N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

COMBINAISON DE DONNÉES PROVENANT D'ENQUÊTES MULTIPLES PAR LA MÉTHODE DE VRAISEMBLANCE EMPIRIQUE

Changbao Wu[1]

RÉSUMÉ

Il est souvent souhaitable de combiner des données recueillies lors d'enquêtes multiples compatibles afin de satisfaire à certaines exigences de cohérence et d'obtenir des estimateurs plus efficaces. Les travaux antérieurs portant sur ce sujet incluent ceux de Zieschang (1990) et de Renssen et Nieuwenbroek (1997) qui ont proposé d'utiliser l'estimateur par la régression généralisée avec un nombre plus grand de variables auxiliaires pour atteindre cet objectif. Un inconvénient de la méthode est que les poids redressés peuvent prendre des valeurs négatives, ce qui est fort peu souhaitable. Dans le présent article, nous utilisons la méthode de la pseudo-vraisemblance empirique mise au point récemment pour construire des estimateurs qui non seulement satisfont aux exigences de cohérence et d'efficacité, mais ont aussi des propriétés plus intéressantes. Les deux méthodes sont asymptotiquement équivalentes, mais la dernière a une interprétation claire, dans le sens du maximum de vraisemblance, ainsi que des poids redressés systématiquement positifs. Nous fournissons aussi des algorithmes efficaces pour calculer les estimateurs proposés, ce qui permet d'appliquer facilement la méthode à des enquêtes réelles.

MOTS CLÉS : Algorithme de Newton-Raphson, contraintes d'étalonnage, estimateur par la régression généralisée, exigences de cohérence, pseudo-vraisemblance empirique.

1. INTRODUCTION

Dans le domaine des sondages, en pratique, on recourt régulièrement au redressement des poids pour tenir compte, entre autres, des exigences de cohérence interne qui intéressent aussi bien les statisticiens d'enquête que les utilisateurs éventuels des données. Les contraintes d'étalonnage imposées le plus fréquemment sont celles où les poids redressés w_i reproduisent les totaux (ou moyennes) de population connus des variables auxiliaires $\mathbf{x}$, c.-à-d. $\sum_{i\in s} w_i \mathbf{x}_i = \mathbf{X}$, où s représente l'ensemble d'unités échantillonnées et $\mathbf{X}$ est le vecteur des totaux de population connus. Ce genre de redressement peut se faire au moyen de l'estimateur par la régression généralisée. Ce dernier est un instrument qui permet non seulement de satisfaire aux contraintes d'étalonnage, mais qui est aussi plus efficace que l'estimateur d'Horvitz-Thompson de référence.

Lorsqu'on réalise deux enquêtes (ou plus) auprès d'une même population, une autre exigence de cohérence peut être nécessaire. Si les données sur certaines variables auxiliaires sont recueillies conjointement au cours des deux enquêtes, mais que les totaux de population sont inconnus, alors il est souhaitable que, à part les contraintes d'étalonnage sur les variables auxiliaires dont on connaît les totaux de population, les poids des deux enquêtes produisent les mêmes estimations pour les totaux de population inconnus des variables auxiliaires communes. Ce problème a été abordé antérieurement par Zieschang (1990) et par Renssen et Nieuwenbroek (1997) qui, dans les deux cas, proposent d'utiliser l'estimateur par la régression généralisée (GREG) en augmentant le nombre de variables auxiliaires pour atteindre cet objectif.

Cependant, l'approche de la régression généralisée a une propriété indésirable, qui avait déjà été reconnue par les auteurs susmentionnés : « Un inconvénient de la méthode est la plus forte possibilité d'obtenir des poids négatifs, à cause du plus grand nombre de variables explicatives. L'obtention de poids négatifs est une caractéristique inhérente de l'estimateur par la régression généralisée et, pour nombre d'utilisateurs, il s'agit d'une caractéristique indésirable. » [Traduction] (Renssen et Nieuwenbroek, 1997).

[1] Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1, cbwu@uwaterloo.ca

Nous proposons d'utiliser la méthode de la pseudo-vraisemblance empirique (EL pour *pseudo empirical likelihood)* mise au point récemment pour construire des estimateurs qui non seulement satisfont aux exigences d'efficacité et de convergence, mais qui ont aussi des caractéristiques plus intéressantes. La méthode EL est un outil d'inférence non paramétrique puissant ayant des applications dans de nombreux domaines statistiques. À cet égard, consulter Owen (2001) pour un compte rendu complet et une vue d'ensemble à jour. Cependant, par le passé, cette méthode a d'abord été utilisée en échantillonnage par Hartley et Rao (1968). Sa nature discrète et non paramétrique est particulièrement séduisante pour la résolution de problèmes en population finie. Dans le présent article, nous démontrons que l'approche EL convient bien dans le contexte courant, et qu'on peut formuler naturellement les exigences de cohérence et d'efficacité entre deux enquêtes ou plus sous forme de contraintes et les intégrer dans le processus d'estimation du maximum de vraisemblance. Les deux approches, GREG et EL, sont asymptotiquement équivalentes, mais la seconde a une interprétation claire, dans le sens du maximum de vraisemblance, et donne des poids qui sont systématiquement positifs.

Dans la suite de l'exposé, nous considérons deux enquêtes, mais notre méthode peut être étendue pour traiter les enquêtes multiples. Une approche logiquement valable comprend une estimation conjointe du maximum de vraisemblance en se servant de deux échantillons. Cette approche est présentée à la section 2. Nous présentons aussi dans cette section deux algorithmes pour le calcul de l'estimateur EL proposé. Le premier algorithme s'appuie sur la méthode du profil de vraisemblance pour la recherche d'une solution et n'est efficace que si la variable auxiliaire commune est unidimensionnelle. Le deuxième s'appuie sur une nouvelle reformulation du problème et peut être appliqué facilement sous des conditions générales utilisant l'algorithme bien développé de Chen, Sitter et Wu (2002). À la section 3, nous suivons l'approche des vraisemblances empiriques individuelles où les estimateurs EL sont calculés séparément pour chaque enquête en estimant les moyennes de population inconnues des variables auxiliaires communes d'après les données d'échantillon combinées et en utilisant ces estimations comme valeurs de contrôle. Dans ce cas, le calcul est simple et direct. Les performances en échantillon fini des estimateurs EL proposés, et leur comparaison aux estimateurs GREG de Zieschang (1990) et de Renssen et Nieuwenbroek (1997) sont présentées à la section 4 à l'aide d'une étude en simulation. Enfin, à la section 5, nous formulons certaines remarques en guise de conclusion.

2. L'APPROCHE DE LA VRAISEMBLANCE EMPIRIQUE COMBINÉE

Supposons que la population finie comprenne N unités identifiables. À la i^{e} unité sont associés les valeurs des variables étudiées y_1 et y_2, et les vecteurs de variables auxiliaires $\mathbf{x}_1, \mathbf{x}_2$ et $\mathbf{z}$, représentés par $y_{1i}, y_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}$, et $\mathbf{z}_i$, respectivement, pour i = 1,2,…,N. L'information sur $(y_1, \mathbf{x}_1)$ est recueillie durant la première enquête et celle sur $(y_2, \mathbf{x}_2)$, durant la seconde. En outre, les données sur les variables auxiliaires communes $\mathbf{z}$ sont recueillies durant les deux enquêtes. Toutefois, les deux enquêtes sont réalisées indépendamment l'une de l'autre. Les deux ensembles de données d'échantillon sont $\{(y_{1i}, \mathbf{x}_{1i}, \mathbf{z}_i), i \in s_1\}$ et $\{(y_{2j}, \mathbf{x}_{2j}, \mathbf{z}_j), j \in s_2\}$, où s_1 et s_2 sont les ensembles d'unités échantillonnées à partir de la première et de la deuxième enquête, respectivement. Les moyennes de population $\overline{\mathbf{X}}_t = N^{-1}\sum_{i=1}^{N}\mathbf{x}_{ti}$ sont connues (t = 1, 2), mais $\overline{\mathbf{Z}} = N^{-1}\sum_{i=1}^{N}\mathbf{z}_i$ est inconnue. Zieschang (1990) et Renssen et Nieuwenbroek (1997) ont fourni d'excellentes motivations et des exemples réels de ces conditions d'étude, y compris une application fort utile aux plans de sondage à questionnaire scindé. Soit $\overline{Y}_t = N^{-1}\sum_{i=1}^{N} y_{ti}, t = 1, 2,$ les quantités de population d'intérêt. Si y_1 et y_2 mesurent la même caractéristique, mais sur des périodes différentes, alors la différence $\Delta = \overline{Y}_2 - \overline{Y}_1$ pourrait aussi être d'intérêt.

En suivant des arguments comparables à ceux de Chen et Sitter (1999), nous pouvons écrire la fonction combinée de pseudo log-vraisemblance empirique fondée sur les deux échantillons comme suit

$$l(\mathbf{p}, \mathbf{q}) = \sum_{i \in s_1} d_{1i} \log(p_i) + \sum_{j \in s_2} d_{2j} \log(q_j),$$

où $\mathbf{p} = (p_1,..., p_{n_1})', \mathbf{q} = (q_1,..., q_{n_2})', p_i = \Pr(y_1 = y_{1i}), q_j = \Pr(y_2 = y_{2j}), d_{ti} = 1/\pi_{ti}, \pi_{ti}$ sont les probabilités d'inclusion de premier ordre et n_t est la taille de l'échantillon de la t^e enquête, t = 1, 2.

Les estimateurs du maximum de pseudo-vraisemblance empirique pour $\overline{Y}_1$ et $\overline{Y}_2$ sont définis comme étant

$$\hat{\overline{Y}}_1 = \sum_{i \in s_1} \hat{p}_i y_{1i} \text{ et } \hat{\overline{Y}}_2 = \sum_{j \in s_2} \hat{q}_j y_{2j},$$

où les $\hat{p}_i$ et $\hat{q}_j$, qui sont les poids redressés, maximisent la fonction conjointe de pseudo-vraisemblance empirique $l(\mathbf{p}, \mathbf{q})$, conditionnellement à un système de contraintes de normalisation et de cohérence :

$$\sum_{i \in s_1} p_i = 1(p_i > 0), \sum_{j \in s_2} q_j = 1(q_j > 0), \tag{1}$$

$$\sum_{i \in s_1} p_i \mathbf{x}_{1i} = \overline{\mathbf{X}}_1, \sum_{j \in s_2} q_j \mathbf{x}_{2j} = \overline{\mathbf{X}}_2, \tag{2}$$

$$\sum_{i \in s_1} p_i \mathbf{z}_i = \sum_{j \in s_2} q_j \mathbf{z}_j, \tag{3}$$

Les deux ensembles de contraintes d'étalonnage donnés dans (2) pourraient comprendre des mesures des mêmes variables **x** et, donc, les mêmes moyennes de population. En l'absence de moyennes de population connues, certaines équations de (2), voire toutes, peuvent être éliminées du système. Le dernier ensemble d'équations (3) impose la cohérence entre les deux enquêtes sur les variables auxiliaires communes. Il rend aussi les estimateurs résultants $\hat{\overline{Y}}_1$ et $\hat{\overline{Y}}_2$ plus efficaces grâce à l'utilisation de l'information combinée provenant des deux enquêtes.

L'une des questions connexes, ici, est l'existence des estimateurs EL combinés définis plus haut. Les estimateurs du maximum de pseudo-vraisemblance empirique $\hat{\overline{Y}}_t$ n'existeront pas si $\overline{\mathbf{X}}_t$ n'est pas un point intérieur de l'enveloppe convexe formée par $\{\mathbf{x}_{ti}, i \in s_t\}$ ou si les deux enveloppes convexes formées par $\{\mathbf{z}_i, i \in s_1\}$ et $\{\mathbf{z}_j, j \in s_2\}$ sont disjointes. Cela se produit avec une probabilité approchant zéro lorsque les tailles des deux échantillons tendent vers l'infini. Il est possible d'esquisser une preuve de ceci en suivant les lignes du lemme 1 de Chen et Sitter (1999).

Une autre question d'ordre pratique importante est celle des calculs que nécessite la méthode EL proposée. Nous présentons deux algorithmes, qui, l'un et l'autre, tirent parti de l'algorithme à bon comportement de Chen et coll. (2002) pour calculer les estimateurs du maximum de vraisemblance empirique sous échantillon simple non stratifié. Le premier algorithme est efficace quand la variable auxiliaire commune est univariée, tandis que le deuxième peut être utilisé sous des conditions générales.

2.1 Premier algorithme

Posons que $\sum_{i \in s_1} p_i \mathbf{z}_i = \sum_{j \in s_2} q_j \mathbf{z}_j = \boldsymbol{\theta}$ est fixe. Il est alors simple de montrer, en utilisant la méthode du multiplicateur de Lagrange, que

$$\hat{p}_i = \frac{d_{1i}^*}{1 + \boldsymbol{\lambda}_1' \mathbf{u}_{1i}(\boldsymbol{\theta})}, \ \hat{q}_j = \frac{d_{2j}^*}{1 + \boldsymbol{\lambda}_2' \mathbf{u}_{2j}(\boldsymbol{\theta})}, \tag{4}$$

où $d_{ti}^* = d_{ti} / \sum_{i \in s_t} d_{ti}$ et

$$\mathbf{u}_{ti}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{x}_{ti} - \overline{\mathbf{X}}_t \\ \mathbf{z}_{ti} - \boldsymbol{\theta} \end{pmatrix},$$

étant entendu que $\mathbf{z}_{ti}$ désigne $\mathbf{z}_i$ pour le t^e échantillon, t = 1, 2. Les multiplicateurs de Lagrange $\boldsymbol{\lambda}_1$ et $\boldsymbol{\lambda}_2$ sont les solutions de

$$\sum_{i \in s_1} \frac{d_{1i}^* \mathbf{u}_{1i}(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_1' \mathbf{u}_{1i}(\boldsymbol{\theta})} = 0 \text{ et } \sum_{j \in s_2} \frac{d_{2j}^* \mathbf{u}_{2j}(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_2' \mathbf{u}_{2j}(\boldsymbol{\theta})} = 0, \tag{5}$$

respectivement. La fonction de vraisemblance profil pour $\boldsymbol{\theta}$ s'obtient alors en plaçant $\hat{p}_i$ et $\hat{q}_j$ dans $l(\mathbf{p}, \mathbf{q})$ et est donnée par (avec un terme constant omis)

$$l(\boldsymbol{\theta}) = -\sum_{i \in s_1} d_{1i} \log\{1 + \boldsymbol{\lambda}_1' \mathbf{u}_{1i}(\boldsymbol{\theta})\} - \sum_{j \in s_2} d_{2j} \log\{1 + \boldsymbol{\lambda}_2' \mathbf{u}_{2j}(\boldsymbol{\theta})\}.$$

Le point maximal de $l(\boldsymbol{\theta})$, représenté par $\hat{\boldsymbol{\theta}}$, peut être trouvé par l'analyse de profil classique. Nous obtenons les poids redressés finaux $\hat{p}_i$ et $\hat{q}_j$ en introduisant $\hat{\boldsymbol{\theta}}$ et le $\boldsymbol{\lambda}_t$ associé dans (4).

Cet algorithme comprend la recherche de $\boldsymbol{\lambda}_t (t = 1, 2)$ comme solutions de (5) pour chaque valeur fixée de $\boldsymbol{\theta}$, puis la recherche de $\hat{\boldsymbol{\theta}}$ qui maximise $l(\boldsymbol{\theta})$. Pour la première partie, Chen et coll. (2002) ont déployé un algorithme simple et stable pour la résolution de (5) en vue d'obtenir le $\boldsymbol{\lambda}_t$ vectoriel. Quant à $\hat{\boldsymbol{\theta}}$, si la variable auxiliaire commune **z** est unidimensionnelle, il est facile de la trouver en utilisant la méthode habituelle du profil de vraisemblance. Si **z** est multidimensionnelle, il en est de même de $\boldsymbol{\theta}$, et cet algorithme devient peu commode. Un algorithme plus souple est nécessaire.

2.2 Second algorithme

Supposons que $\mathbf{z}_i = (z_{1i}, ..., z_{ki})'$ est de dimension *k*. Si nous augmentons $\mathbf{z}_i$ pour passer à *k*+1 dimensions en incluant $z_{(k+1)i} = 1$ comme dernière composante, nous pouvons réécrire le système de contraintes (1), (2) et (3) sous la forme

$$\sum_{i \in s_1} p_i + \sum_{j \in s_2} q_j = 2, \tag{6}$$

$$\begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(2)} \\ \mathbf{Z}^{(1)} & -\mathbf{Z}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \mathbf{0} \end{bmatrix}, \tag{7}$$

où $\mathbf{X}^{(t)} = (\mathbf{x}_{t1}, ..., \mathbf{x}_{tn_t})', \mathbf{Z}^{(t)} = (\mathbf{z}_{t1}, ..., \mathbf{z}_{tn_t})', \mathbf{z}_{ti}$ représente $\mathbf{z}_i$ provenant de la t^e enquête avec 1 comme dernière composante, t = 1, 2. Notons que la toute dernière équation du système (7) est $\sum_{i \in s_1} p_i - \sum_{j \in s_2} q_j = 0$, qui, regroupée à (6), implique que $\sum_{i \in s_1} p_i = 1$ et $\sum_{j \in s_2} q_j = 1$.

Nous pouvons en outre réécrire (7) sous la forme

$$\sum_{i \in s_1} p_i \mathbf{u}_{1i} + \sum_{j \in s_2} q_j \mathbf{u}_{2j} = \mathbf{0}, \tag{8}$$

où

$$\mathbf{u}_{1i} = \begin{bmatrix} \mathbf{x}_{1i} \\ \mathbf{0} \\ \mathbf{z}_{1i} \end{bmatrix} - \frac{1}{2}\begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \mathbf{0} \end{bmatrix}, \mathbf{u}_{2j} = \begin{bmatrix} \mathbf{0} \\ \mathbf{x}_{2j} \\ -\mathbf{z}_{2j} \end{bmatrix} - \frac{1}{2}\begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \mathbf{0} \end{bmatrix}. \tag{9}$$

Il est maintenant clair que maximiser $l(\mathbf{p},\mathbf{q})$ sous les contraintes (1), (2) et (3) équivaut à maximiser $l(\mathbf{p},\mathbf{q})$ conditionnellement à (6) et à (8). En utilisant la méthode du multiplicateur de Lagrange, nous pouvons montrer que

$$\hat{p}_i = \frac{d_{1i}^*}{1+\boldsymbol{\lambda}'\mathbf{u}_{1i}}, \hat{q}_j = \frac{d_{2j}^*}{1+\boldsymbol{\lambda}'\mathbf{u}_{2j}},$$

où $d_{ti}^* = 2d_{ti} / \left(\sum_{i\in s_1} d_{1i} + \sum_{j\in s_2} d_{2j}\right)$ pour t = 1, 2, et le multiplicateur de Lagrange courant $\boldsymbol{\lambda}$ est la solution de

$$\sum_{t=1,2} \sum_{i\in s_t} \frac{d_{ti}^* \mathbf{u}_{ti}}{1+\boldsymbol{\lambda}'\mathbf{u}_{ti}} = \mathbf{0}. \tag{10}$$

L'algorithme de Newton-Raphson modifié par Chen et coll. (2002) peut être utilisé, dans les conditions idéales, pour résoudre (10). Bien qu'une telle modification soit nécessaire pour prouver théoriquement la convergence, selon notre expérience, la procédure d'itération classique de Newton-Raphson qui suit donne de bons résultats dans presque tous les cas :

$$\boldsymbol{\lambda}^{(m+1)} = \boldsymbol{\lambda}^{(m)} + \left\{ \sum_{t=1,2} \sum_{i\in s_t} \frac{d_{ti}^* \mathbf{u}_{ti}\mathbf{u}_{ti}'}{\left(1+\left[\boldsymbol{\lambda}^{(m)}\right]'\mathbf{u}_{ti}\right)^2} \right\}^{-1} \sum_{t=1,2} \sum_{i\in s_t} \frac{d_{ti}^* \mathbf{u}_{ti}}{1+\left[\boldsymbol{\lambda}^{(m)}\right]'\mathbf{u}_{ti}},$$

en choisissant **0** comme valeur initiale de $\boldsymbol{\lambda}$.

Ce second algorithme est applicable sous des conditions générales. Il ne nécessite qu'une seule résolution de (10) à l'aide de l'algorithme à bon comportement de Chen et coll. (2002) et peut être programmé par les utilisateurs des données d'enquête au moyen de logiciels statistiques bien connus, tels que SAS ou R/Splus.

2.3 Comparaison à la méthode de régression de Zieschang

L'approche de la vraisemblance empirique combinée proposée dans le présent article a été élaborée dans le même esprit que l'estimateur par la régression généralisée composite proposé par Zieschang (1990). Ce point est évident si nous comparons les contraintes (7) utilisées ici au système d'équations de régression élargi (3.10) employé par Zieschang. Cependant, la méthode de la vraisemblance empirique offre plusieurs avantages. Outre l'interprétation claire dans le sens du maximum de vraisemblance, l'estimateur EL est calculé en se fondant sur les poids normalisés intrinsèquement positifs, c.-à-d., $\hat{p}_i > 0$ et $\sum_{i\in s_1} \hat{p}_i = 1$. Cette dernière caractéristique est particulièrement intéressante pour les utilisateurs des données d'enquête, puisqu'ils utilisent souvent les poids publiés à diverses fins, dont l'estimation de proportions ou, de façon plus générale, de la fonction de répartition en population finie $F(y)$. L'estimateur EL $\hat{F}_{EL}(y)$ sera, lui-même, une vraie fonction de répartition. Il respecte les intervalles et peut être inversé pour obtenir des estimations de quantile.

Il est possible d'établir une relation explicite entre l'estimateur EL et un estimateur de type régression généralisée.

Théorème 1. *Sous des conditions de régularité appropriées, les estimateurs du maximum de pseudo-vraisemblance empirique* $\hat{\bar{Y}}_1 = \sum_{i\in s_1} \hat{p}_i y_{1i}$ *et* $\hat{\bar{Y}}_2 = \sum_{j\in s_2} \hat{q}_j y_{2j}$ *sont asymptotiquement équivalents à un estimateur de type régression généralisée, c.-à-d.*

$$\hat{\bar{Y}}_t = \bar{y}_t + \hat{\mathbf{B}}'_{t1}\left(\overline{\mathbf{X}}_1 - \bar{\mathbf{x}}_1\right) + \hat{\mathbf{B}}'_{t2}\left(\overline{\mathbf{X}}_2 - \bar{\mathbf{x}}_2\right) + \hat{\mathbf{B}}'_{t3}\left(\bar{\mathbf{z}}_2 - \bar{\mathbf{z}}_1\right) + o_p\left(n^{-1/2}\right), \tag{11}$$

où $\bar{y}_t = \sum_{i\in s_t} d^*_{ti}\, y_{ti}, \bar{\mathbf{x}}_t = \sum_{i\in s_t} d^*_{ti}\, \mathbf{x}_{ti}, \bar{\mathbf{z}}_t = \sum_{i\in s_t} d^*_{ti}\, \mathbf{z}_{ti}, n = n_1 + n_2$ *, et les « coefficients de régression » combinés* $\hat{\mathbf{B}}_t = \left(\hat{\mathbf{B}}'_{ti}, \hat{\mathbf{B}}'_{t2}, \hat{\mathbf{B}}'_{t3}\right)'$ *sont donnés par*

$$\hat{\mathbf{B}}_t = \left(\sum_{t=1,2}\sum_{i\in s_t} d^*_{ti}\, \mathbf{u}_{ti}\, \mathbf{u}'_{ti}\right)^{-1} \sum_{i\in s_t} d^*_{ti}\, \mathbf{u}_{ti} y_{ti},$$

avec les $\mathbf{u}_{ti}$ *définis par (9).*

Les conditions de régularité requises et les preuves des théorèmes figurent dans un rapport technique que l'on peut se procurer auprès de l'auteur. Il convient de souligner que l'information auxiliaire combinée $\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2, \bar{\mathbf{z}}_1$ et $\bar{\mathbf{z}}_2$, ainsi que les poids de sondage de base d_{1i} et d_{2j} provenant des deux enquêtes figurent explicitement dans l'estimateur par la régression généralisée équivalent, c'est-à-dire un estimateur assez unique du point de vue classique. En outre, si les deux plans de sondage satisfont $\sum_{i\in s_1} d_{1i} = \sum_{i\in s_2} d_{2i} = N$, comme cela est le cas sous échantillonnage aléatoire simple ou sous échantillonnage aléatoire stratifié, alors $d^*_{ti} = d_{ti} / N$, et les estimateurs $\bar{y}_t, \bar{\mathbf{x}}_t$ et $\bar{\mathbf{z}}_t$ se réduisent tous aux estimateurs d'Horvitz-Thompson habituels pour les moyennes de population correspondantes.

3. L'APPROCHE DES VRAISEMBLANCES EMPIRIQUES INDIVIDUELLES

Dans l'approche combinée, le vecteur de moyenne de population inconnu $\overline{\mathbf{Z}}$ est estimé implicitement par l'estimateur du maximum de pseudo-vraisemblance empirique $\hat{\boldsymbol{\theta}}$ à partir de l'échantillon regroupé, comme le montre le premier algorithme présenté à la section 2.1. Certaines difficultés de calcul posées par l'approche combinée sont dues uniquement à l'effort fait pour estimer $\overline{\mathbf{Z}}$ en se servant de $\hat{\boldsymbol{\theta}}$.

Un moyen de contourner cette difficulté consiste à suivre une approche en deux étapes. Supposons que nous remplacions $\hat{\boldsymbol{\theta}}$ par un autre estimateur de $\overline{\mathbf{Z}}$, disons $\bar{\mathbf{z}}$, grâce à l'utilisation de données combinées provenant des deux enquêtes. Puis, nous nous servons des valeurs de $\bar{\mathbf{z}}$ comme valeurs de contrôle pour les contraintes imposées pour l'estimation de la vraisemblance empirique pour chacune des deux enquêtes. De cette façon, non seulement nous assurons la cohérence pour les variables auxiliaires ***z*** communes aux deux enquêtes, mais nous améliorons aussi les estimateurs résultants $\hat{\bar{Y}}_1$ et $\hat{\bar{Y}}_2$ à condition que $\bar{\mathbf{z}}$ soit construit de façon appropriée à partir des données d'échantillon combinées. Cette démarche est semblable au cas de l'échantillonnage à deux phases où la quantité de population inconnue $\overline{\mathbf{Z}}$ est estimée en utilisant le grand échantillon de première phase.

L'estimation de $\bar{Y}_1$ et $\bar{Y}_2$ en se servant d'un estimateur $\bar{\mathbf{z}}$ prédéterminé comme valeur de contrôle se réduit à deux problèmes d'estimation distincts. Par exemple, l'estimateur EL de $\bar{Y}_1$ est donné par $\hat{\bar{Y}}_1 = \sum_{i\in s_1} \hat{p}_1 y_{1i}$, où les $\hat{p}_i$ maximisent $l(\mathbf{p}) = \sum_{i\in s_1} d_{1i} \log(p_i)$ sous les contraintes

$$\sum_{i\in s_1} p_i = 1(p_i > 0), \sum_{i\in s_1} p_i \mathbf{x}_{1i} = \overline{\mathbf{X}}_1 \text{ et } \sum_{i\in s_1} p_i \mathbf{z}_{1i} = \overline{\mathbf{z}}\,.$$

Les poids résultants sont calculés sous la forme $\hat{p}_i = d_{1i}^* / (1 + \boldsymbol{\lambda}' \mathbf{u}_{1i})$, où $d_{1i}^* = d_{1i} / \sum_{i\in s_1} d_{1i}$ et le multiplicateur de Lagrange $\boldsymbol{\lambda}$ est la solution de

$$\sum_{i\in s_1} \frac{d_{1i}^* \mathbf{u}_{1i}}{1 + \boldsymbol{\lambda}' \mathbf{u}_{1i}} = 0 \text{, avec } \mathbf{u}_{1i} = \begin{pmatrix} \mathbf{x}_{1i} - \overline{\mathbf{X}}_1 \\ \mathbf{z}_{1i} - \overline{\mathbf{z}} \end{pmatrix}. \tag{12}$$

Ici, nous pouvons utiliser directement l'algorithme de Chen et coll. (2002) pour obtenir $\boldsymbol{\lambda}$ sans aucune modification.

Le problème le plus important sous cette approche des EL individuelles est le choix de $\overline{\mathbf{z}}$. Renssen et Nieuwenbroek (1997) donnent un excellent compte rendu de l'estimation de $\overline{\mathbf{Z}}$ au moyen des données d'échantillon combinées. Ils proposent une classe générale d'estimateurs de la forme $\overline{\mathbf{z}} = \mathbf{P}\,\overline{\mathbf{z}}_1 + \mathbf{Q}\,\overline{\mathbf{z}}_2$, où ***P*** et ***Q*** sont deux matrices dont les dimensions sont compatibles avec ***z*** telles que ***P*** + ***Q*** = ***I***, et $\overline{\mathbf{z}}_t$ est l'estimateur par la régression généralisée de ***Z*** avec $\mathbf{x}_t$ comme variables auxiliaires. En l'absence de $\mathbf{X}_t$, nous pouvons prendre $\overline{\mathbf{z}}_t$ comme estimateur d'Horvitz-Thompson de $\overline{\mathbf{Z}}$ en utilisant les données provenant de la t^e enquête.

Nous examinerons deux choix de la paire de matrices (***P***, ***Q***) dans l'étude en simulation présentée à la section suivante. Le plus simple est la combinaison proportionnelle où $\mathbf{P} = (n_1 + n_2)^{-1} n_1 \mathbf{I}$ et $\mathbf{Q} = (n_1 + n_2)^{-1} n_2 \mathbf{I}$; la combinaison optimale utilise

$$\mathbf{P} = V_p(\overline{\mathbf{z}}_2)\left[V_p(\overline{\mathbf{z}}_1) + V_p(\overline{\mathbf{z}}_2)\right]^{-1} \text{ et } \mathbf{Q} = V_p(\overline{\mathbf{z}}_1)\left[V_p(\overline{\mathbf{z}}_1) + V_p(\overline{\mathbf{z}}_2)\right]^{-1},$$

où $V_p(\overline{\mathbf{z}}_t)$ est la matrice des variances-covariances fondée sur le plan de sondage de $\overline{\mathbf{z}}_t$. Notons que la matrice ***P*** utilisée dans la combinaison optimale peut aussi s'écrire sous la forme $\mathbf{P} = \left\{\left[V_p(\overline{\mathbf{z}}_1)\right]^{-1} + \left[V_p(\overline{\mathbf{z}}_2)\right]^{-1}\right\}^{-1}\left[V_p(\overline{\mathbf{z}}_1)\right]^{-1}$, et il en va de même de ***Q***. Ce choix est optimal, puisqu'il minimise $V_p(\mathbf{a}'\overline{\mathbf{z}})$ pour un vecteur constant arbitraire ***a*** parmi la classe générale d'estimateurs considérée par Renssen et Nieuwenbroek (1997). Quand on utilise un échantillonnage aléatoire simple pour les deux enquêtes et que $\overline{\mathbf{z}}_t$ sont les moyennes simples d'échantillon, la combinaison optimale se réduit à la combinaison proportionnelle si les deux fractions d'échantillonnage sont les mêmes ou peuvent être ignorées. Il convient de souligner que, pour la combinaison optimale, les matrices ***P*** et ***Q*** doivent être remplacées par des estimations fondées sur l'échantillon pour les applications.

L'approche des EL individuelles est moins élégante que l'approche combinée en ce qui concerne l'estimation du maximum de vraisemblance. Toutefois, elle est intuitivement séduisante et les calculs sont directs et simples. Sous des conditions de régularité appropriées comparables à celles utilisées dans le théorème 1, nous pouvons montrer que l'estimateur EL individuel est asymptotiquement équivalent à l'estimateur par la régression décrit par Renssen et Nieuwenbroek (1997), c.-à-d.

$$\hat{\overline{\mathbf{Y}}}_t = \bar{y}_t + \hat{\mathbf{B}}_{t1}'(\mathbf{X}_t - \overline{\mathbf{x}}_t) + \hat{\mathbf{B}}_{t2}'(\overline{\mathbf{z}} - \overline{\mathbf{z}}_t^*) + o_p\left(n^{-1/2}\right), \tag{13}$$

où $\bar{y}_t = \sum_{i\in s_t} d_{ti}^* y_{ti}$, $\overline{\mathbf{x}}_t = \sum_{i\in s_t} d_{ti}^* \mathbf{x}_{ti}$, $\overline{\mathbf{z}}_t^* = \sum_{i\in s_t} d_{ti}^* \mathbf{z}_{ti}$, $d_{ti}^* = d_{ti} / \sum_{i\in s_t} d_{ti}$, et les coefficients de régression $\hat{\mathbf{B}}_t = \left(\hat{\mathbf{B}}_{t1}', \hat{\mathbf{B}}_{t2}'\right)'$ sont donnés par

$$\hat{\mathbf{B}}_t = \left(\sum_{i \in s_t} d_{ti}^{*} \mathbf{u}_{ti} \mathbf{u}_{ti}' \right)^{-1} \sum_{i \in s_t} d_{ti}^{*} \mathbf{u}_{ti} y_{ti} \ ,$$

où $\mathbf{u}_{1i}$ (et $\mathbf{u}_{2i}$ en forme évidente) sont définis par (12). Les deux premiers termes du deuxième membre de (13) peuvent être considérés comme un estimateur par la régression généralisée de $\overline{Y}_t$ fondé sur les variables auxiliaires $\mathbf{x}_t$, et le troisième est un terme d'ajustement destiné à améliorer davantage l'estimateur par la régression grâce à l'utilisation de l'information supplémentaire sur les variables ***z***.

Il mérite d'être souligné que, pour l'estimateur EL individuel, il est facile de tenir compte des tailles d'échantillon différentes pour l'estimation de $\overline{\mathbf{Z}}$. En revanche, l'approche de la EL combinée ne le permet pas et nécessite un redressement spécial de la pondération pour atteindre le même but. Nous ne poursuivrons pas les développements dans cette direction ici, mais cet argument offre une explication possible du fait que l'estimateur EL combiné donne souvent de moins bons résultats que l'estimateur individuel, comme le montrent les résultats de l'étude en simulation présentés à la section suivante.

4. ÉTUDE EN SIMULATION

À la présente section, nous examinons la performance en échantillon fini des estimateurs proposés grâce à une étude en simulation limitée. La population finie utilisée pour l'étude est fondée sur les données réelles pour la province de l'Ontario provenant de l'Enquête sur les dépenses des familles (EDF) de 1996 réalisée par Statistique Canada. L'ensemble de données contient N=2 396 observations couvrant diverses caractéristiques. Les variables pertinentes pour l'étude incluent x_1 : nombre d'enfants (moins de 15 ans); x_2 : nombre de jeunes (15 à 24 ans); x_3 : nombre de personnes dans le ménage; *z*: revenu total après impôt; *y*: dépenses totales.

Dans la simulation, nous traitons l'ensemble de données proprement dit comme une population finie. Nous subdivisons cette population en huit strates conformément au plan de sondage original. Pour la première enquête, nous utilisons le nombre d'enfants (x_1) et le nombre de personnes (x_3), dont les moyennes de population sont connues, comme variables de contrôle et traitons les dépenses totales *y* comme la variable de réponse. Nous supposons aussi que les variables x_2 et x_3 sont utilisées comme variables de contrôle dans la deuxième enquête et que, commodément, l'information sur le revenu total (*z*) est recueillie pour les deux enquêtes, mais que la moyenne de population $\overline{Z}$ est inconnue. Notre objectif est d'estimer la moyenne de population $\overline{Y}$ en utilisant toute l'information utile dont nous disposons, tout en respectant les exigences de cohérence imposées sur la variable *z* pour les deux enquêtes.

Pour chaque exécution de la simulation, nous tirons un échantillon aléatoire stratifié de taille n_t sous répartition proportionnelle pour la t^e enquête, $t = 1, 2$, et nous calculons trois estimateurs du maximum de pseudo-vraisemblance empirique pour $\overline{Y}$. Soit EL(C) l'estimateur EL combiné, EL(SP) l'estimateur EL individuel fondé sur la combinaison proportionnelle pour l'estimation de $\overline{Z}$ et EL(SO) l'estimateur EL individuel utilisant la combinaison optimale pour l'estimation de $\overline{Z}$. Pour chaque simulation, nous calculons aussi les trois estimateurs de type GREG : celui proposé par Zieschang (1990), représenté par GR(Z), qui est équivalent à notre estimateur EL combiné EL(C), et les estimateurs proposés par Renssen et Nieuwenbroek (1997), représentés par GR(RN1) et GR(RN2), qui correspondent à nos estimateurs EL(SP) et EL(SO), respectivement. La matrice Λ utilisée pour formuler l'estimateur GR(Z) est prise comme étant $diag(d_1,...d_n)$. Nous répétons le processus indépendamment $B = 1\ 000$ fois.

Nous évaluons les performances d'un estimateur $\hat{\overline{Y}}$ d'après le biais relatif (RB) et l'efficacité relative (RE) simulés définis comme étant

$$RB = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{\bar{Y}}(b)-\bar{Y}}{\bar{Y}} \text{ et } RE = \frac{MSE\left(\hat{\bar{Y}}_0\right)}{MSE\left(\hat{\bar{Y}}\right)},$$

où $\hat{\bar{Y}}(b)$ est l'estimateur $\hat{\bar{Y}}$ calculé d'après le b^e échantillon simulé, $MSE(\hat{\bar{Y}}) = B^{-1} \sum_{b=1}^{B}\left[\hat{\bar{Y}}(b)-\bar{Y}\right]^2$ et $\hat{\bar{Y}}_0$ est l'estimateur de base utilisé pour la comparaison. Dans notre étude, $\hat{\bar{Y}}_0$ correspond à l'estimateur par la régression généralisée (GREG) de $\bar{Y}$ en utilisant x_1 et x_3 comme variables auxiliaires. Il convient de souligner que l'information d'échantillon sur z ne peut être utilisée ici pour l'estimation par la régression de $\bar{Y}$, puisque nous supposons que $\bar{Z}$ est inconnu. Les tailles d'échantillon n_t = 80, 160 et 240 utilisées dans la simulation représentent des fractions d'échantillonnage typiques de 2,5 %, 5 % et 10 %.

Tableau 1 : Efficacité relative simulée fondée sur les données de l'EDF de 1996 de Statistique Canada

n_1	n_2	GREG	EL(C)	GR(Z)	EL(SP)	GR(RN1)	EL(SO)	GR(RN2)
80	80	1,00	1,19	1,13	1,28	1,28	1,28	1,28
	160	1,00	1,36	1,28	1,31	1,43	1,31	1,42
	240	1,00	1,42	1,31	1,43	1,49	1,48	1,49
160	80	1,00	1,03	0,91	1,15	1,15	1,15	1,15
	160	1,00	1,28	1,10	1,27	1,26	1,27	1,27
	240	1,00	1,33	1,18	1,29	1,28	1,30	1,30
240	80	1,00	0,91	0,80	1,16	1,14	1,15	1,13
	160	1,00	1,11	0,95	1,16	1,14	1,16	1,15
	240	1,00	1,24	1,07	1,22	1,20	1,24	1,23

Les valeurs absolues des biais relatifs simulés sont inférieures à 0,1 % et ne sont pas présentées ici. Le tableau 1 donne l'efficacité relative des estimateurs EL et de type GREG sous divers scénarios de combinaisons de tailles d'échantillon. Les principaux résultats se résument comme suit :

i) les deux estimateurs EL individuels ont des propriétés comparables et donnent tous deux de bons résultats, le gain d'efficacité étant plus prononcé lorsque la taille du deuxième échantillon est plus grande;

ii) l'estimateur EL combiné donne des résultats satisfaisants lorsque la taille du deuxième échantillon est compatible, mais ces résultats pourraient se détériorer autrement (c.-à-d. le cas où n_1 =240 et n_2 =80);

iii) les estimateurs GREG de Renssen et Nieuwenbroek (1997) donnent des résultats comparables à ceux des estimateurs EL individuels, mais l'estimateur GREG de Zieschang (1990) est supplanté par l'estimateur EL combiné dans tous les cas;

iv) l'utilisation de l'information sur la variable auxiliaire commune z produit une amélioration considérable comparativement à l'estimateur GREG de référence lorsque le deuxième échantillon n'est pas trop petit.

5. CONCLUSION

Le redressement des poids pour satisfaire à certaines exigences d'efficacité et de cohérence est un thème constant en échantillonnage, et disposer de poids redressés positivement est une propriété fort souhaitable pour les utilisateurs des fichiers de microdonnées de production, où les poids sont considérés comme étant le nombre d'unités dans la population finie représentées par l'unité échantillonnée. Des poids positifs assurent aussi d'obtenir une estimation positive pour des quantités de population positive connue.

Il convient de souligner qu'en principe, on peut obtenir des poids positifs dans l'estimation par la régression grâce à une minimisation sous contraintes dans le contexte de l'estimation par calage tel que discuté dans Deville et Särndal (1992). L'application pratique d'une telle méthode n'est toutefois pas simple et nécessite souvent des approximations de circonstance. La perte d'efficacité due à ces approximations est habituellement inconnue. La méthode de la vraisemblance empirique, par ailleurs, offre un moyen naturel de procéder à cette application avec des poids redressés finaux qui sont intrinsèquement positifs. Il convient aussi de souligner que la quantité totale de temps requise pour calculer les estimateurs EL proposés reste limitée. Dans notre étude en simulation, il faut moins de 20 secondes sur un poste de travail Sun/Unix à processeur double pour calculer l'estimateur EL combiné quand $n_1 = n_2 = 240$ et que le programme est écrit en R/Splus.

Alors que les poids redressés selon une technique de type GREG ont tendance à présenter certaines valeurs faibles ou négatives, ceux obtenus selon l'approche EL peuvent, à l'occasion, contenir quelques valeurs élevées. Pour les résultats de simulation présentés à la section 4, où nous avons utilisé un scénario de répartition proportionnelle de la taille de l'échantillon pour tirer les deux échantillons aléatoires stratifiés, les poids g donnés par $g_i = w_i / d_i$ se situent tous dans la fourchette de (0,25, 4,00), où w_i représente les poids redressés par la méthode EL et d_i représente les poids de sondage de référence. Plus de 99 % de ces poids g sont en effet compris entre 0,50 et 2,00. Si nous utilisons un scénario de répartition non équilibré où la strate la plus grande $(N_h = 763)$ et la plus petite $(N_h = 33)$ se voient attribuer une même taille d'échantillon, nous observons que quelques poids g peuvent être supérieurs à 4,00, voire 6,00. Théoriquement, cela ne pose pas de problème en ce qui concerne les propriétés statistiques des estimateurs EL. Les utilisateurs que les poids de valeur élevée préoccupent également peuvent appliquer l'idée du relâchement minimal des contraintes présentée dans Chen et coll. (2002) pour obtenir une fourchette plus générale de poids restreints grâce à la méthode EL.

Il est facile de soutenir, tant théoriquement qu'empiriquement, que, dans des conditions idéales, le gain d'efficacité dû à l'utilisation des données d'échantillon combinées est presque garanti. Les cas où l'exigence de cohérence forcée sur les variables communes risque vraisemblablement de nuire aux estimateurs résultants incluent 1) les erreurs non dues à l'échantillonnage graves non contrôlées, 2) les plans d'échantillonnage ou les répartitions de taille d'échantillon fortement déséquilibrés, 3) les populations cibles dont la conception est erronée, 4) une corrélation faible, voire inexistante, entre les variables communes et les variables de réponse et 5) l'utilisation de variables communes douteuses.

La bonne utilisation de la méthode EL proposée pour combiner l'information sur les variables auxiliaires communes dans le cas d'enquêtes réelles exige des examens minutieux à l'étape de la planification et des jugements prudents à l'étape de l'estimation. Comme l'ont souligné Renssen et Nieuwenbroek (1997), il n'est pas facile de trouver des variables communes au sens strict du terme à cause des différences entre les définitions, les méthodes d'observation et les périodes de référence. Il est possible de réduire ce genre de complications en harmonisant les enquêtes visées à l'étape de la conception. Ainsi, dans le cas du plan à questionnaire scindé, où certaines questions communes figurent dans les deux versions du questionnaire, il convient d'accorder de l'attention à l'ordre et au positionnement des questions communes afin de réduire le biais de réponse ou l'effet de report éventuel. Dans le cas des enquêtes sur des populations humaines réalisées régulièrement au cours du temps, des variables comme le sexe, l'âge, le niveau de scolarité, etc., peuvent facilement être considérées comme des variables communes s'il est possible de ne pas tenir compte de l'évolution de la dynamique de la population au cours de certaines périodes. D'autres variables, telles que la « situation d'emploi en mai 2003 », qui peut être mesurée « directement » durant une enquête réalisée en juin 2003, mais nécessite des réponses « remémorées » pour des enquêtes réalisées à une période ultérieure, doivent être traitées avec prudence. Les variables de ce type ne peuvent être considérées comme des variables communes que si les réponses « remémorées » sont aussi exactes que la mesure « directe ».

REMERCIEMENTS

L'étude a été financée par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Nous exprimons notre reconnaissance au professeur Jiahua Chen pour ses commentaires et suggestions.

RÉFÉRENCES

Chen, J. et Sitter, R.R. (1999). A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys. *Statistica Sinica*, 9, 385-406.

Chen, J., Sitter, R.R. et Wu, C. (2002). Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys. *Biometrika*, 89, 230-237.

Deville, J.C. et Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Hartley, H.O. et Rao, J.N.K. (1968). A New Estimation Theory for Sample Surveys. *Biometrika*, 55, 547-557.

Isaki, C.T. et Fuller, W.A. (1982). Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89-96.

Owen, A.B. (1990). Empirical Likelihood Ratio Confidence Regions. *Annals of Statistics*, 18, 90-120.

Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall / CRC.

Renssen, R.H. et Nieuwenbroek, N.J. (1997). Aligning Estimates for Common Variables in Two or More Sample Surveys. *Journal of the American Statistical Association*, 92, 368-374.

Zieschang, K.D. (1990). Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.