Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 339 (2004) 441–444
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Estimation de l'évolution d'un total en présence d'information auxiliaire : une approche par splines de régression

Camelia Goga

Laboratoire de statistique et probabilités, université Paul Sabatier, 31062 Toulouse cedex, France

Reçu le 11 mars 2004 ; accepté après révision le 12 juillet 2004

Disponible sur Internet le 27 août 2004

Présenté par Paul Deheuvels

Résumé

Nous construisons un estimateur pour l'évolution d'un total sur deux occasions différentes en présence d'information auxiliaire. Un modèle de superpopulation est introduit afin d'expliquer la relation entre les variables d'intérêt et les variables auxiliaires. Les fonctions de régression sont estimées à l'aide de splines de régression et par la technique de Horvitz–Thompson. Finalement, on construit un estimateur pour l'évolution qui est asymptotiquement sans biais et convergent et on donne une formule pour la variance sous le plan de sondage. ***Pour citer cet article : C. Goga, C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

Abstract

Finite population total evolution estimation using auxiliary information: a regression spline approach. We build an estimator for the evolution of a finite population total between two periods of time when auxiliary information is available. A superpopulation model is introduced in order to explain the relationship between the study and the auxiliary variables. The regression functions are estimated by regression splines and Horvitz–Thompson technique. Finally, an estimator for the evolution is derived and proved to be asymptotically unbiased and consistent and we compute a design-based variance formula. ***To cite this article: C. Goga, C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

1. Introduction

Le problème de l'estimation d'une combinaison linéaire de totaux quand l'information provient de plusieurs échantillons apparaît dans beaucoup de situations pratiques. L'exemple le plus naturel est celui où deux échantillons correspondent à des sondages effectués à des instants différents du temps. C'est un fait bien connu que

Adresse e-mail : goga@math.ups-tlse.fr (C. Goga).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2004.07.011

le bon usage de l'information auxiliaire améliore la précision des estimateurs dans les enquêtes. Il existe dans la littérature plusieurs approches pour utiliser l'information auxiliaire dans la phase d'estimation : les approches 'model-assisted' [6] et 'predictive' [5] en introduisant un modèle de superpopulation ou bien encore l'approche 'calage' [2].

On va se placer dans la suite dans le cadre de l'approche 'model-assisted' où la relation entre la variable d'intérêt y_k et la variable auxiliaire x_k est de la forme

$$y_k = f(x_k) + \varepsilon_k, \tag{1}$$

la fonction de régression f étant inconnue. On propose d'étendre l'approche par splines de régression introduite par [4] au cas de l'estimation de l'évolution d'un total sur deux périodes différentes de temps. L'estimateur ainsi construit est asymptotiquement sans biais et convergent. Une formule de type Horvitz–Thompson pour la variance sous le plan de sondage est obtenue.

2. Le modèle

On considère une population finie $U = \{1, \ldots, k, \ldots, N\}$ et on s'intéresse à l'estimation de l'évolution du total d'une variable d'intérêt $\mathcal{Y}$ mesurée à deux instants différents T_1 et T_2. On note pour cela $\mathcal{Y}_1$ la variable mesurée à l'instant T_1 et $\mathcal{Y}_2$ à l'instant T_2 avec y_{kt} la valeur de $\mathcal{Y}_t$ pour le k-ème individu et $t = 1, 2$. Nous cherchons à estimer

$$\Delta Y = Y_1 - Y_2 = \sum_U y_{k1} - \sum_U y_{k2}.$$

On suppose connues les valeurs d'une variable auxiliaire unidimensionnelle $\mathcal{X}^{(t)}$ pour toutes les unités k dans U et pour les deux instants ; on note ces valeurs $x_k^{(t)}$ pour $k \in \{1, \ldots, N\}$. Un échantillon bidimensionnel, noté $\mathbf{s} = (s_1, s_2)$, de taille fixe $n_{\mathbf{s}}$ est sélectionné dans $U \times U$ selon un plan de sondage bidimensionnel quelconque $p(\mathbf{s})$ [4]. Les échantillons s_1 et s_2 ne sont pas nécessairement indépendants. La variable d'intérêt $\mathcal{Y}_t$ est observée pour chaque unité dans l'échantillon s_t. Relativement au plan bidimensionnel $p(\mathbf{s})$, nous avons les probabilités d'inclusion bidimensionnelles de premier degré notées $\pi_k^s = Pr(k \in s)$ pour $k \in U$ et s un des échantillons $s_{12} = s_1 \cap s_2$, $s_{1*} = s_1 \setminus s_{12}$ ou $s_{2*} = s_2 \setminus s_{12}$ et de deuxième degré notées $\pi_{ik}^{s,\tau} = Pr(i \in s, k \in \tau)$ pour tous $i, k \in U$ et $s, \tau \in \{s_{1*}, s_{12}, s_{2*}\}$. Nous supposons dans la suite que $\pi_k^s > 0$ pour tous $k \in U$ et $\pi_{ik}^{s,\tau} > 0$ pour tous $i \neq k \in U$. Pour chaque individu k dans la population, on note $I_k^s = \mathbf{1}_{\{k \in s\}}$ la variable indicatrice d'appartenance à l'échantillon $s \in \{s_{1*}, s_{12}, s_{2*}\}$ et n_s la taille de s (pour plus de détails voir [4]). Pour alléger les notations, seuls les indices $1*, 12, 2*$ seront utilisés. L'espérance et la variance par rapport au plan p seront notées E_p et V_p.

On suppose que la population finie U peut être vue comme un échantillon dans une superpopulation infinie ξ dans laquelle la relation entre les variables y_{kt} et x_{kt} est donnée par un modèle de type (1) où $f^{(t)}$ est une fonction inconnue et les erreurs $\varepsilon_k^{(t)}$, $k \in U$ sont des variables aléatoires indépendantes, d'espérance nulle et de variance $v(x_{kt}) = v_k^{(t)}$. En revanche, $\varepsilon_k^{(1)}$ et $\varepsilon_k^{(2)}$ ne sont pas obligatoirement indépendantes. On suppose que les $x_{kt} \in [0, 1]$ pour $k \in U$ et $t = 1, 2$.

3. L'estimateur *B*-splines et l'estimateur pour l'évolution

On construit des estimateurs pour $f^{(t)}$ par splines de régression. Alors, pour $t = 1, 2$ définissons l'espace des fonctions splines d'ordre $m^{(t)}$ ($m^{(t)} \geqslant 2$) avec $K^{(t)}$ noeuds intérieurs équidistants $0 = \xi_0^{(t)} < \xi_1^{(t)} < \cdots < \xi_{K^{(t)}}^{(t)} < \xi_{K^{(t)}+1}^{(t)} = 1$. On note

$$S_{K^{(t)},m^{(t)}} = \big\{u \in C^{m^{(t)}-2}[0,1]\text{: } u(x) \text{ est un polynôme de degré } m^{(t)}-1 \text{ sur } (\xi_j^{(t)}, \xi_{j+1}^{(t)})\big\}.$$

L'espace $S_{K^{(t)},m^{(t)}}$ est un espace de dimension $q^{(t)} = K^{(t)} + m^{(t)}$ dont une base est constituée des fonctions B-splines $(B_j^{(t)}(\cdot))_{j=1}^{q^{(t)}}$ [3]. On estime $f^{(t)}$ par l'estimateur des moindres carrés $\hat{f}^{(t)}(x) \in S_{K^{(t)},m^{(t)}}$ pour $t = 1, 2$. Pour $f^{(t)}(x_{kt}) = f_k^{(t)}$, $k \in U$ et $\boldsymbol{b}'^{(t)} = (B_1^{(t)}(x_{kt}), \ldots, B_{q^{(t)}}^{(t)}(x_{kt}))$, l'estimateur par B-splines de $f_k^{(t)}$ s'écrit pour chaque x_{kt}, $k \in U$:

$$\begin{cases} \hat{f}_k^{(t)} = \boldsymbol{b}'^{(t)}\hat{\boldsymbol{\theta}}^{(t)}, \quad k \in U, \\ \hat{\boldsymbol{\theta}}^{(t)} = \Big(\sum_U \boldsymbol{b}^{(t)}\boldsymbol{b}'^{(t)}\Big)^{-1}\Big(\sum_U \boldsymbol{b}^{(t)} y_{kt}\Big), \end{cases} \tag{2}$$

en supposant que la matrice $\sum_U \boldsymbol{b}^{(t)}\boldsymbol{b}'^{(t)}$ est inversible.

Pour estimer un total Y sur un seul échantillon s, [1] considèrent l'estimateur par la différence généralisée $\hat{Y}_{\text{diff}} = \sum_s \frac{y_k - f_k}{\pi_k} + \sum_U f_k$ avec f_k donné par (1) et obtenu par la différence entre l'estimateur de Horvitz–Thompson $\widehat{Y}_{\text{HT}} = \sum_s \frac{y_k}{\pi_k}$ et son biais par rapport au modèle. En particulier, $\widehat{Y}_{\text{diff}}$ est p-sans biais et ξ-sans biais. Comme f_k est inconnu, il est remplacé par $\hat{f}_k$ estimé à l'aide du modèle (1) par splines de régression dans notre situation. Malheureusement, les $\hat{f}_k$ donnés par (2) sont toujours inconnus, ils dépendent des valeurs de la variable d'intérêt dans toute la population. Par conséquent, ils sont estimés par $\hat{\hat{f}}_k$, les estimateurs de Horvitz–Thompson de $\hat{f}_k$ sur s. Pour estimer la variation ΔY quand l'information provient de deux échantillons s_1 et s_2, [4] construit l'estimateur composite pour α et β deux constantes entre 0 et 1,

$$\widehat{\Delta Y}_{\text{HT}} = \alpha \sum_{s_{1*}} \frac{y_{k1}}{\pi_k^{1*}} + (1-\alpha)\sum_{s_{12}} \frac{y_{k1}}{\pi_k^{12}} - \beta \sum_{s_{2*}} \frac{y_{k2}}{\pi_k^{2*}} - (1-\beta)\sum_{s_{12}} \frac{y_{k2}}{\pi_k^{12}},$$

ce qui nous donne facilement $\widehat{\Delta Y}_{\text{diff}}$ par soustraction du biais de $\widehat{\Delta Y}_{\text{HT}}$ par rapport au modèle. Ensuite, les $f_k^{(t)}$ sont remplacés par $\hat{f}_k^{(t)}$ et on obtient

$$\widehat{\Delta Y}_{\text{reg}} = \alpha \widehat{Y}_{1,\text{reg}}^{1*} + (1-\alpha)\widehat{Y}_{1,\text{reg}}^{12} - \beta \widehat{Y}_{2,\text{reg}}^{2*} - (1-\beta)\widehat{Y}_{2,\text{reg}}^{12}, \tag{3}$$

où $\widehat{Y}_{1,\text{reg}}^{s} = \sum_s \frac{y_{k1} - \hat{f}_k^{(1)}}{\pi_k^s} + \sum_U \hat{f}_k^{(1)}$ avec $\hat{f}_k^{(1)}$ donné par (2) et $s \in \{s_{1*}, s_{12}\}$. On construit de la même manière $\widehat{Y}_{2,\text{reg}}^{s}$ avec $\hat{f}_k^{(2)}$ et $s \in \{s_{12}, s_{2*}\}$. L'évolution est finalement estimée en remplaçant dans (3) les $\hat{f}_k^{(t)}$ par $\hat{\hat{f}}_k^{(t),s} = \boldsymbol{b}'^{(t)}\hat{\hat{\boldsymbol{\theta}}}_s^{(t)} = \boldsymbol{b}'^{(t)}((\sum_s \boldsymbol{b}^{(t)}\boldsymbol{b}'^{(t)})/\pi_k^s)^{-1}((\sum_s \boldsymbol{b}^{(t)} y_{kt})/\pi_k^s)$ pour $s \in \{s_{1*}, s_{12}\}$ si $t = 1$ et $s \in \{s_{2*}, s_{12}\}$ si $t = 2$. On obtient pour $\widehat{Y}_{t\pi}^{s} = \sum_s \frac{y_{kt} - \hat{\hat{f}}_k^{(t),s}}{\pi_k^s} + \sum_U \hat{\hat{f}}_k^{(t),s}$ et notre estimateur de l'évolution

$$\widehat{\Delta Y}_{\pi} = \alpha \widehat{Y}_{1\pi}^{1*} + (1-\alpha)\widehat{Y}_{1\pi}^{12} - \beta \widehat{Y}_{2\pi}^{2*} - (1-\beta)\widehat{Y}_{2\pi}^{12}. \tag{4}$$

4. Résultats

Nous présentons dans cette section les propriétés asymptotiques de l'estimateur. Pour cela nous devons supposer que les tailles de la population U et des échantillons s_1, s_2 deviennent de plus en plus grandes [4].

C1 $\lim_{N\to\infty} \frac{n_s}{N} = \pi_s \in (0,1)$ pour $s \in \{s_{1*}, s_{12}, s_{2*}\}$.
C2 La fonction $f^{(t)}$ est $m^{(t)}$ fois continûement dérivable sur $[0,1]$.
C3 $\frac{1}{N}\sum_{k\in U} y_{kt}^2 < \infty$ avec ξ probabilité 1 [6].

C4 $\min_{k\in U}\pi_k^s \geqslant \lambda > 0$, $\min_{i,k\in U}\pi_{ik}^s \geqslant \lambda* > 0$, $\overline{\lim}_{N\to\infty} n_s \max_{i\neq k\in U}|\pi_{ik}^s - \pi_i^s\pi_k^s| < \infty$ pour $s \in \{s_{1*}, s_{12}, s_{2*}\}$.

C5 Il existe une fonction de répartition $Q^{(t)}(x)$ admettant une densité strictement positive sur $[0,1]$ telle que $\sup_{x\in[0,1]}|Q_N^{(t)}(x) - Q^{(t)}(x)| = \mathrm{o}(1/K^{(t)})$ où $Q_N^{(t)}(x)$ est la distribution empirique de $(x_{kt})_{k=1}^N$ pour $t = 1, 2$.

Lemme 4.1. *Sous les conditions* C1, C3 *et* C4, *l'estimateur* $\widehat{\Delta Y}_{\mathrm{HT}}$ *est asymptotiquement sans biais et convergent vers* ΔY.

Lemme 4.2. *Sous les conditions* C1, C3, C4 *et* C5, *l'estimateur* $\widehat{\Delta Y}_{\mathrm{reg}}$ *est asymptotiquement sans biais et convergent vers* ΔY.

Proposition 4.3. *Soit* $n_{1,min} = \min(n_{1*}, n_{12})$ *et* $n_{2,\min} = \min(n_{2*}, n_{12})$. *Sous les conditions* C1–C5 *et pour* $K^{(t)} = \mathrm{o}(N)$, $K^{(t)} < \sqrt{n_{t,\min}}$

$$\frac{1}{N}\left(\widehat{\Delta Y}_\pi - \widehat{\Delta Y}_{\mathrm{HT}}\right) = \mathrm{O}_p\left(\sqrt{\frac{K^{(1)}}{n_{1,\min}}}\right) + \mathrm{O}_p\left(\sqrt{\frac{K^{(2)}}{n_{2,\min}}}\right)$$

et par conséquent, $\widehat{\Delta Y}_\pi$ *est asymptotiquement sans biais et convergent vers* ΔY.

Proposition 4.4. *Sous les conditions* C1–C5 *et pour* $K^{(t)} = \mathrm{o}(n_{t,\min}^{1/3})$,

$$\frac{1}{N}\left(\widehat{\Delta Y}_\pi - \widehat{\Delta Y}_{\mathrm{reg}}\right) = \mathrm{o}_p\left(\frac{1}{\sqrt{n_{1,\min}}}\right) + \mathrm{o}_p\left(\frac{1}{\sqrt{n_{2,\min}}}\right).$$

Alors, la variance de $\widehat{\Delta Y}_\pi$ *sous le plan de sondage* p *est asymptotiquement équivalente à la variance de* $\widehat{\Delta Y}_{\mathrm{reg}}$, *c'est à dire pour* $\mathbf{c}' = (\alpha \boldsymbol{E}'^{(1)}, -\beta \boldsymbol{E}'^{(2)}, (1-\alpha)\boldsymbol{E}'^{(1)} - (1-\beta)\boldsymbol{E}'^{(2)})$ *où* $\boldsymbol{E}^{(t)}$ *sont les vecteurs des résidus dans la population, de composantes* $E_k^{(t)} = y_{kt} - \hat{f}_k^{(t)}$, *on a*

$$E_p\left(\frac{\widehat{\Delta Y}_\pi - \Delta Y}{N}\right)^2 \simeq \frac{1}{N^2}\,\mathbf{c}'\left(\frac{\Delta_{kl}^{s,\tau}}{\pi_k^s\pi_l^\tau}\right)_{k,l\in U}^{s,\tau\in\{s_{1*},s_{12},s_{2*}\}}\mathbf{c}.$$

Références

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

[2] J.C. Deville, C.E. Särndal, Calibration estimators in survey sampling, J. Amer. Statist. Assoc. 87 (1992) 376–382.

[3] P. Dierckx, Curves and Surface Fitting with Splines, Clarendon Press, Oxford, 1993.

[4] C. Goga, Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques, Thèse, Université Rennes 2, 2003.

[5] R.M. Royal, W.G. Cumberland, Variance estimation in finite population sampling, J. Amer. Statist. Assoc. 73 (1978) 351–358.

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