Estimation non-paramétrique par des B-splines pour des paramètres non-linéaires

Camelia GOGA[1] *et Anne Ruiz-GAZEN*[2]

Résumé : Ce travail concerne l'estimation de paramètres non-linéaires de totaux tels que le ratio, le coefficient de régression ou de corrélation, quand l'information auxiliaire est disponible pour chaque individu d'une population finie. Nous proposons une nouvelle classe d'estimateurs par substitution qui consiste à remplacer chaque total par un estimateur assisté par un modèle de superpopulation fondé sur une régression non-paramétrique par des fonctions B-splines (Goga, 2005). La statistique complexe ainsi obtenue est une fonction non-linéaire des estimateurs pondérés avec des poids calés sur les fonctions B-splines et indépendants des variables d'intérêt. Nous dérivons la variance asymptotique de l'estimateur par substitution en utilisant la méthode de linéarisation par la fonction d'influence développée par Deville (1999).

INTRODUCTION

Soit une population finie U de taille N et une variable auxiliaire Z connue pour chaque individu de la population. Notons par z_k la valeur de Z pour l'individu $k, k \in U$. Un échantillon s est sélectionné dans U selon un plan $p(s)$. Les probabilités d'inclusion de premier et second degrés sont notées π_k et π_{kl}. Considérons par exemple deux variables d'intérêt X et Y, l'objectif de cette étude est de prendre en compte Z dans l'estimation d'un paramètre non-linéaire des totaux t_X et t_Y : $\Phi(t_X, t_Y)$.

Nous introduisons un modèle qui explique la relation entre la variable d'intérêt X, respectivement Y, et la variable auxiliaire Z. Chaque total est estimé par un estimateur par la différence généralisée (Cassel *et al.*, 1976). Les fonctions de régression sont estimées dans un premier temps par leurs meilleures approximations dans l'espace de fonctions spline. Les fonctions spline d'ordre $m \geq 2$ avec K nœuds intérieurs sont des fonctions $m-2$ continûment dérivables et polynômiales par morceaux. Cet espace est de dimension $q = K + m$ et une base est constituée de fonctions B-splines $B_1, \ldots, B_q$ (Schumaker, 1981). Les projections des fonctions de régression sont ensuite estimées à l'aide du plan d'échantillonnage. On note : $b'(z_k) = \left(B_j(z_k)\right)_{j=1}^{q}$.

[1] IMB, Université de Bourgogne, 9 avenue Alain Savary, 21078 Dijon, France
- camelia.goga@u-bourgogne.fr
[2] Toulouse School of Economics, Université Toulouse 1, 21 allée de Brienne, 31000 Toulouse
- ruiz@cict.fr

On obtient : $\hat{t}_{X,BS} = \sum_s w_{ks} x_k$ et : $\hat{t}_{Y,BS} = \sum_s w_{ks} y_k$, pour des poids $w_{ks} = \frac{1}{\pi_k}\left(\sum_{i\in U} b'(z_i)\right)\left(\sum_{i\in s}\frac{b(z_i)b'(z_i)}{\pi_i}\right) b(z_k)$ indépendants des variables d'intérêt X et Y et qui contiennent l'information auxiliaire Z. Ces poids satisfont également les équations de calage : $\sum_s w_{ks}\, b(z_k) = \sum_U b(z_k)$.

L'estimateur par substitution de Φ est :

$$\hat{\Phi}_{BS} = \Phi\left(\hat{t}_{X,BS}, \hat{t}_{Y,BS}\right).$$

Pour obtenir la variance asymptotique de $\hat{\Phi}_{BS}$, nous proposons d'utiliser la méthode de linéarisation par la fonction d'influence (Deville, 1999). Soit u_k la variable linéarisée de Φ du total $t_U = \sum_U u_k$ estimé par : $\hat{t}_{U,BS} = \sum_s w_{ks} u_k$.

Résultat : Sous certaines conditions, l'estimateur $\hat{\Phi}_{BS}$ satisfait :

$$N^{-\alpha}\left(\hat{\Phi}_{BS} - \Phi\right) = N^{-\alpha}\left(\hat{t}_{U,BS} - t_U\right) + o_p\left(n^{-\frac{1}{2}}\right) = N^{-\alpha}\left(\hat{t}_{diff,u} - t_U\right) + o_p\left(n^{-\frac{1}{2}}\right),$$

avec : $\hat{t}_{diff,u} = \sum_{k\in s} \frac{u_k - b'(z_k)\hat{\theta}_u}{\pi_k} + \sum_{k\in U} b'(z_k)\hat{\theta}_u$ et : $\hat{\theta}_u = \left(\sum_U b(z_k)\right)^{-1} \sum_U b(z_k) u_k$.

L'estimateur $\hat{\Phi}_{BS}$ sera plus efficace que l'estimateur par substitution utilisant des estimateurs de type Horvitz-Thompson des totaux, en termes de variance asymptotique, si la variable auxiliaire explique bien la variable linéarisée, ce qui revient à dire que les résidus $u_k - b'(z_k)\hat{\theta}_u$ sont « petits » pour tout $k \in U$.

Réferences

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population, *Biometrika*, **63**, 615-620.

Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey Methodology*, **25,** 193-203.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche non paramétrique par splines de régression, *The Canadian Journal of Statistics*, **33,** 1-18.

Schumaker, L. L. (1981). Spline functions: basic theory. Wiley, New-York.