots, \tilde{y}_k, \ldots, \tilde{y}_n)'$, $\tilde{\mathbf{x}}_k = \mathbf{x}_k / p_k$, $\tilde{y}_k = y_k / p_k$, $\hat{\mathbf{X}}$ est l'estimateur de Horvitz-Thompson (1952) de $\mathbf{X}$, et $\mathbf{W}_s$ est une matrice $n \times n$ positive. Le coefficient de régression $\hat{\mathbf{b}}_{y|x}$ peut être vu comme un estimateur de

$$\mathbf{b}_{y|x} = (\tilde{\mathbf{X}}_U' \mathbf{W}_U \tilde{\mathbf{X}}_U)^{-1} \tilde{\mathbf{X}}_U' \mathbf{W}_U \tilde{\mathbf{y}}_U. \qquad (1.1)$$

où $\tilde{\mathbf{X}}_U = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k, \ldots, \tilde{\mathbf{x}}_N)'$. Särndal, Swensson et Wretman (1992) définissent $\mathbf{W}_U$ comme une matrice $N \times N$ diagonale contenant des poids positifs.

Considérons maintenant l'expression

$$\hat{Y}_{greg} = \hat{Y} + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \mathbf{b}_{y|x},$$

$\hat{Y}_{greg}$ n'est pas un estimateur car $\mathbf{b}_{y|x}$ dépend de $\tilde{\mathbf{y}}_U$ dont les valeurs pour les unités $k \notin s$ sont inconnues. Montanari (1987) montre que la variance de $\hat{Y}$ est minimisée avec $\mathbf{b}_{y|x}$ lorsque $\mathbf{W}_U = \mathbf{?}_U$, où $\mathbf{?}_U = [\Delta_{k\ell}]$ et $\Delta_{k\ell} = p_{k\ell} - p_k p_\ell$ si $k \neq \ell$ et $\Delta_{k\ell} = p_k(1 - p_k)$ si $k = \ell$.

Puisque l'estimateur optimal de Montanari dépend de $\mathbf{?}_U$, notre objectif est de développer une approximation convenable de $\mathbf{?}_U$ afin d'estimer le coefficient de régression de l'estimateur de Montanari et par la suite une bonne approximation de celui-ci.

2. Le plan de sondage équilibré

Considérons que les valeurs de H variables auxiliaires sont connues pour toutes les unités de la population et sont données par les vecteurs

$$\mathbf{z}_k = (z_{k1}, \ldots, z_{kH})'.$$

Ces variables sont appelées variables d'équilibrage. Le vecteur des totaux sur la population des variables d'équilibrage,

$$\mathbf{Z} = \sum_{k \in U} \mathbf{z}_k$$

est supposé connu. Ces variables peuvent être aussi représentées par la matrice

$$\mathbf{Z}_U = (\mathbf{z}_1, \ldots, \mathbf{z}_N)',$$

appelée matrice d'équilibrage.

On suppose que l'échantillon sélectionné est tel que chacune de ces variables vérifie l'équation d'équilibrage. C'est-à-dire,

$$\sum_{k \in s} \tilde{\mathbf{z}}_k = \mathbf{Z},$$

où $\tilde{\mathbf{z}}_k = \mathbf{z}_k / p_k$. Deville et Tillé (2004) proposent une méthode générale pour sélectionner des échantillons équilibrés.

3. Approximation des probabilités d'inclusion d'ordre deux

Comme l'estimateur de Montanari dépend des probabilités d'inclusion d'ordre deux, notre objectif est de construire une approximation adéquate de ces probabilités à travers l'approximation de la matrice $\mathbf{?}_U$. Sous des conditions de régularité, Berger, Tirari et Tillé (2003) proposent une approximation de $\mathbf{?}_U$ donnée par

$$\mathbf{?}_U^* = \mathbf{C}_U(\mathbf{I} - \mathbf{P}_U) \tag{3.1}$$

où $\mathbf{I}$ est la matrice identitée $(N \times N)$ et

$$\mathbf{P}_U = \mathbf{A}\mathbf{C}_U,$$
$$\mathbf{A} = \breve{\mathbf{Z}}_U(\breve{\mathbf{Z}}_U'\mathbf{C}_U\breve{\mathbf{Z}}_U)^{-1}\breve{\mathbf{Z}}_U',$$
$$\mathbf{B}_z = \breve{\mathbf{Z}}_U(\breve{\mathbf{Z}}_U'C_U\breve{\mathbf{Z}}_U)^{-1}\breve{\mathbf{Z}}_U'\mathbf{C}_U\breve{\mathbf{y}}_U,$$
$$\mathbf{P}_U = \mathbf{A}\mathbf{C}_U.$$

$\mathbf{C}_U$ est une matrice diagonale dont les éléments diagonaux sont notés par c_k, où les c_k sont la solution de

$$p_k(1-p_k) = c_k(1-a_{kk}c_k).$$

où a_{kk} est le k-ème élément diagonal de $\mathbf{A}$. Pour des populations de grande taille, les c_k peuvent être approximés par $p_k(1-p_k)$. Berger, Tirari et Tillé (2003) donnent une preuve de (3.1) avec une description détaillée des conditions de régularité.

4. Approximation de l'estimateur optimal de Montanari

En remplaçant $\mathbf{W}_U$ par $\mathbf{?}_U^*$ dans (1.1), on peut obtenir l'approximation suivante de l'estimateur optimal de Montanari :

$$\tilde{Y}_{opt} = \breve{Y} + \left(\mathbf{X} - \breve{\mathbf{X}}\right)'\mathbf{b}_{zx-regr},$$

où $\mathbf{b}_{zx-regr,}$ est composé des M premiers éléments du vecteur

$$\mathbf{b}_{y|xz} = \left[\left(\breve{\mathbf{X}}_U, \breve{\mathbf{Z}}_U\right)'\mathbf{C}_U\left(\breve{\mathbf{X}}_U, \breve{\mathbf{Z}}_U\right)\right]^{-}\left(\breve{\mathbf{X}}_U, \breve{\mathbf{Z}}_U\right)'\mathbf{C}_U\breve{\mathbf{y}}_U.$$

En d'autres termes, $\tilde{Y}_{opt}$ peut être réécrit comme un estimateur par la régression généralisée dans lequel les deux groupes de variables auxiliaires sont utilisés dans le coefficient de régression : les variables d'équilibrage et les variables de calage. Ce résultat simplifie considérablement le calcul de l'estimateur optimal. L'estimateur $\tilde{Y}_{opt}$ paraît indépendant des probabilités $p_{k\ell}$. En fait, l'effet des $p_{k\ell}$ est dissimulé dans la régression grâce aux variables d'équilibrage.

Comme $\mathbf{b}_{zx-regr,}$ est un paramètre de la population, il doit être estimé. On propose de l'estimer par $\hat{\mathbf{b}}_{zx-regr,}$ qui est composé des M premiers éléments du vecteur

$$\left[\left(\breve{\mathbf{X}}_s, \breve{\mathbf{Z}}_s\right)'\breve{\mathbf{C}}_s\left(\breve{\mathbf{X}}_s, \breve{\mathbf{Z}}_s\right)\right]^{-}\left(\breve{\mathbf{X}}_s, \breve{\mathbf{Z}}_s\right)'\breve{\mathbf{C}}_s\breve{\mathbf{y}}_s,$$

où $\breve{\mathbf{C}}_s$ est la matrice diagonale avec c_k / p_k comme éléments diagonaux.

En plus de la simplicité d'utilisation de $\tilde{Y}_{opt}$, les simulations qui ont été effectuées montrent que cet estimateur est plus précis que celui de Montanari et que l'utilisation des vraies valeurs des probabilités d'inclusion d'ordre deux donnent des estimateurs moins précis que l'estimateur GREG basé sur $\hat{\mathbf{b}}_{zx-regr}$. On peut donc conclure que la connaissance des probabilités d'inclusion d'ordre deux est inutile pour construire un estimateur GREG optimal quand les variables d'équilibrage sont utilisées dans notre estimateur GREG.

5. Conclusion

On propose une approximation de l'estimateur asymptotiquement optimal de Montanari (1987). L'estimateur proposé est un estimateur GREG qui utilise, en même temps, les variables d'équilibrage et les variables de calage comme variables de régression. On a seulement besoin d'utiliser les poids c_k / p_k au moment du calcul des coefficients de régression.

En ce qui concerne les probabilités d'inclusion d'ordre deux, nous avons montré qu'elles correspondent à une correction due à l'absence des variables d'équilibrage dans l'estimateur GREG. En utilisant les variables d'équilibrage comme variables de régression, une connaissance exacte des probabilités d'inclusions d'ordre deux n'est plus nécessaire.

Bibliographie

[1] Berger, Y.G., Tirari, M.E.H., Tillé, Y., « Toward optimal regression estimation in sample surveys », *Australian and New Zealand Journal of Statistics*, 45, pp. 319-329, 2003.

[2] Deville, J.C., Tillé, Y., « Efficient balanced sampling : The cub method », *Biometrika*, 91, p. 893-912, 2004.

[3] Montanari, G.E., « Post sampling efficient QR-prediction in large sample survey », *International Statistical Review*, 55, 191-202, 1987.

[4] Särndal, C.E., Swenson, B., Wretman, J.H., *Model Assisted Survey Sampling,* New York, Springer-Verlag, 1992.