Estimation de paramètres non linéaires par des méthodes non-paramétriques en population finie

Camelia Goga, Anne Ruiz-Gazen

Estimation de paramètres non linéaires par des méthodes non-paramétriques en population finie

CAMELIA GOGA & ANNE RUIZ-GAZEN

IMB, Université de Bourgogne, 9 Avenue Alain Savary, 21078 DIJON, France
camelia.goga@u-bourgogne.fr
Toulouse School of Economics (Gremaq and IMT),
Université Toulouse 1, 21 allée de Brienne, 31000 Toulouse, France
ruiz@cict.fr

Abstract

In this paper we consider the estimation of nonlinear finite population parameters when auxiliary information is available for each individual from the finite population. We propose a new class of substitution estimators obtained by replacing each total with model-assisted estimators based on nonparametric regression. In order to obtain the asymptotic variance, the complex statistic obtained is linearized using the influence function approach proposed by Deville (1999).

Résumé

Nous considérons dans cet article l'estimation de paramètres non-linéaires de totaux en population finie quand une variable auxiliaire est disponible pour chaque individu de la population. Une nouvelle classe d'estimateurs par substitution est obtenue en remplaçant chaque total par un estimateur assisté par un modèle et basé sur une régression non-paramétrique. Pour obtenir la variance asymptotique, la statistique complexe obtenue est ensuite linéarisée par la technique de la fonction d'influence proposée par Deville (1999).

1 Introduction

L'estimation de paramètres non-linéaires Φ tels que le ratio, le coefficient de régression ou de corrélation, les indices économiques (indice de Gini, de pauvreté) sont l'objet de nombreuses enquêtes par sondages. L'estimation peut être améliorée si l'information auxiliaire est disponible pour chaque individu de la population. Nous proposons dans ce papier une nouvelle classe d'estimateurs pour prendre en compte l'information auxiliaire à l'aide d'un modèle non-paramétrique.

2 Estimation d'un total par régression non-paramétrique

On considère une population finie $U = \{1, \ldots, k, \ldots, N\}$ et on suppose connues les valeurs d'une variable auxiliaire unidimensionnelle $\mathcal{Z}$ pour toutes les unités k dans U; on note ces valeurs z_k pour $k \in \{1, \ldots, N\}$. Un échantillon s de taille fixe n est sélectionné dans U selon un plan de sondage non-informatif quelconque $p(s)$ et la valeur de la variable d'intérêt $\mathcal{Y}$ est observée pour chaque unité dans l'échantillon; on obtient y_k pour $k \in s$. Pour chaque individu $k \in U$, la probabilité d'inclusion dans s, $\pi_k = Pr(k \in s)$, est supposée strictement positive; de même, $\pi_{kl} = Pr(k, l \in s) > 0$ pour tous $k, l \in U$. Pour chaque individu k dans la population, on note $I_k = \mathbf{1}_{\{k \in s\}}$ la variable indicatrice d'appartenance à l'échantillon s. Nous voulons estimer le total t_y d'une variable $\mathcal{Y}$ sur U, $t_y = \sum_{k \in U} y_k$. En l'absence d'information auxiliaire, ce total est estimé par l'estimateur de Horvitz-Thompson, $\hat{t}_{y,HT} = \sum_s \frac{y_k}{\pi_k}$. En présence d'information auxiliaire, on peut introduire un modèle de superpopulation

$$\xi : \quad y_k = f(z_k) + \varepsilon_k \tag{1}$$

et considérer la classe d'estimateurs assistés par un modèle introduite par Cassel *et al.* (1976),

$$\hat{t}_y \;=\; \sum_{k \in s} \frac{y_k - f_k}{\pi_k} + \sum_{k \in U} f_k \tag{2}$$

avec $f_k = f(z_k)$. Dans le cas d'un modèle linéaire, on obtient la classe d'estimateurs GREG présentée par Särndal *et al.*, 1992. Néanmoins, si la vraie relation n'est pas linéaire, l'efficacité, en terme de variance, de l'estimateur par la régression généralisée $\hat{t}_{GREG}$ sous un modèle linéaire peut se révéler mauvaise, même comparée à l'estimateur de Horvitz-Thompson qui pourtant ne prend pas en compte l'information auxiliaire.
L'utilisation de modèles non-paramétriques permet de couvrir une classe beaucoup plus large de relations entre information auxiliaire et variable d'intérêt, en imposant uniquement des conditions de régularité (dérivabilité) sur la fonction de régression. Contrairement au modèle linéaire, les estimateurs basés sur des modèles nonparamétriques nécessitent que l'information auxiliaire soit connue sur toute la population (ce qui permet d'utiliser aussi la variable auxiliaire pour établir le plan de sondage).
Récemment, Breidt & Opsomer (2000) ont proposé des estimateurs assistés par un modèle non-paramétrique en utilisant une approche par polynômes locaux pour des plans de sondage à une ou deux phases. Ultérieurement, Breidt and Opsomer (2005) ont utilisé une décomposition de la fonction de régression dans une base de fonctions splines des polynômes tronqués (P-splines) et Goga (2005) a utilisé une décomposition de f_k dans une base de $B-$splines.

Soit f_k estimé par $\hat{f}_{y,k}$ qui dépend de la variable $\mathcal{Y}$ lorsqu'on utilise une méthode non-paramétrique telle que les polynômes locaux, les P ou les B-splines. Si on remplace dans (2) les f_k par $\hat{f}_{y,k}$, on obtient un estimateur pour le total t_y qui est toujours p-sans biais mais approximativement ξ-sans biais :

$$\hat{t}_{y,\text{diff}} = \sum_{k\in s} \frac{y_k - \hat{f}_{y,k}}{\pi_k} + \sum_{k\in U} \hat{f}_{y,k}. \tag{3}$$

Les $\hat{f}_{y,k}$ étant inconnus pour $k \in U - s$, ils sont estimés par $\tilde{f}_{y,k}$ obtenus à l'aide du plan d'échantillonnage p qui fournit s et on obtient

$$\hat{t}_{y,np} = \sum_{k\in s} \frac{y_k - \tilde{f}_{y,k}}{\pi_k} + \sum_{k\in U} \tilde{f}_{y,k}. \tag{4}$$

Ces trois estimateurs assistés par un modèle non-paramétrique (polynômes locaux, P et B-splines) peuvent s'écrire comme une somme pondérée des valeurs de $\mathcal{Y}$ avec des poids indépendants de la variable d'intérêt et contenant l'information auxiliaire,

$$\hat{t}_{y,np} = \sum_s w_{ks} y_k \tag{5}$$

où l'expression de w_{ks} dépend de la méthode utilisée.

Résultat 1 *Sous certaines hypothèses (voir Breidt & Opsomer, 2000, 2005 et Goga, 2005), l'estimateur $\hat{t}_{y,np}$ satisfait*

$$\frac{1}{N}(\hat{t}_{y,np} - t_y) = O_p(n^{-1/2}) \quad \textit{et}$$

$$n^{1/2}N^{-1}(\hat{t}_{y,np} - t_y) = n^{1/2}N^{-1}(\hat{t}_{y,diff} - t_y) + o_p(1). \tag{6}$$

Par conséquent, il est asymptotiquement sans biais et convergent pour t_y. Si la distribution asymptotique de $\sqrt{n}N^{-1}(\hat{t}_{y,diff} - t_y)$ est normale, alors la variance asymptotique de $n^{1/2}N^{-1}(\hat{t}_{y,np} - t_y)$ est donnée par la variance de $n^{1/2}N^{-1}(\hat{t}_{y,diff} - t_y)$, dont l'expression est donnée par :

$$\frac{n}{N^2} \sum_U \sum_U \Delta_{kl} \frac{y_k - \hat{f}_k}{\pi_k} \frac{y_k - \hat{f}_l}{\pi_l}.$$

La variance asymptotique est d'autant plus petite que les résidus $y_k - \hat{f}_k$ sont petits en valeur absolue. Ce résultat justifie l'usage des techniques non-paramétriques qui permettent de fournir de bons estimateurs d'une large classe de fonctions f qui ne sont pas toujours paramétrisables simplement.

L'estimateur $\hat{t}_{y,np}$ dans le cas d'une régression par des P ou B splines possède la plupart des propriétés des estimateurs GREG sous un modèle linéaire, notamment le fait que l'estimateur de Horvitz-Thompson pour les résidus $y_k - \tilde{f}_k$ est nul. Par conséquent, $\hat{t}_{y,np} = \sum_{k\in U} \tilde{f}_k$ dans ce cas.

3 Estimation d'un paramètre non-linéaire par régression non-paramétrique

Nous souhaitons prendre en compte l'information auxiliaire $\mathcal{Z}$ pour l'estimation d'un paramètre Φ fonction non-linéaire de totaux. Pour simplifier, nous supposons que Φ est une fonction de seulement deux totaux, $\Phi = \Phi(t_x, t_y)$.
Les poids w_{ks} donnés dans la relation (5) ont été obtenus en utilisant $\mathcal{Z}$ pour estimer le total t_y. Ces poids ne dépendent pas de la variable d'intérêt et par conséquent, ils peuvent être utilisés pour estimer d'autres totaux. Dans la suite, nous estimons les totaux t_x et t_y par des estimateurs pondérés avec les mêmes poids w_{ks},

$$\hat{t}_{x,np} = \sum_{k\in s} w_{ks}x_k, \quad \hat{t}_{y,np} = \sum_{k\in s} w_{ks}y_k.$$

L'estimateur par substitution de Φ est donné par :

$$\widehat{\Phi}_{np} = \Phi(\hat{t}_{x,np}, \hat{t}_{y,np}). \tag{7}$$

Pour calculer la variance asymptotique de $\widehat{\Phi}_{np}$, on utilise l'approche par linéarisation basée sur la fonction d'influence (Deville, 1999). On considère la mesure discrète $M = \sum_U \delta_{(x_k,y_k)}$ et on suppose que Φ peut s'écrire comme une fonctionnelle T de M, $\Phi = T(M)$. La mesure M est estimée par $\widehat{M}_{np} = \sum_s w_{ks}\delta_{(x_k,y_k)}$ avec des poids w_{ks} donné par (5). L'estimateur par substitution non-paramétrique $\widehat{\Phi}_{np}$ donné par la relation (7) est obtenu en remplaçant M avec $\hat{M}_{np}$,

$$\widehat{\Phi}_{np} = T(\widehat{M}_{np}). \tag{8}$$

Pour obtenir la variance asymptotique de $\widehat{\Phi}_{np}$, nous faisons un développement au premier ordre de la fonctionnelle T. La différentielle de T s'appelle fonction d'influence et elle est définie ci-dessous.

Definition 1 *: La fonction d'influence $IT(M,x)$ de $T(M)$ est définie comme la dérivée au sens de Gateaux de T par rapport à M dans la direction de la masse de Dirac en x,*

$$IT(M,x) = \lim_{\varepsilon \longrightarrow 0} \frac{T(M+\varepsilon\delta_x) - T(M)}{\varepsilon}$$

lorsque cette limite existe.

Definition 2 *La variable linéarisée u_k pour $k \in U$ est la valeur de IT dans (x_k, y_k),*

$$u_k = IT(M, (x_k, y_k)), \quad k \in U.$$

Soient les estimateurs $\hat{t}_{u,\text{diff}}$ et $\hat{t}_{u,np}$ obtenus en remplaçant dans les expressions (3) and (4) les valeurs y_k de la variable $\mathcal{Y}$, par celles des variables linéarisées, u_k. Le résultat suivant donne une linéarisation de la statistique complexe $\widehat{\Phi}_{np}$ par un estimateur par la différence généralisée du total des variables linéarisées, $t_u = \sum_U u_k$.

Résultat 2 *Supposons que la fonctionnelle T est dérivable au sens de Fréchet et de degré α, c'est à dire $T(rM) = r^\alpha T(M)$ et de plus $\lim_{N\to\infty} T(M/N) < \infty$. Supposons que les variables $N^{1-\alpha}u_k$ satisfont la relation (6), c'est à dire $N^{-\alpha}(\hat{t}_{u,np} - \hat{t}_{u,diff}) = o_p(n^{-1/2})$. Alors,*

$$N^{-\alpha}\left(\widehat{\Phi}_{np} - \Phi\right) \;=\; N^{-\alpha}(\hat{t}_{u,np} - t_u) + o_p(n^{-1/2}) = N^{-\alpha}(\hat{t}_{u,diff} - t_u) + o_p(n^{-1/2}).$$

Supposons que la distribution de $N^{-\alpha}(\hat{t}_{u,diff} - t_u)$ est normale, alors la variance asymptotique de $\sqrt{n}N^{-\alpha}\left(\widehat{\Phi}_{np} - \Phi\right)$ est donnée par $\dfrac{\sqrt{n}}{N^{2\alpha}}\displaystyle\sum_{k\in U}\sum_{i\in U}\Delta_{kl}\frac{u_k - \hat{f}_{u,k}}{\pi_k}\frac{u_l - \hat{f}_{u,l}}{\pi_l}$ avec $\hat{f}_{u,k}$ donné par la relation (3) pour la variable u.

Les conditions dans lesquelles l'approximation $N^{-\alpha}(\hat{t}_{u,np} - \hat{t}_{u,\text{diff}}) = o_p(n^{-1/2})$ est valable, dépendent du type d'estimateur non-paramétrique utilisé (par polynômes locaux, par P ou B-splines).
La variance asymptotique donnée par le résultat 2 sera d'autant plus petite que les résidus $u_k - \hat{f}_{u,k}$ serons petits ou que le modèle explique bien la variable linéarisée. Ce résultat surprend un peu car la variable linéarisée ou son total n'est pas le but d'une enquête.

4 Estimation d'un ratio par régression non-paramétrique basée sur des B-splines

Considérons maintenant l'estimation d'un ratio $R = \dfrac{t_y}{t_x}$ quand une variable auxiliaire $\mathcal{Z}$ est disponible pour chaque individu dans la population. Cette situation a déjà été traitée par Särndal *et al.* (1992) en utilisant un modèle de régression multiple pour améliorer l'estimation des totaux t_x et t_y. Nous proposons ici une alternative en introduisant un modèle non paramétrique et une estimation par des B-splines.
Soit $B_1, \ldots, B_q$ une base de B-spline et les poids w_{ks} qui donnent l'estimateur $\hat{t}_{y,np}$ dans la relation (4) deviennent

$$w_{ks} = \frac{1}{\pi_k}\left(\sum_U \boldsymbol{b}'(z_i)\right)\left(\sum_{i\in s}\boldsymbol{b}(z_i)\boldsymbol{b}'(z_i)/\pi_i\right)^{-1}\boldsymbol{b}(z_k), \tag{9}$$

pour $\boldsymbol{b}'(z_i) = (B_1(z_k), \ldots, B_q(z_k))$. Nous utilisons ces poids pour estimer le ratio R par

$$\hat{R}_{BS} = \frac{\sum_s w_{ks} y_k}{\sum_s w_{ks} x_k} = \frac{\sum_{k\in s}(y_k - \boldsymbol{b'}(z_k)\tilde{\boldsymbol{\theta}}_y)/\pi_k + \sum_U \boldsymbol{b'}(z_k)\tilde{\boldsymbol{\theta}}_y}{\sum_{k\in s}(x_k - \boldsymbol{b'}(z_k)\tilde{\boldsymbol{\theta}}_x)/\pi_k + \sum_U \boldsymbol{b'}(z_k)\tilde{\boldsymbol{\theta}}_x}$$

$$\text{avec } \tilde{\boldsymbol{\theta}}_y = \left(\sum_s \boldsymbol{b}(z_k)\boldsymbol{b'}(z_k)/\pi_k\right)^{-1}\left(\sum_s (\boldsymbol{b}(z_k)y_k)/\pi_k\right)$$

$$\text{et } \tilde{\boldsymbol{\theta}}_x = \left(\sum_s (\boldsymbol{b}(z_k)\boldsymbol{b'}(z_k))/\pi_k\right)^{-1}\left(\sum_s (\boldsymbol{b}(z_k)x_k)/\pi_k\right).$$

La variable linéarisée associée à R est $u_k = \frac{1}{t_x}(y_k - Rx_k)$ et en utilisant le résultat (2), la variance asymptotique de $\hat{R}_{BS}$ est :

$$\sum_{k\in U}\sum_{i\in U} \Delta_{ki} \frac{u_k - \hat{f}_u(z_k)}{\pi_k}\frac{u_l - \hat{f}_u(z_l)}{\pi_l} \quad \text{avec} \quad \hat{f}_u(z_k) = \boldsymbol{b'}(z_k)\left(\sum_U \boldsymbol{b}(z_k)\boldsymbol{b'}(z_k)\right)^{-1}\sum_U \boldsymbol{b}(z_k)u_k.$$

Bibliographie

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

[2] Breidt, F.J. and Opsomer, J. (2000), Local Polynomial Regression Estimators in Survey Sampling, *The Annals of Statistics*, **28**, 1026-1053.

[3] Breidt, F.J. and Opsomer, J. (2005), Model-assisted estimation for complex surveys using penalised splines, *Biometrika*, **92**, 831-846.

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

[5] 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.*

[6] Särndal C.E. , Swensson B. and Wretman J. (1992), Model Assisted Survey Sampling, Springer, Berlin.