Calage non linéaire

Éric Lesage
Ensai-Crest, Laboratoire de Statistique d'Enquête

Résumé

Dans la méthode du calage de Jean-Claude Deville et Carl-Erik Särndal, les équations de calage ne prennent en compte que les estimations exactes de totaux des variables auxiliaires.
L'objectif de cet article est de s'intéresser à d'autres paramètres que le total pour caler ; on regardera notamment le ratio, la médiane, la variance et l'indice de Gini des variables auxiliaires.

1 Introduction

En statistique d'enquête, à la phase d'estimation, deux grandes familles d'estimateurs sont utilisées : d'une part les estimateurs dits "assistés par un modèle" (comme l'estimateur par la regression ou l'estimateur par le ratio) et d'autre part les estimateurs par calage (comme le raking ratio) de Deville et Särndal (1992). Ces deux familles ont d'ailleurs une certaine proximité comme le montre l'estimateur par la régression qui correspond à l'estimateur par calage avec la distance du χ^2.
L'objet de cet article est d'étendre la famille des estimateurs par calage. Actuellement la méthode du calage pourrait être qualifiée de méthode **linéaire**, car elle ne permet de caler que sur des totaux (et plus généralement des paramètres qui sont des fonctions linéaires des observations du type $\theta = \sum_{k \in U} \theta_k$).
L'idée est de pouvoir prendre en compte dans les contraintes de calage des statistiques ou paramètres non linéaires, comme un ratio, une médiane, une moyenne géométrique ou un indice de Gini.
Bien entendu, il existe une façon directe mais coûteuse de caler sur une statistique θ complexe ; on écrit le programme d'optimisation de la méthode du calage dans lequel la contrainte linéaire est remplacée par la contrainte non linéaire $\hat{\theta} = \theta$. Plikusas(2006) explore cette voie dans sa définition générale du calage non linéaire. Ainsi, le problème devient un problème d'optimisation complexe. Toutefois ceci a trois inconvénients : d'abord ce problème peut être complexe à résoudre, ensuite les méthodes numériques à mettre en jeu necessitent beaucoup de temps de calcul et enfin l'utilisateur ne peut pas les implémenter dans son logiciel actuel de calage.
Notre objectif est de conserver une méthode simple à mettre en oeuvre et sans augmentation du temps de calcul informatique.
On évite également la solution qui consisterait à caler sur davantage de totaux. Prenons l'exemple du ratio, on peut obtenir une estimation exacte de ce ratio en calant sur le numérateur et le dénominateur de ce dernier. Là encore, ce serait

une technique coûteuse, puisqu'elle fait augmenter le nombre de contraintes. Par ailleurs, on pourrait se trouver dans les cas où l'on connaît bien le ratio mais pas exactement les totaux du numérateur et du dénominateur.

Nous allons montrer, dans cet article, comment procéder à un calage non linéaire en ayant recours à une équation de calage linéaire.

Dans la deuxième partie de cet article, nous rappellerons le fonctionnement de la méthode du calage et présenterons des cas simples de calage sur paramètre non linéaire. Dans la troisième partie, nous nous intéresserons aux paramètres qui peuvent se définir comme solution d'une équation estimante (Godambe et Thompson, 1986). Nous y évoquerons aussi la notion de paramètre de nuisance. Dans la quatrième partie, nous verrons le cas des paramètres qui s'écrivent comme des fonctions de totaux et qui permettent d'obtenir un calage exact. Enfin, en partie cinq, nous donnerons une méthode générale de calage sur paramètres non linéaires. Cette méthode reposera sur la méthode de linéarisation (Deville, 1999) et donnera dans certains cas un calage approché.

2 Rappel de la méthode du calage et cas simples

2.1 L'estimateur par calage

Soit U, une population finie de taille N, les unités statistiques de cette population sont indexées par une étiquette k avec $k \in \{1, ..., N\}$. Le paramètre d'intérêt est t_y la somme sur U de la caractéristique (variable) y qui prend la valeur y_k pour l'unité k : $t_y = \sum_{k \in U} y_k$.
Un échantillon s est tiré suivant un plan de sondage $p(s)$. Sa taille est notée n, elle est aléatoire. On note π_k la probabilité d'inclusion de k dans l'échantillon s et $d_k = \frac{1}{\pi_k}$ son poids d'échantillonnage. L'estimateur de Horvitz-Thompson s'écrit : $\hat{t}_{y\pi} = \sum_{k \in s} d_k y_k$.
Soient $x_1, ..., x_P$, P variables auxiliaires, connues sur s et $t_{x_1}, ..., t_{x_P}$ les totaux de ces variables auxiliaires également connus.

L'estimateur par calage de t_y est : $\hat{t}_y = \sum_{k \in s} w_k y_k$, avec une série de poids $\{w_k\}_{(k \in s)}$ obtenue par résolution du programme d'optimisation suivant :

$$\min_{\{w_k\}_{(k \in s)}} \sum_{k \in s} d(w_k, d_k)$$

sous contraintes :

$$\begin{cases} \hat{t}_{x_1} = t_{x_1} \\ \dots \\ \hat{t}_{x_P} = t_{x_P} \end{cases}$$

$d(w_k, d_k)$ est une distance, i.e. une fonction qui mesure l'écart entre le poids d'échantillonnage et le poids de calage. Le programme se résout à l'aide d'un Lagrangien. Dans le cas où la distance utilisée est la distance du χ^2 (i.e.

$d(w_k, d_k) = \frac{1}{2}\frac{(w_k - d_k)^2}{d_k}$), on trouve comme solution : $w_k = d_k(1 + x_k^{'}\lambda)$ (avec λ le vecteur de taille P des multiplicateurs de Lagrange).

Nous allons montrer, dans cet article, comment procéder à un calage non linéaire :

$$\hat{\theta}(\{x_{1,k}, ..., x_{P,k}\}_{k\in s}) = \theta(\{x_{1,k}, ..., x_{P,k}\}_{k\in U})$$

en ayant recours à une équation de calage linéaire

$$\sum_{k\in s} w_k \Phi(\theta, x_{1,k}, ..., x_{P,k}) = \sum_{k\in U} \Phi(\theta, x_{1,k}, ..., x_{P,k})$$

Dit autrement, on passe du calage non linéaire $\hat{\theta} = \theta$ au calage sur le total de la nouvelle variable auxiliaire $z_k = \Phi(\theta, x_{1,k}, ..., x_{P,k})$.

2.2 Cas simples où le paramètre peut se ramener à l'estimation d'un total

Il n'est pas évident de savoir d'emblée si un paramètre va pouvoir s'écrire sous la forme d'un total de termes ne dépendant que de l'unité k. Ainsi, cela fonctionne pour tous les moments d'une variable auxiliaire x (on suppose ici que le plan de sondage permet d'estimer parfaitement la taille de la population N) : $t_{x^m} = \frac{1}{N}\sum_{k\in U} x_k^m$.

Par contre, ça ne marche pas pour la variance : $\sigma_x^2 = \frac{1}{N}\sum_{k\in U}\left(x_k - \left(\frac{\sum_{l\in U} x_l}{N}\right)\right)^2$.

Le terme générique de la première somme fait intervenir tous les x_k, ce qui est rédhibitoire. Nous reviendrons sur le cas de la variance ultérieurement dans cet article.

Un autre exemple est celui de la moyenne géométrique (toujours dans le cas où le plan de sondage permet d'estimer parfaitement la taille de la population N) : $\mu_{Geo,x} = \left(\prod_{k\in U} x_k\right)^{1/N}$. En prenant le logarithme de la moyenne géométrique, on a : $ln(\mu_{Geo,x}) = \frac{1}{N}\sum_{k\in U} ln(x_k)$ qui est un total de la nouvelle variable auxiliaire $z_k = ln(x_k)$ sur le total de laquelle on peut caler.

3 Paramètres définis par une équation estimante

3.1 Principe d'estimation par équation estimante

Certains paramètres se définissent, ou peuvent se définir, comme solution d'une fonction implicite appelée **équation estimante sur** U (Godambe et Thompson, 1986), i.e. :

$$\sum_{k\in U} \Phi(\theta, x_{1,k}, ..., x_{P,k}) = 0.$$

Dans ce contexte, on propose de construire un estimateur de θ, noté $\hat{\theta}$, qui soit la solution de **l'équation estimante sur** s :

$$\sum_{k\in s} d_k \Phi(\hat{\theta}_\pi, x_{1,k}, ..., x_{P,k}) = 0.$$

On peut donner des exemples de paramètres définis par une équation estimante :

moyenne	$\sum_{k\in U}(x_k - \mu) = 0$
ratio	$\sum_{k\in U}(x_{1k} - Rx_{2k}) = 0$
médiane	$\sum_{k\in U}(\mathbb{1}_{x_k\leq m} - \frac{1}{2}) = 0$
coefficients de régression	équations normales

Revenons sur le cas de la variance. La variance peut être définie par deux équations estimantes :

$$\begin{cases} \sum_{k\in U}(x_k - \mu_x) = 0 \\ \sum_{k\in U}\left((x_k - \mu_x)^2 - \sigma_x^2\right) = 0 \end{cases}$$

On voit que dans ce cas la moyenne μ_x apparaît comme un paramètre de nuisance (Binder, 1991). Il y aurait plusieurs solutions pour s'en affranchir. On peut soit utiliser la vraie valeur de μ_x (sans pour autant caler dessus), soit remplacer la moyenne par une estimation de celle-ci construite à partir de l'échantillon s.

3.2 Calage dans le cas de paramètres définis par des équations estimantes

Proposition 1 *Dans le cas où θ est défini par une équation estimante, caler sur θ revient à imposer que θ soit la solution de l'équation estimante sur s :*

$$\sum_{k\in s} w_k \Phi(\theta, x_{1,k}, ..., x_{P,k}) = 0.$$

D'un point de vue pratique, on construit une nouvelle variable auxiliaire $z_k = \Phi(\theta, x_{1,k}, ..., x_{P,k})$ sur le total de laquelle on cale. On a donc comme équation de calage : $\hat{t}_z = t_z = 0$.

4 Paramètres définis comme des fonctions de totaux et qui offrent un calage exact

4.1 Fonction de totaux homogènes de degré 0

On considère le cas où le paramètre θ est défini comme une fonction de totaux : $\theta = f(t_{x_1}, ..., t_{x_P})$, avec f une fonction homogène de degré 0. Par ailleurs, on prend comme estimateur de θ l'estimateur par substitution $\hat{\theta} = f\left(\hat{t}_{x_1}, ..., \hat{t}_{x_P}\right)$.

Définition 1 *Soit f une fonction de $\mathbb{R}^P$ dans $\mathbb{R}$. f est homogène de degré 0 si pour tout λ de $\mathbb{R}$, $f(\lambda x_1, ..., \lambda x_P) = f(x_1, ..., x_P)$.*

4.2 Fonction de 2 totaux

Puisque f est homogène de degé 0, on a : $\theta = f(t_{x_1}, t_{x_2}) = f\left(\frac{t_{x_1}}{t_{x_2}}, 1\right) = g\left(\frac{t_{x_1}}{t_{x_2}}\right)$ (prendre $\lambda = \frac{1}{t_{x_2}}$).

Proposition 2 *Si θ est une fonction bijective de deux totaux ($\theta = f(t_{x1}, t_{x2})$), homogène de degré 0, alors caler sur θ est équivalent à caler sur le ratio $R = \frac{t_{x1}}{t_{x2}}$. Ce qui est équivalent à caler sur le total de la variable $z_k = x_{1k} - Rx_{2k}$.*

On peut reprendre les exemples de la 2e partie. En effet, dans le cas où N n'est pas estimé parfaitement, on introduit un total supplémentaire dans les paramètres tels que les moments, la moyenne généralisée ou la moyenne géométrique.

4.3 Fonction d'un ratio de combinaisons linéaires de totaux

Proposition 3 *Si θ est fonction bijective d'un ratio de combinaisons linéaires de totaux :*

$$\theta = f\left(\frac{\alpha' \cdot \mathbf{t_x}}{\beta' \cdot \mathbf{t_x}}\right)$$

avec α, β et $\mathbf{t_x}$ des vecteurs colonnes de taille P,
et si on définit : $\hat{\theta} = f\left(\frac{\alpha' \cdot \hat{\mathbf{t}}_\mathbf{x}}{\beta' \cdot \hat{\mathbf{t}}_\mathbf{x}}\right)$
alors caler sur θ est équivalent à caler sur le total de la variable $z_k = \left(\alpha' - f^{-1}(\theta)\beta'\right) \cdot \mathbf{x_k}$.

Preuve 1

$$\hat{\theta} = \theta \Longleftrightarrow \frac{\alpha' \cdot \hat{\mathbf{t}}_\mathbf{x}}{\beta' \cdot \hat{\mathbf{t}}_\mathbf{x}} = f^{-1}(\theta) \Longleftrightarrow \left(\alpha' - f^{-1}(\theta)\beta'\right) \cdot \hat{\mathbf{t}}_\mathbf{x} = 0$$

5 Paramètre quelconque : méthode de linéarisation

Dans cette section on examine le cas des paramètres qui n'entrent pas dans les catégories précédemment exposées (i.e. paramètres définis par une équation estimante ou paramètres définis comme des fonctions de totaux idoines). On peut également voir ce paragraphe comme une méthode générale. Toutefois, cette dernière ne donne qu'un calage approché sur le paramètre non linéaire.

5.1 Principe de linéarisation d'estimateur complexe

Le principe des techniques de linéarisation est d'approcher des formulations d'estimateurs non linéaires par des expressions linéaires auxquelles on les assimile. Ces méthodes ont été développées initialement pour permettre des calculs analytiques approchés de variance d'estimateurs complexes. On trouve une présentation claire et pédagogique de la linéarisation chez Ardilly,

2006. Pour une présentation plus formelle et riche en références bibliographiques, on peut consulter Dell et d'Haultfœuille, 2006.

Donc, pour un estimateur $\hat{\theta}$ de θ linéarisable, on aura l'expression (sous certaines conditions, dont n grand) :

$$\hat{\theta} \approx \theta - t_z + \hat{t}_z$$

Avec z_k la variable dite linéarisée de $\hat{\theta}$, également notée $lin_k\theta$ et $\hat{t}_z = \sum_{k \in s} w_k z_k$ l'estimateur par calage de la variable z_k.

Il existe plusieurs méthodes pour trouver z_k qui sont plus ou moins complexes en fonction de la régularité du paramètre θ. Toutefois, toutes reposent sur des notions de dérivabilité, de développement de taylor à l'ordre 1 et de convergence asymptotique.

5.1.1 Linéarisation d'un paramètre θ fonction régulière de totaux

Commençons avec le cas de paramètres qui sont des fonctions régulières de totaux : $\theta = f\left(t_{x_1}, ..., t_{x_P}\right)$, avec f une fonction deux fois dérivable.

Nous avons :

$$\begin{aligned}
\hat{\theta} &= f\left(\hat{t}_{x_1}, ..., \hat{t}_{x_P}\right) \\
&= f\left(t_{x_1}, ..., t_{x_P}\right) \\
&\quad + \frac{\partial f}{\partial x_1}\left(t_{x_1}, ..., t_{x_P}\right)\left(\hat{t}_{x_1} - t_{x_1}\right) + ... + \frac{\partial f}{\partial x_P}\left(t_{x_1}, ..., t_{x_P}\right)\left(\hat{t}_{x_P} - t_{x_P}\right) \\
&\quad + K_n
\end{aligned}$$

Avec K_n, une variable aléatoire d'expression complexe qui a un ordre de grandeur de $\frac{1}{n}$ en "probabilité" (voir Isaki et Fuller, 1982, pour le cadre théorique précis).

On trouve donc comme variable linéarisée de $\hat{\theta}$:

$$lin_k\theta = z_k = \sum_{j=1}^{P} \frac{\partial f}{\partial x_j}\left(t_{x_1}, ..., t_{x_P}\right) x_k$$

Exemple de la variance
La variance de la variable auxiliaire x se définit par la formule :

$$\sigma_x^2 = \frac{1}{N}\sum_{k \in U} x_k^2 - \left(\frac{\sum_{k \in U} x_k}{N}\right)^2 = f(t_x, t_x^2, N)$$

Son estimateur "naturel" (ou estimateur par substitution) est :

$$\hat{\sigma}_x^2 = \frac{1}{\hat{N}}\sum_{k \in s} x_k^2 - \left(\frac{\sum_{k \in s} x_k}{\hat{N}}\right)^2$$

La variable linéarisée de σ_x^2 est :

$$z_k = -2\frac{t_x}{N^2}x_k + \frac{1}{N}x_k^2 - \frac{1}{N}\left\{\sigma_x^2 - \left(\frac{t_x}{N}\right)^2\right\}$$

5.1.2 Linéarisation d'un paramètre θ qui n'est pas une fonction régulière de totaux

On peut utiliser dans ce cas la méthode de linéarisation basée sur la fonction d'influence d'une fonctionnelle développée par Deville, 1999.
Je ne donne ici qu'une présentation rapide de cette méthode que j'ai empruntée à Dell et al., 2002.

Le calcul d'un paramètre (ou statistique) sur la population U consiste à mettre un poids égal à 1 sur chaque individu k de U. L'estimation de ce même paramètre sur l'échantillon consiste à mettre un poids égal à w_k sur les individus k de s et un poids nul pour les autres. Il est donc intéressant de se demander quel est l'effet d'une variation de poids de l'individu k sur l'estimation du paramètre (mesure de l'**influence** de k !).
D'un point de vue mathématique, notons M la mesure qui met un point égal à 1 sur chaque individu, $\hat{M}$ la mesure associée au sondage et $M + t\delta_k$ la mesure qui met un poids de 1 à tous les individus sauf k, qui, lui, a un poids de $(1 + t)$. Notons $\theta = T(M)$, $\hat{\theta} = T(\hat{M})$ et $T(M+t\delta_k)$. La dérivée de T liée à une variation infinitésimale de poids associée à l'individu k est appelée fonction d'influence et est égale à :

$$lin_k T = z_k = \lim_{t\to 0} \frac{T(M + t\delta_k) - T(M)}{t}$$

Exemple de l'indice de Gini Il existe plusieurs définitions de l'indice de Gini. Prenons la définition suivante :

$$G = 2\frac{\sum_{k\in U}\left(\sum_{l\in U}\mathbb{1}_{x_l\leq x_k}\right)x_k}{N\sum_{k\in U}x_k} - 1$$

L'estimateur par substitution de G est :

$$\hat{G} = 2\frac{\sum_{k\in s}\left(\sum_{l\in s}w_l\mathbb{1}_{x_l\leq x_k}\right)w_k x_k}{\hat{N}\sum_{k\in s}w_k x_k} - 1$$

Pour simplifier les notations, on notera :

$$r(k) = \sum_{l\in U}\mathbb{1}_{x_l\leq x_k}$$

$$\hat{r}(k) = \sum_{l\in s}w_l\mathbb{1}_{x_l\leq x_k}$$

La linéarisée de l'indice de Gini vaut :

$$z_k = \frac{1}{Nt_x}\left((2r(k) - (G+1)N)\, x_k + 2\sum_{l\in U} \mathbb{1}_{x_k \leq x_l} x_l - (G+1)\,(t_x)\right)$$

5.2 Calage dans le cas de paramètres linéarisables

Proposition 4 *Soit θ, un paramètre non linéaire, estimé par $\hat{\theta}$, son estimateur par substitution.*
Si $\hat{\theta}$ est linéarisable, alors caler sur t_z (avec z_k la variable linéarisée de θ) est équivalent à un calage approché (ou asymptotiquement exact) sur θ.

$$\hat{t}_z = t_z \Leftrightarrow \hat{\theta} \approx \theta$$

Exemple du calage (approché) sur la variance
En faisant le calage sur la variable linéarisée, on n'obtient pas exactement $\hat{\sigma}_x^2 = \sigma_x^2$, mais :

$$\hat{\sigma}_x^2 = \sigma_x^2 - \left(\frac{t_x}{N} - \frac{\hat{t}_x}{\hat{N}}\right)^2$$

.

6 Conclusion

Les estimateurs par calage, qui mettent à profit l'information auxiliaire, sont très largement connus et utilisés pour améliorer la précisions des estimations des paramètres d'intérêt d'une enquête. Or, les équations de calage ne mobilisaient jusqu'à présent que des estimations exactes de totaux.
On a pu voir dans cet article qu'il est assez facile de caler sur des paramètres non linéaires (i.e. qui ne sont pas des totaux) lorsque ceci peuvent se définir comme solution d'équations estimantes ou lorsqu'ils sont fonctions de ratios de totaux.
Par ailleurs, dans les autres cas, on peut procéder à un calage approché (ou asymptotiquement exact), en ayant recours aux techniques de linéarisation.

Références

[1] Ardilly P. (2006). *Les Techniques de Sondage*. Paris : Technip, 583-595.

[2] Binder D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the survey research methods section*, American Statistical Association, 34-42.

[3] Dell, F. , d'Haultfoeuille, X., Février, Ph. et Massé, E. (2002). Mise en œuvre du calcul de variance par linéarisation. *Insee-Méthodes : Actes des Journées de Méthodologie Statistique 2002*.

[4] Dell, F. and d'Haultfoeuille, X. (2006). Measuring the Evolution of Complex Indicators Theory and Application to the Poverty Rate in France. *Série des Documents de Travail du CREST*, INSEE.

[5] Deville, J.-C. (1999). Estimation de variance pour des staistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

[6] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey samling. *Journal of the American Statistical Association*, 87, 376-382.

[7] Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population : Their relationship and estimation. *Revue International de Statistique*, 54, 127-138.

[8] Isaki, C.T. and Fuller, W.A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77, 89-96.

[9] Plikusas, A. (2006). Non-linear calibration. *Proceedings, Workshop on survey sampling*, Venspils, Latvia. Riga : Central Statistical Bureau of Latvia.

[10] Krapavickaite, D., and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16, 347-364.