Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherché économique et sociale

MÉTHODES DU CALAGE OPTIMAL ET DE LA VRAISEMBLANCE EMPIRIQUE EN ÉCHANTILLONNAGE

C. Wu[1]

RÉSUMÉ

La méthode de calage est mentionnée de plus en plus souvent dans les publications sur l'échantillonnage et nombre d'organismes d'enquête calculent régulièrement des estimateurs par calage. Cependant, quelle que soit la méthode utilisée, le choix des variables de calage demeure ponctuel. Dans le présent article, nous montrons que l'estimateur par calage d'un modèle de la moyenne d'une population finie qu'ont proposé Wu et Sitter (2001) par raisonnement intuitif est en effet optimal parmi une classe d'estimateurs par calage. En outre, nous présentons des estimateurs par calage optimaux pour la fonction de distribution d'une population finie, la variance de population, la variance d'un estimateur linéaire et d'autres fonctions quadratiques de population finie établis dans un cadre de référence unifié. Au moyen d'une étude en simulation limitée, nous montrons que l'amélioration de ces estimateurs optimaux par rapport aux estimateurs conventionnels peut être considérable. Nous abordons clairement la question de savoir quand et comment on peut utiliser des données auxiliaires aussi bien pour l'estimation de la moyenne de population au moyen de l'estimateur par la régression généralisée que pour celle de la variance de cet estimateur par calage dans le contexte de la méthode générale proposée. Nous examinons aussi certaines questions fondamentales relatives à l'utilisation d'information auxiliaire provenant de données d'enquête dans le contexte de l'estimation optimale.

Certains mots-clés : Variance asymptotique due au plan de sondage; information auxiliaire; calage du modèle; estimation optimale; superpopulation.

1. INTRODUCTION

La notion des estimateurs par calage a été introduite par Deville et Särndal (1992) dans le contexte de l'utilisation d'information auxiliaire tirée de données d'enquête. Supposons que U = {1, 2, ... , N} est l'ensemble d'étiquettes pour la population finie. Soit (y_i, x_i) les valeurs de la variable étudiée y et du vecteur de variables auxiliaires x liées à la i^{e} unité. La question est celle de savoir comment estimer efficacement $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ en utilisant les totaux connus de population $X = \sum_{i=1}^{N} x_i$ à l'étape de l'estimation. Soit s = {1, 2, ..., n} l'ensemble d'unités échantillonnées sous un plan général d'échantillonnage, p, et $\pi_i = P(i \in s)$ les probabilités d'inclusion de premier ordre. L'estimateur par calage classique pour $\bar{Y}$ est défini par $\hat{\bar{Y}}_C = N^{-1}\sum_{i\in s} w_i y_i$, où les w_i sont modifies par rapport aux poids d'échantillonnage de base $d_i = 1/\pi_i$ par minimisation d'une mesure de distance, Φ_s entre les w_i et les d_i , conditionnellement aux contraintes

$$\sum_{i\in s} w_i x_i = \sum_{i=1}^{N} x_i \tag{1.1}$$

La mesure de distance utilisée le plus couramment est la distance du chi-deux

$$\Phi_s = \sum_{i\in s} (w_i - d_i)^2 / (q_i d_i),$$

où les qi sont des constants positives connues non corrélées aux d_i. D'autres mesures de distance peuvent également être envisagées. Voir Deville et Särndal (1992) pour une discussion détaillée.

[1] Department of Statistics and Actuarial Science, University of Waterloo, cbwu@uwaterloo.ca

La construction d'estimateurs par calage comporte deux composantes fondamentales, à savoir une mesure de distance et un ensemble d'équations de calage. Le choix de la mesure de distance n'est pas aussi critique que celui des équations de calage, puisque les estimateurs résultants sont tous asymptotiquement équivalents à ceux fondés sur une distance du chi-deux pour un certain q_i (Deville et Särndal, 1992). Les équations de calage (1.1) sont utilisées régulièrement par nombre d'organismes d'enquête qui les qualifient de contraintes de base. En pratique, ces contraintes sont souvent imposées pour deux raisons : i) l'enquêteur estime que les poids qui produisent des estimations parfaites pour les variables auxiliaires devraient également donner de bonnes estimations pour la variable étudiée; ii) l'information auxiliaire n'est disponible qu'au niveau agrégé, c.-à-d. uniquement si X est connu. Parfois, les statisticiens qui se spécialisent dans des domaines tels que la démographie insistent pour que le calage afin d'établir une concordance avec les totaux connus d'après le recensement se fasse sur un grand nombre de variables, au risque de réduire l'efficacité des estimateurs. Par ailleurs, si l'on dispose de données auxiliaires $x_1, \ldots, x_N$ complètes, ce qui est le cas dans nombre d'enquêtes, une question fort intéressante est celle de savoir quelle est la meilleure équation de calage à utiliser pour construire un estimateur par calage.

Soit $u_i = u(x_i)$, $i = 1, \ldots, N$, où u(·) est une fonction réelle. Si nous remplaçons (1.1) par

$$\sum_{i \in s} w_i u(x_i) = \sum_{i=1}^{N} u(x_i), \tag{1.2}$$

la question qui se pose alors est de savoir quel choix de $u(\cdot)$ rendra $\hat{\bar{Y}}_C$ plus efficace. Notons que les contraintes de base (1.1) correspondent à k équations, où k est le nombre de composantes dans x, tandis que la contrainte (1.2) ne comporte qu'une seule équation contenant la variable unique de réduction des données $u = u(x)$. Wu et Sitter ont montré (2001, partie (3) du théorème 1) que l'estimateur par calage classique de la moyenne ou du total d'une population finie au moyen de (1.1) est identique à l'estimateur par calage fondé sur (1.2) lorsqu'on choisit une $u(\cdot)$ particulière. La contrainte à une seule équation (1.2) où $u(\cdot)$ est non spécifiée est plus générale que les contraintes de base à k équations fixes.

Il est bien connu, en échantillonnage, qu'il n'existe pas d'estimateur uniformément sans biais et à variance minimale dans le cadre de travail fondé sur le plan de sondage. En effet, le seul choix de $u(\cdot)$ qui aboutit à un $\hat{\bar{Y}}_C$ dont la variance est minimale est $u(x_i) \equiv y_i$, ce qui n'est, naturellement, d'aucune utilité en pratique.

Les estimateurs optimaux assistés par modèle fondés sur le critère de variance due au plan de sondage prévue minimale $E_\xi\left\{V_p\left(\hat{\bar{Y}}\right)\right\}$ sous une superpopulation ont été discutés par plusieurs auteurs. Consulter, par exemple, les travaux de Godambe (1955), Godambe et Thompson (1973), Cassel, Särndal et Wretman (1976), et Isaki et Fuller (1982). La variance due au plan de sondage prévue a également été appelée « variance anticipée » par Isaki et Fuller (1982). Notons que E_p et V_p représentent l'espérance et la variance sous le plan d'échantillonnage, p, et E_ξ et V_ξ représentent l'espérance et la variance sous le modèle de superpopulation ξ.

Dans le présent article, nous utilisons un critère comparable. Les estimateurs par calage appartiennent à la classe des estimateurs non linéaires et ni leur variance due au plan de sondage exacte ni leur erreur quadratique moyenne n'a une expression explicite. Aux fins d'optimalité, une solution naturelle consiste à minimiser l'espérance fondée sur le modèle de la variance asymptotique due au plan de sondage $E_\xi\left\{AV_p\left(\hat{\bar{Y}}\right)\right\}$, où AV_p représente la variance asymptotique fondée sur le plan de sondage. Puisque le biais $B_p\left(\hat{\bar{Y}}_C\right) = E_p\left(\hat{\bar{Y}}_C - \bar{Y}\right)$ d'un estimateur par calage $\hat{\bar{Y}}_C$ satisfait à $B_p\left(\hat{\bar{Y}}_C\right) = o\left(n^{-1/2}\right)$ et $V_p\left(\hat{\bar{Y}}_C\right) = O\left(n^{-1}\right)$, minimiser $E_\xi\left\{AV_p\left(\hat{\bar{Y}}_C\right)\right\}$ équivaut à minimiser $E_\xi\left\{E_p\left(\hat{\bar{Y}}_C - \bar{Y}\right)^2\right\}$ asymptotiquement.

À la section 2, nous montrons que l'estimateur par calage d'un modèle de la moyenne de population finie qu'ont propose Wu et Sitter (2001) par raisonnement intuitif est effectivement optimal parmi une classe d'estimateurs par calage, au sens de la variance asymptotique due au plan de sondage prévue minimale sous un modèle de superpopulation et tout plan d'échantillonnage régulier. Le résultat offre un cadre unifié pour la construction d'estimateurs par calage optimaux de la fonction de distribution d'une population finie, de la variance de population, de la variance d'un estimateur linéaire et d'autres fonctions quadratiques de population finie. Les estimateurs par calage optimaux de la fonction de distribution sont présentés à la section 3 et ceux d'une quantité de deuxième ordre générale de population finie, au paragraphe 4. Toujours dans ce paragraphe, nous abordons clairement la question de savoir quand et comment l'information auxiliaire peut être utilisée pour l'estimation de la moyenne de population au moyen d'un estimateur par la régression généralisée, ainsi que pour l'estimation de la variance de cet estimateur par calage dans le contexte du cadre de travail unifié. Les estimateurs optimaux du pseudo-maximum de vraisemblance empirique, qui sont asymptotiquement équivalents aux estimateurs par calage optimaux, sont particulièrement utiles pour estimer la fonction de distribution, la variance de population et d'autres quantités connues non négatives. Les résultats d'une étude en simulation limitée des propriétés en échantillon fini fondées sur le plan de sondage de ces estimateurs optimaux comparativement à celle des estimateurs classiques sont présentés à la section 5. Enfin, à la section 6, nous examinons certains problèmes fondamentaux que pose l'utilisation d'information auxiliaire tirée de données d'enquête dans le cadre de travail proposé et nous présentons certains conclusions.

2. OPTIMALITÉ DE L'ESTIMATEUR PAR CALAGE D'UN MODÈLE

Pour le cadre de travail asymptotique, nous supposons qu'il existe une série de populations finies, représentées par l'indice v. La taille de population et la taille d'échantillon pour ces populations sont représentées par N_v et n_v. quand $v \to \infty$, $N_v \to \infty$ et $n_v \to \infty$. Tous les processus de limite devraient être interprétés comme signifiant $v \to \infty$. Pour alléger la notation, nous supprimerons l'indice v dans la suite de l'exposé. Pour une formulation détaillée de ce cadre asymptotique, consulter Isaki et Fuller (1982). Nous considérons des situations où l'information auxiliaire complète $x_1, \cdots, x_N$ est disponible.

Supposons que $y_1, y_2, \cdots, y_N$ est un échantillon aléatoire d'une superpopulation ξ tel que

$$E_\xi(y_i \mid x_i) = \mu(x_i, \theta), V_\xi(y_i \mid x_i) = \{v(x_i)\}^2 \sigma^2, i = 1,2,..., N. \tag{2.1}$$

Ici, θ et σ^2 sont des paramètres du modèle et θ pourrait être évalué sur vecteur $\mu(\cdot, \cdot)$ et $v(\cdot)$ sont des fonctions connues, $y_1, y_2,..., y_N$ sont conditionnellement indépendants les uns des autres pour les x_i donnés.

Soit $\hat{\bar{Y}}_C$ un estimateur par calage de $\bar{Y}$ quand on utilise $C = \{u(x_1), u(x_2),...\}$ dans (1·2). Soit $\boldsymbol{L}$ l'ensemble de séries $C = \{u(x_1), u(x_2),...\}$ pour toutes les fonctions imaginables $u(\cdot)$ tels que $N^{-1}\sum_{i=1}^N \{u(x_1)\}^2 \to c \neq 0$ quand $N \to \infty$. Cette condition de moment de deuxième ordre fini sur la série $C \in \boldsymbol{L}$, qui n'est pas très contraignante, est nécessaire dans les preuves. Nous supposons que $\{\mu(x_1, \theta), \mu(x_2, \theta),...\} \in \boldsymbol{L}$ et que $\{v(x_1), v(x_2),...\} \in \boldsymbol{L}$.

Un plan d'échantillonnage est dit régulier s'il donne lieu à une taille fixe d'échantillon, que les probabilités d'inclusion π_i et π_{ij} sont indépendantes de la variable dépendante y et qu'il satisfait à

(C1) $\max_{i \in s} nd_i / N = O(1)$.

(C2) $N^{-1}\sum_{i \in s} d_i u_i - N^{-1}\sum_{i=1}^N u_i = O_p\left(n^{-1/2}\right)$ pour toute série $(u_1, u_2,...) \in \boldsymbol{L}$.

La condition (C1) énonce simplement qu'aucun poids d'échantillonnage de base n'est anormalement grand. La condition (C2) en découle si l'estimateur d'Horvitz-Thompson pour $\bar{u}_N = N^{-1}\sum_{i=1}^N u_i$ suit asymptotiquement une loi normale.

Théorème 1. Parmi la classe d'estimateur par calage $\hat{\bar{Y}}_C$ avec $C=\{u(x_1),u(x_2),...\} \in \boldsymbol{L}$, *le choix de* $C=\{\mu(x_1,\theta),\mu(x_2,\theta),...\}$ *minimise* $E_\xi\left\{AV_p\left(\hat{\bar{Y}}_C\right)\right\}$ *sous le modèle (2·1) et tout plan d'échantillonnage régulier.*

Preuve: Voir l'annexe.

En pratique, le paramètre du modèle θ doit être remplacé par une estimation fondée sur l'échantillonnage $\hat{\theta}$. On peut montrer que le remplacement de θ par un estimateur convergent par rapport au plan de sondage $\hat{\theta}$ ne modifie pas l'estimateur $\hat{\bar{Y}}_C$ asymptotiquement. Wu et Sitter (2001) ont appelé l'estimateur résultant, $\hat{\bar{Y}}_{MC}$, estimateur par calage d'un modèle de $\bar{Y}$. Ils ont proposé $\hat{\bar{Y}}_{MC}$ par raisonnement intuitif et ont montré qu'il s'agit d'un moyen plus général et plus efficace de construire des estimateurs par calage si l'on dispose d'information auxiliaire complète. Toutefois, ils n'ont pas étudié l'optimalité de $\hat{\bar{Y}}_{MC}$, dans leur article.

Cette approche du calage optimal peut être appliquée à la méthode de la pseudo-vraisemblance empirique (Chen et Sitter, 1999) pour obtenir un estimateur optimal présentant des caractéristiques intéressantes. Soit $\hat{\bar{Y}}_{EC}$ l'estimateur du pseudo-maximum de vraisemblance empirique de $\bar{Y}$ (Chen & Sitter, 1999) obtenu par calage sur $C=\{u(x_1),u(x_2),...\}$. Autrement dit, $\hat{\bar{Y}}_{EC}=\sum_{i\in s}\hat{p}_i y_i$, où les $\hat{p}_i$ maximisent la fonction de la pseudo log-vraisemblance empirique $l(p)=\sum_{i\in s} d_i \log(p_i)$ conditionnellement aux contraintes

$$\sum_{i\in s} p_i = 1\,(0<p_i<1),\ \sum_{i\in s} p_i u(x_i)=\frac{1}{N}\sum_{i=1}^{N} u(x_i). \tag{2.2}$$

On obtient l'estimateur à modèle calé du pseudo-maximum de vraisemblance empirique $\hat{\bar{Y}}_{ME}$ de $\bar{Y}$ lorsqu'on utilise $C=\{\mu(x_1,\hat{\theta}),\mu(x_2,\hat{\theta}),...\}$ dans les contraintes (2.2).

Théorème 2. Parmi la classe d'estimateurs du pseudo-maximum de vraisemblance empirique $\hat{\bar{Y}}_{EC}$ *pour* $\bar{Y}$ *où* $C \in L$, *le choix de* $C=\{\mu(x_1,\theta),\mu(x_2,\theta),...\}$ *minimise* $E_\xi\left\{AV_p\left(\hat{\bar{Y}}_{EC}\right)\right\}$ *sous le modèle (2.1) et tout plan d'échantillonnage régulier.*

Preuve: Voir l'annexe.

L'estimateur à modèle calé du pseudo-maximum de vraisemblance empirique pour $\bar{Y}$ est asymptotiquement équivalent à l'estimateur par calage d'un modèle et est optimal sous les mêmes conditions. Chen et coll. (2002) ont développé des algorithme simples pour le calcul de l'estimateur $\hat{\bar{Y}}_{ME}$. Cependant, la caractéristique la plus intéressante de cet estimateur $\hat{\bar{Y}}_{ME}$, tient aux propriétés intrinsèques des poids, c.-à-d. $\hat{p}_i>0$ et $\sum_{i\in s}\hat{p}_i=1$. Ces propriétés sont particulièrement utiles quand on étend la méthode à l'estimation de la fonction de distribution et à d'autres quantités connues non négatives. L'approche du calage optimal fournit aussi un cadre de travail unifié pour l'estimation efficace de la fonction de distribution, de la variance de population et d'autres fonctions quadratiques de population finie. Nous décrivons cet aspect en détail à la section 3 pour la fonction de distribution et à la section 4 pour l'estimation de la variance et d'autres fonctions quadratiques.

3. ESTIMATEURS PAR CALAGE OPTIMAUX DE LA FONCTION DE DISTRIBUTION

La fonction de distribution d'une population finie $F_Y(t) = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$ est aussi une moyenne de population finie définie sur une variable indicatrice $z_i = I(y_i \le t)$. Ici, $z_i = 1$ si $y_i \le t$ et 0 autrement. Si l'on utilise aucune information auxiliaire, l'estimation de $F_Y(t)$ devient un cas particulier de l'estimation de la moyenne de population et est habituellement simple. En présence d'information auxiliaire, il faut accorder une attention particulière aux éléments suivants :

(a) si l'imposition des contraintes de base (1.1) calées directement sur les x variables se justifient parfois pour l'estimation de $\bar{Y}$, cette exigence de convergence n'est pas nécessaire pour l'estimation de $F_Y(t)$, car l'efficacité est la préoccupation principale.

(b) c'est avec la variable indicatrice $z_i = I(y_i \le t)$ qu'il faut travailler; l'estimation $F_Y(t)$ pose aussi un problème d'efficacité locale (valeur particulière de t) par opposition à l'efficacité globale (une valeur arbitraire de t).

(c) il est souhaitable qu'un estimateur de $F_Y(t)$, disons $\hat{F}_Y(t)$, soit aussi une fonction de distribution, de sorte qu'on puisse obtenir les estimations des quantiles par inversion directe de $\hat{F}_Y(t)$.

Appliquées directement à l'estimation de $F_Y(t)$, nombre de techniques d'estimation de $\bar{Y}$ produisent des résultats insatisfaisants. Par exemple, dans le cas d'une variable scalaire x, un estimateur par la régression de $F_Y(t)$ aura la forme $\hat{F}_{RE}(t) = \hat{F}_Y(t) + \{F_X(t) - \hat{F}_X(t)\}\hat{B}$, où $\hat{F}_Y(t)$ et $\hat{F}_X(t)$ sont les estimateurs de type Horvitz-Thompson de $F_Y(t)$ et $F_X(t) = N^{-1}\sum_{i=1}^{N} I(x_i \le t)$. Le terme $\hat{B}$ est la pente estimée de la régression de $I(y_i \le t)$ sur $I(x_i \le t)$. $\hat{F}_{RE}(t)$ présente plusieurs inconvénients, le plus évident étant qu'il n'est pas une fonction de distribution et qu'il peut prendre des valeurs en dehors de [0, 1].

Ici, nous pouvons appliquer facilement la méthode d'estimation à modèle calé de la pseudo-vraisemblance empirique pour obtenir des estimateurs de $F_Y(t)$ qui sont non seulement efficaces, mais aussi, eux-mêmes, de vraies fonctions de distribution. La variable de calage optimal $\mu(x_i, \theta)$ devrait maintenant être remplacée par $g(x_i, t) = E_\xi\{I(y_i \le t) \mid x_i\} = P(y_i \le t \mid x_i)$. Nous pouvons considérer deux types de modèles de travail pour obtenir $g(x_i, t)$: ceux qui établissent la relation entre y_i et x_i ou ceux qui établissent la relation entre la variable indicatrice $I(y_i \le t)$ et les x_i.

Sous le modèle par la régression utilisé habituellement,

$$y_i = x_i'\theta + v(x_i)\varepsilon_i,\ i = 1,2,\ldots,N, \tag{3.1}$$

où les ε_i sont des variables aléatoires indépendantes et identiquement distribuées de moyenne 0 et de variance σ^2. Soit $G(\cdot)$ la fonction de distribution cumulative des ε_i. Nous avons

$$g(x_i, t) = P(y_i \le t \mid x_i) = G\{(t - x_i'\theta)/v(x_i)\}.$$

Comme dans le cas de la moyenne, dans les applications, il faudra remplacer le paramètre du modèle θ par une estimation sur échantillon convergente par rapport au plan de sondage.

Notons que les $g_i = g(x_i, t)$ sont des probabilités; un autre processus de modélisation consiste à utiliser un modèle de régression logistique

$$\log\left(\frac{g_i}{1-g_i}\right) = x_i'\theta, \tag{3.2}$$

avec la fonction de variance habituelle V *(g) = g(1 - g).* Sous le modèle (3.2) nous avons $g(x_i,t)=\exp(x_i'\theta)/\{1+\exp(x_i'\theta)\}$. Soit $\hat{F}_{EC}(t)=\sum_{i\in s}\hat{p}_i I(y_i\le t)$, où les $\hat{p}_i$ maximisent *l(p)* conditionnellement à la contrainte (2.2) avec $C=\{u(x_1),u(x_2),...\}$.

Théorème 3. L'estimateur du pseudo-maximum de vraisemblance empirique $\hat{F}_{ME}(t)$ *calé sur* $\{g(x_1,t),g(x_2,t),...\}$ *est optimal parmi la classe d'estimateurs* $\hat{F}_{EC}(t)$ *avec C* $\in$ ***L*** *sous le modèle de travail (3.1) ou (3.2) et tout plan d'échantillonnage régulier.*

Preuve: Le résultat découle directement du théorème 2 si l'on remplace y_i par $I(y_i\le t)$ et $\mu(x_i,\theta)$ par $g(x_i, t)$.

Les propriétés fondées sur le plan de sondage et les propriétés en petit échantillon de ces estimateurs, ainsi que le problème connexe de l'estimation des quantiles ont été étudiés par Chen et Wu (2002) dans un autre article. Nous qualifierons l'estimateur de F_Y *(t)* calé directement sur les *x* variables (Chen et Sitter, 1999) d'estimateur par calage classique. Comme nous le verrons au paragraphe 5, les estimateurs par calage optimaux sont nettement plus efficaces.

Il convient de souligner que les deux modèles de travail (3.1) et (3.2) ne sont pas compatibles. Par conséquent, l'optimalité de l'estimateur résultant n'est significative que sous le modèle choisi. Il faut aussi noter que la variable optimale de calage $g(x_i, t)$ dépend de *t.* Aucun ensemble unique de poids $\hat{p}_i$ ne produira un estimateur optimal pour une valeur arbitraire de *t.* Chen et Wu (2002) proposent d'utiliser une valeur fixe t_o dans $g(x_i, t)$ et d'utiliser les poids résultants pour tout *t* dans $\hat{F}_{ME}(t)$. Dans ces conditions, $\hat{F}_{ME}(t)$ est une fonction de distribution réelle. Chen et Wu (2002) démontrent par une étude en simulation que l'estimateur $\hat{F}_{ME}(t)$ ainsi construit est très efficace pour les valeurs de *t* dans un voisinage étendu de t_o. Il est facile de déterminer la valeur réelle de t_o qui maximise l'efficacité de l'estimateur résultant lorsqu'on s'intéresse à un certain voisinage de t_o.

4. ESTIMATION OPTIMALE DE LA VARIANCE ET D'AUTRES FONCTIONS QUADRATIQUES

L'estimation de la variance et d'autres quantités de population finie de deuxième ordre en se servant d'information auxiliaire a été abordée par nombre de spécialistes des enquêtes. Ceux-ci ont essayé d'appliquer diverses techniques, dont l'estimation par la régression, l'estimation par le quotient et l'estimation par calage. Voir Sitter et Wu (2002) pour une synthèse de la littérature. Un point faible commun à ces méthodes est l'argument ponctuel consistant à appliquer certaines techniques, mises au point au départ pour estimer $\bar{Y}$, à l'estimation de la variable ou d'autres paramètres de population de deuxième ordre sans adopter de cadre de travail commun unifiant les deux types de paramètres de population finie.

La méthode du calage optimal d'un modèle et celle du pseudo-maximum de vraisemblance empirique fondées sur un modèle calé peuvent être étendues au calcul des variances et d'autres paramètres de population finie de deuxième ordre selon une approche de traitement par lot. Pour les paramètres dont la forme générale est $Q=\sum_{i=1}^{N}\sum_{j=i+1}^{N}\phi(y_i,y_j)$, qui inclut la variance de population $S^2=N^{-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2=\{N(N-1)\}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N}(y_i-y_j)^2$ et la variance de l'estimateur d'Horvitz-Thompson $V_p(\hat{Y}_{HT})=\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\pi_i\pi_j-\pi_{ij})(y_i/\pi_i-y_j/\pi_j)^2$ en tant que cas particuliers, il est possible d'élaborer une stratégie unifiée d'estimation comme suit.

Q peut être considéré comme un total sur une population fini synthétique, c.-à-d. $Q=\sum_{\alpha=1}^{N^*}t_\alpha$ où $\alpha = (ij) = 1,2,\cdots,N^*$, $t_\alpha=\phi(y_i, y_j)$ pour $\alpha = (ij)$, et $N^* = N(N - 1)/2$ est le nombre total de paires. Les données d'échantillon sur la population synthétique comprennent toutes les paires provenant de l'échantillon original: $s^* = \{(ij) : i < j, i, j \in s\}$. Sous ces conditions, les probabilités d'inclusion « de premier ordre » sont $\pi_{ij}=P(i,j\in s)$, et les "poids de sondage

de base" sont $d_{ij} = 1/\pi_{ij}$. Le terme $\mu(x_i, \theta) = E_\xi(y_i|x_i)$ devrait maintenant être remplacé par $E_\xi \{\phi (y_i, y_j)|x_i, x_j\}$.

En utilisant l'indice de paire original (ij), nous définissons l'estimateur par calage d'un modèle de Q comme étant $\hat{Q}_{MC} = \sum_{i\in s}\sum_{j>i} w_{ij}\phi(y_i, y_j)$ où les poids w_{ij} minimisent la mesure de la distance chi-deux modifiée

$$\Phi_{s^*} = \sum_{i\in s}\sum_{j>i} (w_{ij} - d_{ij})^2 / (d_{ij} q_{ij})$$

conditionnement à

$$\sum_{i\in s}\sum_{j>i} w_{ij} E_\xi\{\phi(y_i, y_j)| x_i, x_j\} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} E_\xi\{\phi(y_i, y_j)| x_i, x_j\} \quad (4.1)$$

Soit $\hat{Q}_C$ un estimateur par calage de Q quand on utilise C* = $\{u(x_i, x_j), i, j = 1, 2, \cdots\}$ dans (4·1) comme variable de calage. Soit $\boldsymbol{L}^*$ l'ensemble de toutes les séries possibles C* = $\{u(x_i, x_j), i, j = 1, 2, \cdots\}$ satisfaisant à une condition de moment de deuxième ordre fini semblable à celle utilisée pour définir $\boldsymbol{L}$. Si nous redéfinissons le plan d'échantillonnage régulier en remplaçant les d_i dans (C1) et (C2) de la section 2 par d_{ij} reformulé de façon appropriée, nous obtenons le résultat suivant.

Théorème 4. Parmi la classe d'estimateurs par calage $\hat{Q}_C$ avec C = $\{u(x_i, x_j), i, j = 1, 2, \cdots\} \in \boldsymbol{L}^*$, l'estimateur par calage de modèle $\hat{Q}_{MC}$ atteint la valeur minimale de $E_\xi\{AV_p(\hat{Q}_C)\}$ sous le modèle (2.1) et tout plan d'échantillonnage régulier.*

Preuve : Le résultat du théorème 1 ne s'applique pas directement ici, à cause d'une faible corrélation dans la série de $t_\alpha = \phi(y_i, y_j)$, $\alpha = (ij) = 1, 2, \cdots, N^*$: t_α et $t_{a'}$ ne sont pas indépendants sous le mode (2.1) si $\alpha = (ij)$ et $\alpha' = (lm)$ ont un indice en commun. Puisque le nombre total de paires $(t_a, t_{a'})$ dont la covariance est éventuellement non nulle est d'ordre $O(N^3) = O\{(N^*)^{3/2}\}$ et que le nombre total de paires dont la covariance est nulle et d'ordre $O\{(N^*)^2\}$, en utilisant une notation similaire à celle de la preuve du théorème 1, nous pouvons montrer que $E_p\{V_\xi(T_1)\} = O\{(N^*)^{-1}\}$ et $E_p\{V_\xi(T_2)\} = O\{(n^*)^{-1}(N^*)^{-1/2}\}$. Le reste de la preuve découle directement de celle du théorème 1.

Nous modifions la fonction de pseudo (log) vraisemblance empirique pour tenir compte de toutes les paires (ij) en utilisant les d_{ij}. Soit

$$l^*(p) = \sum_{i\in s}\sum_{j>i} d_{ij} \log(p_{ij}).$$

L'estimateur à modèle calé du pseudo-maximum de vraisemblance empirique de Q est défini comme étant $\hat{Q}_{ME} = N^* \sum_{i\in s}\sum_{j>i} \hat{p}_{ij}\phi(y_i, y_j)$ où les $\hat{p}_{ij}$ maximisent $l^*(p)$ conditionnement à

$$\sum_{i\in s}\sum_{j>i} p_{ij} = 1(p_{ij} > 0), \sum_{i\in s}\sum_{j>i} p_{ij} E_\xi\{\phi(y_i, y_j)| x_i, x_j\} = \frac{1}{N^*}\sum_{i=1}^{N}\sum_{j=i+1}^{N} E_\xi\{\phi(y_i, y_j)| x_i, x_j\}. \quad (4.2)$$

Nous pouvons établir de la même façon un théorème concernant l'optimalité de $\hat{Q}_{ME}$ qu'il fait pendant au théorème 2.

Comme d'habitude, nous remplaçons tout paramètre du modèle figurant dans les contraintes (4.1) ou (4.2) par une estimation sur échantillon convergente par rapport au plan de sondage. À la section 4.1, nous discutons de certains détails des estimateurs proposés pour la variance de population. À la section 4.2, nous résolvons la question de savoir quand et comment nous pouvons utiliser l'information auxiliaire pour l'estimation de la moyenne de population au moyen d'un estimateur par la régression généralisée, ainsi que pour l'estimation de la variance de cet estimateur par

calage sous cette approche unifiée.

4·1 Estimation de la variance de population

Notons que la variance de population peut être réécrite sous la forme $S^2 = \{N(N-1)\}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N}(y_i - y_j)^2$. Sous le modèle (2.1), $E_\xi\{(y_i - y_j)^2 \mid x_i, x_j\} = \{\mu(x_i,\theta) - \mu(x_j,\theta)\}^2 + \sigma^2\{v^2(x_i) + v^2(x_j)\}$, qui devrait être utilisée dans la contrainte (4.1) pour l'estimation optimale. Nous pouvons aussi remplacer (4.1) de manière suffisante par les deux équations qui suivent.

$$\sum_{i\in s}\sum_{j>i} w_{ij}\{\mu(x_i,\theta) - \mu(x_j,\theta)\}^2 = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\{\mu(x_i,\theta) - \mu(x_j,\theta)\}^2 \tag{4.3}$$

$$\sum_{i\in s}\sum_{j>i} w_{ij}\{v^2(x_i) + v^2(x_j)\} = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\{v^2(x_i) + v^2(x_j)\}. \tag{4.4}$$

Dans de nombreuses applications, $v(x_i) \equiv 1$. Dans ce cas, la deuxième équation de calage (4.4) devient $\sum_{i\in s}\sum_{j>i} w_{ij} = N^*$. L'estimateur résultant $\hat{S}^2_{MC}$ se réduit à celui propose par Sitter et Wu (2002). Sous un modèle de travail linéaire où $\mu(x_i,\theta) = x_i'\theta$, cet estimateur prend la forme élégante $\hat{S}^2_{MC} = \hat{S}^2_{HT} + \hat{\theta}'(S_x^2 - S_x^2)\hat{\theta}\,\hat{B}$, où $\hat{S}^2_{HT} = \{N(N-1)\}^{-1}\sum_{i\in s}\sum_{j>i} d_{ij}(y_i - y_j)^2$, $S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})(x_i - \bar{X})'$ $s_x^2 = \{N(N-1)\}^{-1}\sum_{i\in s}\sum_{j>i} d_{ij}(x_i - x_j)(x_i - x_j)'$, et $\hat{\mathrm{B}}$ est le coefficient de régression estimé de la régression de $v_{ij} = (y_i - y_j)^2$ sur $u_{ij} = \hat{\theta}'(x_i - x_i)(x_i - x_j)'\hat{\theta}$.

L'estimateur à modèle calé du pseudo-maximum de vraisemblance empirique est plus utile dans ce contexte. Notons que, sous le modèle (2.1), nous pouvons remplacer les contraintes (4.2) par

$$\sum_{i\in s}\sum_{j>i} p_{ij} = 1 \quad (p_{ij} > 0) \tag{4.5}$$

$$\sum_{i\in s}\sum_{j>i} p_{ij}\{\mu(x_i,\theta) - \mu(x_j,\theta)\}^2 = \frac{1}{N^*}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\{\mu(x_i,\theta) - \mu(x_j,\theta)\}^2 \tag{4.6}$$

$$\sum_{i\in s}\sum_{j>i} p_{ij}\{v^2(x_i) + v^2(x_j)\} = \frac{1}{N^*}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\{v^2(x_i) + v^2(x_j)\} \tag{4.7}$$

Quand $v(x_i) \equiv 1$, la dernière équation (4.7) se réduit à (4.5), et l'estimateur résultant $\hat{S}^2_{ME}$ est également réduit à celui proposé par Sitter et Wu (2002). Ces derniers décrivent un algorithme simple et stable pour le calcul des poids $\hat{p}_{ij}$. Puisque $\hat{p}_{ij} > 0$, l'estimateur à modèle calé du pseudo-maximum de vraisemblance empirique est toujours positif, propriété qui est désirable pour les applications pratiques.

Sous un modèle à variance non homogène, l'introduction de la contrainte (4.4) ou (4.7) augmente habituellement l'efficacité des estimateurs résultants, comme le montre l'étude en simulation présentée à la section 5.

4·2 Estimation de la variance de l'estimateur par la régression généralisée

L'estimation par la régression généralisée du total (ou de la moyenne) de population est l'une des méthodes les plus fréquemment adoptées pour utiliser l'information auxiliaire provenant d'une enquête. En supposant que les totaux *X*

sont connus, nous calculons l'estimateur par la régression généralisée de *Y* sous la forme $\hat{Y}_{GR} = \hat{Y}_{HT} + \left(X - \hat{X}_{HT}\right)'\hat{\theta}$, où $\hat{Y}_{HT} = \sum_{i\in s} d_i y_i$ et $\hat{X}_{HT} = \sum_{i\in s} d_i x_i$ sont les estimateurs classiques d' Horvitz-Thompson, et $\hat{\theta}$ est le coefficient estimés de la régression de *y* sur *x*. La variance due au plan de sondage asymptotique de cet estimateur est donnée par

$$AV_p\left(\hat{Y}_{GR}\right) = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2 ,$$

où $e_i = y_i - x_i'\theta_N$ et θ_N est le coefficient de régression en population finie qui est estimé par $\hat{\theta}$.

Nous pouvons maintenant répondre clairement à la question de savoir quand et comment nous pouvons utiliser l'information auxiliaire pour l'estimation du total de population au moyen de l'estimateur par la régression généralisée, ainsi que pour l'estimation de la variance de cet estimateur dans le cadre de la méthode du calage optimal du modèle. Notons que $AV_p\left(\hat{Y}_{GR}\right)$ a la forme de Q avec $\phi\left(y_i, y_j\right) = \left(\pi_i\pi_j - \pi_{ij}\right)\left(e_i/\pi_i - e_j/\pi_j\right)^2$. Sous le modèle (3.1), qui a motivé l'estimateur par régression généralisée, la variable optimale de calage qu'il convient d'utiliser dans (4.1) est

$$E_\xi\left\{\phi\left(y_i, y_j\right) \mid x_i, x_j\right\} \approx \left(\pi_i\pi_j - \pi_{ij}\right)\left\{\frac{v^2\left(x_i\right)}{\pi_i^2} + \frac{v^2\left(x_j\right)}{\pi_j^2}\right\}\sigma^2 .$$

Ici, nous avons utilisé le fait que $E_\xi(e_i) = 0$. Il est maintenant évident que, si la structure de variance du modèle (3.1) est homogène, c.-à-d. $v(x_i) \equiv 1$, la variable de calage sera indépendante des x_i. Nous ne pouvons pas utiliser la même information auxiliaire pour améliorer l'estimation de la variance de l'estimateur par la régression généralisée. Par contre, sous un modèle par la régression linéaire dont la variance n'est pas homogène, nous pourrons apporter certaines améliorations. La contrainte qu'il conviendrait d'utiliser pour construire l'estimateur par calage d'un modèle est donnée par

$$\sum_{i\in s}\sum_{j>i} w_{ij}\left(\pi_i\pi_j - \pi_{ij}\right)\left\{\frac{v^2\left(x_i\right)}{\pi_i^2} + \frac{v^2\left(x_j\right)}{\pi_j^2}\right\} = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left\{\frac{v^2\left(x_i\right)}{\pi_i^2} + \frac{v^2\left(x_j\right)}{\pi_j^2}\right\} .$$

Une contrainte semblable devrait être utilisée lorsqu'on estime $AV_p\left(\hat{Y}_{GR}\right)$ par la méthode de la pseudo-vraisemblance empirique fondée sur le modèle calé.

5. SIMULATION

Nous établissons l'optimalité des estimateurs proposés sous le modèle réel sur grands échantillons. À la présente section, nous décrivons une étude en simulation limitée réalisée pour étudier en échantillon fini les propriétés fondées sur le plan de sondage de ces estimateurs comparativement à celles des estimateurs par calage classiques.

Dans la simulation, nous avons généré une population finie fixe de taille $N = 2000$ à partir du modèle de régression $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\varepsilon_i$ où les x_{1i} et les x_{2i} suivent une loi gamma standard et une loi log-normale standard, respectivement. Les deux variables auxiliaires prennent des valeurs non négatives et leur distribution est étalée à droite, ce qui est assez courant dans les applications réelles. Notons que nous utilisons $v(x_i) = x_{1i}$ dans le modèle. De façon fort commode, la valeur de tous les β_i est fixée à 1. Les ε_i sont indépendantes et identiquement distribuées car $N(0,\sigma^2)$. Nous avons choisi quatre valeurs distinctes de σ^2 de sorte que les coefficients de corrélation en population finie, *p*, entre *y* et $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ soient égaux à 0.9, 0.8, 0.7 et 0.6, respectivement. Lors de chaque exécution de la simulation, nous commençons par tirer un échantillon aléatoire simple de taille n = 100 à partir de la population finie, puis nous estimons les paramètres du modèle (β_0, β_1, β_2) et σ^2 par la méthode des moindres carrés habituelle, et nous

calculons les divers estimateurs à partir des données d'échantillon. Le processus est répété B = 1000 fois.

Sous le modèle de la régression linéaire, l'estimateur par calage optimal du modèle de la moyenne de population calé sur $u_i = x_i'\hat{\beta}$ se réduit à l'estimateur par calage classique avec utilisation des contraintes (1.1) (Wu et Sitter, 2001) et il n'est donc pas inclus dans la simulation.

Pour la fonction de distribution $F_Y(t)$, nous calculons quatre estimateurs : un estimateur par calage optimal du modèle de la vraisemblance empirique $\hat{F}_{ME1}(t)$ sous le modèle de régression (3.1), l'estimateur par calage optimal du modèle de la vraisemblance empirique $\hat{F}_{ME2}(t)$ sous le modèle de régression logistique (3.2), l'estimateur du pseudo-maximum de vraisemblance empirique $\hat{F}_{CC}(t)$ sous l'équation de calage classique $\sum_{i\in s} p_i x_i = \overline{X}$, et l'estimateur de référence d'Horvitz-Thompson $\hat{F}_{HT}(t)$. Nous calculons tous les estimateurs pour cinq quantiles de population différents t_α en fixant α = 0,10, 0,30, 0,50, 0,70, et 0,90, et nous calculons les poids optimaux $\hat{p}_i$ en utilisant la valeur particulière de t_α. Nous évaluons la performance d'un estimateur $\hat{F}(t)$ au moyen du biais relatif en pourcentage (RB%) et de l'efficacité relative (RE) définis comme étant

$$RB\% = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{F}_b(t) - F_Y(t)}{F_Y(t)} \text{ et } RE + \frac{MSE\left(\hat{F}_{HT}(t)\right)}{MSE\left(\hat{F}(t)\right)},$$

où $MSE\left(\hat{F}(t)\right) = B^{-1}\sum_{b=1}^{B}\left[\hat{F}_b(t) - F_Y(t)\right]^2$ et $\hat{F}_b(t)$ est calculé à partir de la b^e exécution de la simulation. Nous utilisons l'estimateur d'Horvitz-Thompson comme valeur de référence.

Tableau 1. *Efficacité relative des estimateurs pour la fonction de distribution*

	α	0,10	0,30	0,50	0,70	0,90
$\hat{F}_{ME1}(t)$		1,15	1,79	2,17	1,99	2,18
$\hat{F}_{ME2}(t)$		1,07	1,43	1,72	1,73	1,85
$\hat{F}_{CC}(t)$		1,04	1,12	1,27	1,39	1,53
$\hat{F}_{HT}(t)$		1,00	1,00	1,00	1,00	1,00

Le tableau 1 résume les données sur l'efficacité relative simulée pour la population quand p = 0,80. Les résultats pour d'autres valeurs de p produisent des profils comparables à celui du tableau 1, où l'efficacité relative des trois estimateurs décroît lorsque p diminue. Les valeurs du biais relatif en pourcentage sont toutes dans les 2%. L'estimateur optimal $\hat{F}_{ME1}(t)$ sous le modèle de régression donne de meilleurs résultats que tous les autres pour tous les quantiles de population et le gain d'efficacité par rapport à l'estimateur par calage classique $\hat{F}_{CC}(t)$ peut être très important. L'estimateur $\hat{F}_{ME2}(t)$ sous un modèle de régression logistique donne d'assez bons résultats, mais est un peu moins efficace que $\hat{F}_{ME1}(t)$ qui est calculé sous le modèle de régression pure. L'un des avantages de l'utilisation de $\hat{F}_{ME2}(t)$ est que la fonction de variance $v(x)$ qui joue un rôle crucial sous le modèle de régression ne pose aucun problème ici et le modèle de régression logistique offre une solution de rechange intéressante pour nombre d'applications du monde réel.

Comme pour la variance de population, nous calculons l'estimateur par calage d'un modèle $\hat{S}^2_{MC1}$ et l'estimateur à modèle calé de la vraisemblance empirique $\hat{S}^2_{ME1}$ en utilisant la contrainte unique (4.3) ou (4.6); lorsque nous utilisons les deux contraintes (4.3) et (4.4), ou (4.6) et (4.7), nous représentons les estimateurs résultants par $\hat{S}^2_{MC2}$ et

$\hat{S}^2_{ME2}$. Nous définissons le biais relatif en pourcentage et l'efficacité relative de la même façon que précédemment et nous comparons les résultats à ceux obtenus pour l'estimateur d'Horvitz-Thompson de référence $\hat{S}^2_{HT}$.

Tableau 2. *Efficacité relative des estimateurs pour la variance de population*

P	$\hat{S}^2_{MC1}$	$\hat{S}^2_{ME1}$	$\hat{S}^2_{MC2}$	$\hat{S}^2_{ME2}$	$\hat{S}^2_{HT}$
0,90	6,36	5,45	7,05	4,27	1,00
0,80	2,27	2,42	3,17	2,34	1,00
0,70	1,25	1,56	2,67	1,95	1,00
0,60	0,89	1,02	2,01	1,83	1,00

L'efficacité relative simulée de chaque estimateur pour les quatre valeurs de *p* est présentée au tableau 2. Les valeurs absolues du biais relatif en pourcentage simulé sont toutes inférieures à 4% et ne sont pas présentées ici faute d'espace. Les estimateurs $\hat{S}^2_{MC1}$ et $\hat{S}^2_{ME1}$ donnent d'excellents résultats lorsque la variable de calage unique $\mu(x_i, \hat{\beta}) = x_i' \hat{\beta}$ est un prédicteur puissant de la variable dépendante (c.-à-d. valeur élevée de *p*), mais leur performance peut être médiocre lorsque ce genre de relation est faible (p. ex. p = 0,60). Les estimateurs optimaux $\hat{S}^2_{MC2}$ et $\hat{S}^2_{ME2}$ qui utilisent de l'information auxiliaire provenant à la fois de la fonction de moyenne $\mu(x_i, \hat{\beta})$ et de la fonction de variance $v(x_i)$ donnent de bons résultats dans tous les cas et leur perte d'efficacité causée par la réduction de la valeur de *p* est moins importante.

6. CONCLUSION

Un nombre important de travaux ont été publiés récemment sur l'utilisation de données d'enquêtes auxiliaires au stade de l'estimation. Selon ces travaux, un bon moyen de le faire à recourir à la minimisation sous contrainte (comme dans la méthode de calage) ou à la maximisation sous contrainte (comme dans la méthodes de la vraisemblance empirique) d'une fonction objective. Cependant, pour la plupart des méthodes existantes, le choix des variables de calage demeure ponctuel. Il en est notamment ainsi lorsqu'on doit estimer la fonction de distribution d'une population finie ou une dimension de population de deuxième ordre comme la variance de population.

La méthode de calage optimal proposée nécessite la spécification de la fonction de moyenne $\mu(x_i, \theta)$ et (ou) de la fonction de variance $v(x_i)$ à partir du modèle. Une discussion générale de la construction des modèles et des tests diagnostiques en s'appuyant sur des données d'enquête complexes dépasse le cadre du présent article et le sujet nécessite une étude plus approfondie. Dans nombre d'applications, il est probable que l'on utilise le modèle paramétrique de régression linéaire (3.1). Dans le cadre d'une variable unique *x*, Breidt et Opsomer (2000) applique une méthode de lissage non paramétrique pour trouver les espérances fondées sur le modèle de la variable dépendante. L'extension de la méthode à des variables *x* multiples paraît possible.

Il convient de souligner que les estimateurs par calage optimaux assistés par un modèle tirent parti de ce dernier, mais ne dépendent pas excessivement de son exactitude. Les estimateurs proposés donnent leurs meilleurs résultats sous un modèle qui décrit adéquatement la population étudiée et demeurent convergents par rapport au plan de sondage même si la spécification du modèle est incorrecte. Les résultats d'optimalité, qui sont établis sous le modèle supposé conceptuellement comme étant vrai, offrent une orientation pratique pour la construction des estimateurs par calage sous le modèle de travail.

Une caractéristique importante des résultats présentés dans l'article est que l'optimalité de l'estimateur par calage d'un modèle ou de l'estimateur à modèle calé du pseudo-maximum de vraisemblance empirique ne dépend pas du plan d'échantillonnage, contrairement aux résultats de Godambe et Thompson (1973), ou de Cassel et coll. (1976), où un estimateur optimal est apparié à un plan d'échantillonnage particulier. En pratique, le fait que l'optimalité d'un estimateur ne dépend pas du plan d'échantillonnage est une caractéristique intéressante lorsqu'il faut construire ce genre d'estimateur au stade de l'estimation.

Nos résultats fournissent aussi un cadre de travail unifié pour l'estimation optimale de la moyenne ou du total de population, de la fonction de distribution, de la variance de population, de la variance d'un estimateur linéaire ou de quantités de deuxième ordre d'une population finie. L'application de cette méthode de calage optimal permet d'examiner plus clairement certaines questions fondamentales que pose l'utilisation de données auxiliaires provenant d'enquêtes.

(i) L'utilisation efficace de l'information auxiliaire tirée de données d'enquête dépend à la fois des paramètres qu'il faut estimer et de la relation réelle entre la variable dépendante et les covariables. Le calage aveugle sur les variables auxiliaires n'est habituellement pas une bonne méthode.

(ii) Les contraintes de base utilisées dans (1.1) sont justifiables si la relation entre *y* et *x* est presque linéaire et que le paramètre étudié est la moyenne ou le total de population. Dans ces conditions, l'estimateur par calage (classique) résultant de $\overline{Y}$ est identique à l'estimateur par calage optimal du modèle obtenu en utilisant $\hat{\mu}_i = x_i' \hat{\theta}$ comme variable de calage (Wu et Sitter, 2001). Donc, l'étalonnage sous-entend une estimation efficace.

(iii) Si la relation entre *y* et *x* est linéaire, connaître $\overline{X}$ est une « condition suffisante » pour que l'estimation de la moyenne $\overline{Y}$ ou du total *Y* de la population soit efficace. Si la relation est non linéaire ou que les paramètres d'intérêt comportent une fonction non linéaire, il est essentiel de disposer de données auxiliaires complètes et (ou) d'un modèle plus poussé pour obtenir une estimation « optimale ».

(iv) La fonction de variance $v(x_i)$ provenant du modèle (2.1) ne joue aucun rôle dans la construction des estimateurs par calage optimaux de la moyenne ou du total de population. Cependant, il n'en est pas ainsi pour l'estimation optimale de la fonction de distribution de la population finie, de la variance de population ou d'autres quantités de population de deuxième ordre où $v(xi)$ est aussi importante que la fonction de moyenne $\mu(x_i, \theta)$.

(v) L'information auxiliaire peut parfois être utilisée triplement, pour l'élaboration du plan d'échantillonnage, pour l'estimation de la moyenne ou du total de population à l'aide d'un estimateur par la régression généralisée et pour l'estimation de la variance de cet estimateur par calage. Ce genre de situations peuvent être définies sous l'approche du calage optimal.

Dans les cas où l'estimation optimale nécessite des données auxiliaires complètes, mais que celles-ci ne sont pas disponibles, la méthode du calage optimal peut être combinée à l'échantillonnage à deux degrés où on traite les mesures des covariables obtenues sur le grand échantillon de premier degré comme des données auxiliaires « complètes ». Nous étudions à l'heure actuelle certaines questions pratiques et le gain d'efficacité lié à cette approche.

REMERCIEMENTS

La présente étude a été financée par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada. L'auteur remercie le professeur Randy R. Sitter, M. Steve Drekic, le rédacteur adjoint et deux examinateurs de leurs commentaires et suggestions constructifs qui lui ont permis d'améliorer considérablement l'article.

ANNEXE : PREUVES DES THÉORÈMES 1 ET 2

Preuve du théorème 1 : Sans perte de généralités, nous considérons la mesure de la distance du chi-deux avec les poids q_i satisfaisant à $N^{-1}\sum_{i=1}^{N} q_i^2 = O(1)$ et à $q_i \geq q$ pour une certaine constante $q > 0$. Il est facile de montrer que la minimisation de Φ conditionnellement à (1.2) mène à

$$\hat{\bar{Y}}_C = \frac{1}{N}\sum_{i\in s} d_i y_i + \frac{1}{N}\left(\sum_{i=1}^{N} u_i - \sum_{i\in s} d_i u_i\right)\hat{B},$$

où $u_i = u(x_i)$, $\hat{B} = \left(\sum_{i\in s} d_i q_i u_i y_i\right)/\left(\sum_{i\in s} d_i q_i u_i^2\right)$.

Sous un plan d'échantillonnage régulier, $AV_p\left(\hat{\bar{Y}}_C\right) = V_p(T)$, où

$$T = \frac{1}{N}\sum_{i\in s} d_i y_i + \frac{1}{N}\left(\sum_{i=1}^{N} u_i - \sum_{i\in s} d_i u_i\right)B_N ,$$

et $B_N = \left(\sum_{i=1}^{N} u_i q_i y_i\right)/\left(\sum_{i=1}^{N} q_i u_i^2\right)$.

Soit $\mu_i = \mu(x_i, \theta)$, $\bar{\mu} = E_\xi\left(\bar{Y}\right) = N^{-1}\sum_{i=1}^{N}\mu_i$, $B_\xi(T) = E_\xi(T) - \bar{\mu}$. Puisque $E_p(T) = \bar{Y}$, $V_p(T) = E_p\left(T - \bar{Y}\right)^2$, il est facile de montrer que

$$E_\xi\left\{V_p(T)\right\} = E_p\left\{V_\xi(T)\right\} + E_p\left[\left\{B_\xi(T)\right\}^2\right] - V_\xi\left(\bar{Y}\right).$$

Notons que E_ξ et V_ξ sont conditionnels pour les x_i donnés. Soit $U^2 = N^{-1}\sum_{i=1}^{N} q_i u_i^2$, $D = N^{-1}\left(\sum_{i=1}^{N} u_i - \sum_{i\in s} d_i u_i\right)$. Nous pouvons réécrire T sous la forme T_1 + T_2, où $T_1 = N^{-1}\sum_{i\in s} d_i y_i$, $T_2 = DU^{-2}N^{-1}\sum_{i=1}^{N} q_i u_i y_i$. Nous avons

$$E_p\left\{V_\xi(T)\right\} = E_p\left\{V_\xi(T_1)\right\} + E_p\left\{V_\xi(T_2)\right\} + 2E_p\left\{COV_\xi(T_1, T_2)\right\},$$

où $COV_\xi(T_1, T_2)$ représente la covariance sous le modèle. On peut voir que

$$E_p\left\{V_\xi(T_1)\right\} = \frac{1}{N^2}\sum_{i=1}^{N} d_i v^2(x_i)\sigma^2 = O\left(\frac{1}{N}\right),$$

$$E_p\left\{V_\xi(T_2)\right\} = \left\{E_p\left(D^2\right)\right\}U^{-4}\frac{1}{N^2}\sum_{i=1}^{N} q_i^2 u_i^2 v^2(x_i)\sigma^2 = O\left(\frac{1}{nN}\right).$$

Ici, nous avons utilisé le fait que $\max_{i\in s} nd_i / N = O(1)$, $N^{-1}\sum_{i=1}^{N} v^2(x_i) = O(1)$,
$E_p\left(D^2\right) = V_p\left(N^{-1}\sum_{i\in s} d_i u_i\right) = O\left(n^{-1}\right)$, $N^{-1}\sum_{i=1}^{N} q_i^2 = O(1)$, $q_i \geq q > 0$ pour tout *i*, et $N^{-1}\sum_{i=1}^{N} q_i u_i^2 \to c^* \neq 0$.

Il découle aussi de $|COV_\xi(T_1, T_2)| \leq \left\{V_\xi(T_1)\right\}^{1/2}\left\{V_\xi(T_2)\right\}^{1/2}$ que

$$\left\{E_p\left|COV_\xi(T_1, T_2)\right|\right\}^2 \leq E_p\left\{V_\xi(T_1)\right\}E_p\left\{V_\xi(T_2)\right\},$$

qui implique $E_p\left\{COV_\xi(T_1, T_2)\right\} = O\left(n^{-3/2}\right)$. Quand *n* est grand, le terme principal dans $E_p\left\{V_\xi(T)\right\}$ est $E_p\left\{V_\xi(T_1)\right\}$, qui est indépendant du choix de la série, C. Le terme $V_\xi\left(\bar{Y}\right)$ est également indépendant de *C*.

Pour le terme $E_p\left[\left\{B_\xi(T)\right\}^2\right]$, notons que

$$B_\xi(T) = \frac{1}{N}\sum_{i\in s} d_i(\mu_i - u_i B) - \frac{1}{N}\sum_{i=1}^{N}(\mu_i - u_i B),$$

où $B = \sum_{i=1}^{N} q_i u_i \mu_i \,/\, \sum_{i=1i}^{N} q_i u_i^2$. Il s'ensuit que $E_p\{B_\xi(T)\} = 0$ et

$$E_p\left[\{B_\xi(T)\}^2\right] = V_p\{B_\xi(T)\} = V_p\left\{N^{-1}\sum_{i\in s} d_i(\mu_i - u_i B)\right\} = O\left(n^{-1}\right).$$

Minimiser $E_\xi\left\{AV_p\left(\hat{\bar{Y}}_C\right)\right\}$ revient minimiser $E_p\left[\{B_\xi(T)\}^2\right]$. Le choix de $C = (\mu_1, \mu_2, ...)$ aboutit à *B = 1* et $E_p\left[\{B_\xi(T)\}^2\right] = 0$.

Preuve du théorème 2 : D'après le théorème 1 de Chen et Sitter ((1999), nous avons

$$\hat{\bar{Y}}_{ME} = \left(\sum_{i\in s} d_i\right)^{-1}\sum_{i\in s} d_i y_i + \left\{\frac{1}{N}\sum_{i=1}^{N} u(x_i) - \left(\sum_{i\in s} d_i\right)^{-1}\sum_{i\in s} d_i u(x_i)\right\}\hat{B} + o_p\left(n^{-1/2}\right),$$

où $\hat{B}$ est défini de la même façon que dans le théorème 1 avec $q_i = 1$.

Le terme $T_1^* = \left(\sum_{i\in s} d_i\right)^{-1}\sum_{i\in s} d_i y_i$ est un estimateur de type quotient et sa variance fondée sur le plan de sondage $V_p\left(T_1^*\right)$ n'est pas la même que $V_p(T_1)$, où $T_1 = N^{-1}\sum_{i\in s} d_i y_i$. Cependant, puisque $\sum_{i\in s} d_i$ est une constante sous le modèle de superpopulation, la conclusion au sujet de $E_p\{V_\xi(T_1)\}$ dans le théorème 1 peut aussi être réénoncée ici en ce qui concerne T_1^* . Le reste de la preuve étant semblable à la preuve du théorème 1, nous l'omettons ici.

RÉFÉRENCES

Breidt, F.J. et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics.* 28, 1026-1053.

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

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 method to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Chen, J. et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica***,** 12, 1223-1239.

Deville, J.C. et Särndal, C.E. (1992). Calibration estimators in survey sampling. J. *Am. Statist. Assoc.* 87, 376-82.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. J. *Roy. Statist.* Soc.. Ser. B, 17, 267-278.

Godambe, V.P. et Thompson, M.E. (1973). Estimation in sampling theory with exchangeable prior distributions. *The Annals of Statistics.* 1, 1212-1221.

Isaki, C.T. et Fuller, W.A. (1982). Survey design under the regression superpopulation model. J. *Am. Statist. Assoc.* 77, 89-96.

Sitter, R.R. et Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. J. *Am. Statist. Assoc.*, 97, 535-543.

Wu, C. et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. J. *Am. Statist. Assoc.* 96, 185-93.