Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Estimation linéaire optimale dans un échantillonnage à deux phases

par Takis Merkouris

Date de diffusion : le 15 décembre 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 481-505
Statistique Canada, nº 12-001-X au catalogue

Estimation linéaire optimale dans un échantillonnage à deux phases

Takis Merkouris[1]

Résumé

L'échantillonnage à deux phases est un plan de sondage rentable couramment utilisé dans les enquêtes. Le présent article propose une méthode optimale d'estimation linéaire des totaux dans un échantillonnage à deux phases, qui exploite au mieux l'information auxiliaire de l'enquête. Tout d'abord, on calcule formellement un meilleur estimateur linéaire sans biais (MELSB) de tout total sous une forme analytique, et on démontre qu'il s'agit d'un estimateur par calage. Ensuite, la reformulation appropriée du MELSB et l'estimation de ses coefficients inconnus permettent de construire un estimateur par la régression « optimal », qui peut également être obtenu au moyen d'une procédure de calage adéquate. Ce calage présente une caractéristique distinctive : l'alignement des estimations des deux phases dans une procédure en une étape comprenant les échantillons combinés de la première et de la deuxième phase. L'estimation optimale est faisable pour certains plans à deux phases souvent employés dans les enquêtes à grande échelle. Pour les plans généraux à deux phases, une autre procédure de calage donne un estimateur par la régression généralisée comme estimateur optimal approximatif. L'approche générale proposée d'estimation optimale permet d'utiliser le plus efficacement possible l'information auxiliaire disponible dans toute enquête à deux phases. Les avantages de cette méthode par rapport aux méthodes existantes d'estimation dans un échantillonnage à deux phases sont démontrés théoriquement et au moyen d'une étude par simulations.

Mots-clés : Information auxiliaire; meilleure estimation linéaire sans biais; calage; estimation par la régression généralisée; échantillonnage double.

1. Introduction

Le plan d'échantillonnage à deux phases, aussi appelé échantillonnage double, a toujours été utilisé dans les enquêtes par sondage en tant que méthode d'enquête rentable. Pendant la première phase, un échantillon relativement important de la population cible est sélectionné afin de fournir des renseignements auxiliaires peu coûteux à obtenir. Cet échantillon constitue une base très informative à partir de laquelle un sous-échantillon est sélectionné à la deuxième phase pour recueillir des renseignements sur les éléments d'intérêt. De plus, l'échantillonnage à deux phases est de plus en plus utilisé comme mécanisme de traitement de la non-réponse. Särndal, Swensson et Wretman (1992) présentent un compte rendu exhaustif de ce type d'emploi de l'échantillonnage à deux phases. Groves et Heeringa (2006) ainsi que Brick et Tourangeau (2017) traitent du rôle important de l'échantillonnage à deux phases dans les plans de sondage réactifs quand des mesures coûteuses sont prises pour réduire le biais de non-réponse. D'autres applications de l'échantillonnage à deux phases – apparues récemment dans la pratique des enquêtes – comportent diverses formes d'intégration d'enquêtes distinctes. Dans une de ces formes, un échantillon de première phase sert de base de sondage à l'échantillon de deuxième phase pour une multitude d'enquêtes semblables (Turmelle et Beaucage, 2013). Dans un autre cas, une enquête primaire de grande envergure est utilisée comme base de sondage pour une autre enquête de plus petite envergure comportant un plus grand nombre d'éléments (Australian Bureau of Statistics, 2004).

1. Takis Merkouris, Department of Statistics, Athens University of Economics and Business, 2, rue Trias, Athènes 11362, Grèce. Courriel : merkouris@aueb.gr.

Dans un échantillonnage à deux phases, l'information auxiliaire peut se trouver à différents niveaux. Certains renseignements sont au niveau de l'ensemble de la population, alors que d'autres sont au niveau de l'échantillon de la première ou de la deuxième phase. De nombreuses études portent sur l'utilisation de ces renseignements en vue de l'amélioration de l'estimation des totaux ou des moyennes de la population; voir Särndal et coll. (1992), Hidiroglou et Särndal (1998), Hidiroglou (2001), Estevao et Särndal (2002, 2009), Wu et Luan (2003), Chen et Kim (2014), et les références qui s'y trouvent. En général, la littérature présente deux méthodes d'intégration de l'information auxiliaire au processus d'estimation : l'approche par régression généralisée et l'approche par calage; les deux phases de l'échantillonnage supposent deux modèles de régression ou deux calages successifs. Dans certaines conditions, les deux approches donnent des estimateurs identiques, mais en général, ce n'est pas le cas. L'estimation de la variance de ces estimateurs à deux phases a fait l'objet de nombreuses études; voir par exemple Sitter (1997), Fuller (1998), Kim et Sitter (2003), Kim, Navarro et Fuller (2006), Hidiroglou, Rao et Haziza (2008), Kim et Yu (2011), Beaumont, Beliveau et Haziza (2015).

Que ce soit une formulation par la régression ou par calage des procédures d'estimation existantes, les estimateurs obtenus pour une variable cible sont en fait des combinaisons linéaires d'estimateurs de Horvitz-Thompson de divers totaux (ou moyennes), y compris l'estimateur de la variable cible calculé à partir de l'échantillon de la deuxième phase et des estimateurs de variables auxiliaires calculés à partir à la fois des échantillons de première et de deuxième phase. Dans le présent article, en adoptant une approche formelle de l'estimation optimale, nous considérons la combinaison linéaire la plus efficace des estimateurs disponibles à partir des deux phases, selon le principe de la meilleure estimation linéaire sans biais. Nous montrons que le meilleur estimateur linéaire sans biais (MELSB) calculé sous une forme analytique possède une propriété d'orthogonalité utile et qu'il peut être construit autrement comme un estimateur par calage, qui est linéaire dans les valeurs de la variable associée et intègre l'information auxiliaire dans les poids de sondage calés. L'estimation des coefficients inconnus de ce MELSB, au moyen de toute l'information auxiliaire disponible des deux phases d'échantillonnage, donne un estimateur « optimal », analogue à l'estimateur par la régression optimale en une phase de Montanari (1987) et Rao (1994). Cet estimateur est une approximation à grand échantillon du MELSB, par des coefficients estimés minimisant sa variance approximative estimée (grand échantillon) et préservant la propriété d'orthogonalité du MELSB. Au moyen d'une reformulation adéquate du MELSB, l'estimateur optimal peut également être obtenu par une procédure de calage appropriée. La caractéristique distinctive de ce calage est la procédure pratique en une étape consistant à aligner les estimations des deux phases au moyen des échantillons combinés de la première et de la deuxième phase. L'estimation optimale est faisable pour certains plans à deux phases souvent employés dans les enquêtes à grande échelle. Pour les plans généraux, une autre procédure de calage à une étape donne un nouvel estimateur par la régression généralisée comme approximation pratique de l'estimateur optimal.

La méthode générale d'estimation proposée guide la construction d'estimateurs par calage dans tout cas particulier d'enquête à deux phases, en s'appuyant le plus efficacement possible sur l'information

auxiliaire disponible. Elle permet aussi de donner un aperçu des méthodes d'estimation moins efficaces quand elles sont placées dans le cadre d'une estimation optimale. Les avantages de la méthode proposée par rapport aux méthodes existantes sont démontrés théoriquement et au moyen d'une étude par simulations.

Le présent article est organisé comme suit. La structure du plan d'échantillonnage à deux phases et la notation sont présentées dans la section 2. Le calcul du MELSB pour l'information auxiliaire de type standard dans l'échantillonnage à deux phases et son autre construction comme estimateur par calage, sont décrits à la section 3. L'estimateur optimal à deux phases et son équivalent par calage sont présentés à la section 4. L'approximation de l'estimateur optimal par un estimateur par la régression généralisée est traitée dans la section 5. Des comparaisons avec les méthodes existantes sont données à la section 6. Une étude par simulations est présentée dans la section 7. L'article se conclut par une discussion à la section 8.

2. Plan d'échantillonnage à deux phases : structure et notation

Soit $U=\{1,\ldots,k,\ldots,N\}$ une population finie de N unités. Un échantillon de la première phase s_1 de taille n_1 est tiré de la population U, au moyen d'un plan de sondage qui définit la probabilité d'inclusion $\pi_{1k}=P(k\in s_1)$ pour l'unité $k\in U$ et la probabilité d'inclusion conjointe $\pi_{1kl}=P(k,l\in s_1)$ pour les unités $k,l\in U$. Ensuite, un échantillon de la deuxième phase s_2 de taille n_2 est tiré de s_1 au moyen d'un plan de sondage qui définit la probabilité d'inclusion conditionnelle $\pi_{2k}=P(k\in s_2\,|\,s_1)$ pour $k\in s_1$, et la probabilité d'inclusion conditionnelle conjointe $\pi_{2kl}=P(k,l\in s_2\,|\,s_1)$ pour les unités $k,l\in s_1$. En supposant que $\pi_{1k}>0$ pour tous les $k\in U$ et $\pi_{2k}>0$ pour tous les $k\in s_1$, le poids de sondage de la première phase pour $k\in s_1$ est $w_{1k}=1/\pi_{1k}$, le poids de sondage conditionnel de la deuxième phase pour $k\in s_2$ est $w_{2k}=1/\pi_{2k}$, et le poids de sondage global pour $k\in s_2$ est $w_k=w_{1k}w_{2k}$.

Le type standard de variables auxiliaires dans un échantillonnage à deux phases (voir, par exemple, Särndal et coll. (1992)) comporte un vecteur de variables auxiliaires $\mathbf{x}$, partitionné comme étant $\mathbf{x}=(\mathbf{x}_1',\mathbf{x}_2')'$ par p et q de ses composantes, avec un total de population $\mathbf{t}_\mathbf{x}=\sum_U\mathbf{x}_k$ et un total connu $\mathbf{t}_{\mathbf{x}_1}=\sum_U\mathbf{x}_{1k}$ de $\mathbf{x}_1$. La valeur $\mathbf{x}_k$ est observée pour chaque unité $k\in s_1$, alors que pour un vecteur à d dimensions de variables cibles $\mathbf{y}$, avec un total $\mathbf{t}_\mathbf{y}=\sum_U\mathbf{y}_k$, la valeur $\mathbf{y}_k$ est observée seulement pour les unités $k\in s_2$. Dans certaines enquêtes, les composantes du vecteur $\mathbf{x}_2$ sont aussi des variables cibles. Un estimateur sans biais du total $\mathbf{t}_\mathbf{y}$, l'estimateur commun de Horvitz-Thompson (HT), donné par $\tilde{\mathbf{t}}_\mathbf{y}=\sum_{s_2}w_k\mathbf{y}_k$, est obtenu au moyen de l'échantillon de la deuxième phase s_2, tandis que deux estimateurs de HT du total $\mathbf{t}_\mathbf{x}$, donnés par $\hat{\mathbf{t}}_\mathbf{x}=\sum_{s_1}w_{1k}\mathbf{x}_k$ et $\tilde{\mathbf{t}}_\mathbf{x}=\sum_{s_2}w_k\mathbf{x}_k$, sont obtenus respectivement au moyen des échantillons s_1 et s_2. Dans le calcul des résultats comprenant ces estimateurs, nous utilisons la notation de vecteur $\tilde{\mathbf{t}}_\mathbf{y}=\mathbf{Y}_2'\mathbf{w}$, $\hat{\mathbf{t}}_\mathbf{x}=\mathbf{X}_1'\mathbf{w}_1$, $\tilde{\mathbf{t}}_\mathbf{x}=\mathbf{X}_2'\mathbf{w}$, $\hat{\mathbf{t}}_{\mathbf{x}_1}=\mathbf{X}_{11}'\mathbf{w}_1$, où $\mathbf{w}_1$ et $\mathbf{w}$ désignent les vecteurs des poids de sondage pour les échantillons s_1 et s_2, respectivement, $\mathbf{X}_1$ et $\mathbf{X}_{11}$ désignent les matrices de l'échantillon s_1 de $\mathbf{x}$ et $\mathbf{x}_1$ de dimensions $n_1\times(p+q)$ et $n_1\times p$, respectivement, et $\mathbf{Y}_2$, $\mathbf{X}_2$ désignent les matrices de l'échantillon s_2 de $\mathbf{y}$ et $\mathbf{x}$ de dimensions $n_2\times d$ et $n_2\times(p+q)$.

La principale cible de l'estimation est le total $\mathbf{t}_\mathbf{y}$. Cependant, pour mieux comprendre la construction des estimateurs proposés, et puisque les composantes du vecteur $\mathbf{x}_2$ peuvent également être des variables cibles, nous adopterons une approche unifiée pour estimer à la fois $\mathbf{t}_\mathbf{y}$ et $\mathbf{t}_\mathbf{x}$.

3. Meilleure estimation linéaire sans biais dans un échantillonnage à deux phases

3.1 Forme analytique du meilleur estimateur linéaire sans biais

Pour estimer plus efficacement les totaux $\mathbf{t}_\mathbf{y}$ et $\mathbf{t}_\mathbf{x}$, en incorporant toute l'information disponible des deux phases par la corrélation de $\mathbf{y}$ et $\mathbf{x}$, nous examinons les meilleurs estimateurs linéaires sans biais (MELSB), désignés par $\hat{\mathbf{t}}_\mathbf{y}^B$ et $\hat{\mathbf{t}}_\mathbf{x}^B$, qui sont des combinaisons linéaires sans biais à variance minimale des quatre estimateurs $\tilde{\mathbf{t}}_\mathbf{y}$, $\hat{\mathbf{t}}_\mathbf{x}$, $\tilde{\mathbf{t}}_\mathbf{x}$, $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ et qui sont donnés sous forme matricielle par :

$$\left(\hat{\mathbf{t}}_\mathbf{y}^{B'}, \hat{\mathbf{t}}_\mathbf{x}^{B'}\right)' = \boldsymbol{\mathcal{P}}\left(\tilde{\mathbf{t}}_\mathbf{y}', \hat{\mathbf{t}}_\mathbf{x}', \tilde{\mathbf{t}}_\mathbf{x}', \mathbf{t}_{\mathbf{x}_1}' - \hat{\mathbf{t}}_{\mathbf{x}_1}'\right)', \tag{3.1}$$

où $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}$, la matrice $\mathbf{W}$ comporte des entrées 1 et 0 et satisfait à $E\left[(\tilde{\mathbf{t}}_\mathbf{y}', \hat{\mathbf{t}}_\mathbf{x}', \tilde{\mathbf{t}}_\mathbf{x}', \mathbf{t}_{\mathbf{x}_1}' - \hat{\mathbf{t}}_{\mathbf{x}_1}')'\right] = \mathbf{W}(\mathbf{t}_\mathbf{y}', \mathbf{t}_\mathbf{x}')'$, et où $\mathbf{V}$ est la matrice de covariance de $(\tilde{\mathbf{t}}_\mathbf{y}', \hat{\mathbf{t}}_\mathbf{x}', \tilde{\mathbf{t}}_\mathbf{x}', \mathbf{t}_{\mathbf{x}_1}' - \hat{\mathbf{t}}_{\mathbf{x}_1}')'$. Il s'ensuit que $\mathrm{Var}\left[(\hat{\mathbf{t}}_\mathbf{y}^{B'}, \hat{\mathbf{t}}_\mathbf{x}^{B'})'\right] = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}$. Cette formulation typique de la meilleure estimation linéaire sans biais a été étudiée dans deux autres domaines de l'échantillonnage d'enquête, comme dans Wolter (1979), Jones (1980), Fuller (1990) et Chipperfield et Steel (2009). Dans le contexte actuel, une formulation plus pratique, qui conduit également à la représentation du MELSB comme estimateur par calage, se présente comme suit.

En formulant les deux combinaisons linéaires dans (3.1) sous une forme développée et en utilisant la condition d'absence de biais $E(\hat{\mathbf{t}}_\mathbf{y}^B) = \mathbf{t}_\mathbf{y}$ et $E(\hat{\mathbf{t}}_\mathbf{x}^B) = \mathbf{t}_\mathbf{x}$, on montre facilement que la matrice $\boldsymbol{\mathcal{P}}$ des coefficients dans ces combinaisons linéaires satisfait à :

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{B}_{1\mathbf{y}} & \mathbf{B}_{2\mathbf{y}} & \mathbf{B}_{3\mathbf{y}} & \mathbf{B}_{4\mathbf{y}} \\ \mathbf{B}_{1\mathbf{x}} & \mathbf{B}_{2\mathbf{x}} & \mathbf{B}_{3\mathbf{x}} & \mathbf{B}_{4\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{B}_{2\mathbf{y}} & -\mathbf{B}_{2\mathbf{y}} & \mathbf{B}_{4\mathbf{y}} \\ \mathbf{0} & \mathbf{B}_{2\mathbf{x}} & \mathbf{I} - \mathbf{B}_{2\mathbf{x}} & \mathbf{B}_{4\mathbf{x}} \end{pmatrix},$$

et qu'ensuite les deux composants du MELSB dans (3.1) s'expriment sous la forme d'une régression :

$$\begin{aligned} \hat{\mathbf{t}}_\mathbf{y}^B &= \tilde{\mathbf{t}}_\mathbf{y} + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{t}}_\mathbf{x} - \tilde{\mathbf{t}}_\mathbf{x}) + \mathbf{B}_{4\mathbf{y}}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_\mathbf{x}^B &= \tilde{\mathbf{t}}_\mathbf{x} + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{t}}_\mathbf{x} - \tilde{\mathbf{t}}_\mathbf{x}) + \mathbf{B}_{4\mathbf{x}}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.2}$$

Nous pouvons maintenant formuler (3.1) comme suit :

$$\begin{pmatrix} \hat{\mathbf{t}}_\mathbf{y}^B \\ \hat{\mathbf{t}}_\mathbf{x}^B \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{t}}_\mathbf{y} \\ \tilde{\mathbf{t}}_\mathbf{x} \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{t}}_\mathbf{x} - \tilde{\mathbf{t}}_\mathbf{x} \\ \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1} \end{pmatrix}, \tag{3.3}$$

où la matrice $\boldsymbol{\mathcal{B}}$ se compose des deuxième et quatrième colonnes de $\boldsymbol{\mathcal{P}}$, et a la valeur de minimisation de variance qui se calcule facilement :

$$\mathcal{B} = -\text{Cov}\left[\begin{pmatrix}\tilde{\mathbf{t}}_{\mathbf{y}}\\ \tilde{\mathbf{t}}_{\mathbf{x}}\end{pmatrix}, \begin{pmatrix}\hat{\mathbf{t}}_{\mathbf{x}}-\tilde{\mathbf{t}}_{\mathbf{x}}\\ \mathbf{t}_{\mathbf{x}_1}-\hat{\mathbf{t}}_{\mathbf{x}_1}\end{pmatrix}\right]\left[\text{Var}\begin{pmatrix}\hat{\mathbf{t}}_{\mathbf{x}}-\tilde{\mathbf{t}}_{\mathbf{x}}\\ \mathbf{t}_{\mathbf{x}_1}-\hat{\mathbf{t}}_{\mathbf{x}_1}\end{pmatrix}\right]^{-1}. \tag{3.4}$$

Nous posons ensuite :

$$\mathbf{w}^* = \begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}\end{pmatrix}, \quad \boldsymbol{\mathcal{X}} = \begin{pmatrix}-\mathbf{X}_1 & \mathbf{X}_{11}\\ \mathbf{X}_2 & \mathbf{0}\end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{Y}_2 & \mathbf{X}_2\end{pmatrix}, \tag{3.5}$$

de sorte que :

$$\boldsymbol{\mathcal{X}}'\mathbf{w}^* = \begin{pmatrix}\tilde{\mathbf{t}}_{\mathbf{x}}-\hat{\mathbf{t}}_{\mathbf{x}}\\ \hat{\mathbf{t}}_{\mathbf{x}_1}\end{pmatrix}, \qquad \boldsymbol{\Psi}'\mathbf{w}^* = \begin{pmatrix}\tilde{\mathbf{t}}_{\mathbf{y}}\\ \tilde{\mathbf{t}}_{\mathbf{x}}\end{pmatrix}, \tag{3.6}$$

et $\mathcal{B}$ peut ensuite être exprimé comme étant $\mathcal{B} = \text{Cov}(\boldsymbol{\Psi}'\mathbf{w}^*, \boldsymbol{\mathcal{X}}'\mathbf{w}^*)\left[\text{Var}(\boldsymbol{\mathcal{X}}'\mathbf{w}^*)\right]^{-1}$. Pour le calcul des variances et des covariances, nous définissons $\mathbf{w}^*$ au niveau de la population comme étant $\mathbf{w}_U^* = (\mathbf{w}_{1U}', \mathbf{w}_U')'$, où le k^{e} élément de $\mathbf{w}_{1U}$ est $w_{1U_k} = (1/\pi_{1k})\, I_{1k}$, la variable indicatrice I_1 désignant l'inclusion d'une unité de population dans s_1, et le k^{e} élément de $\mathbf{w}_U$ est $w_{U_k} = \left[1/(\pi_{1k}\pi_{2k})\right] I_{1k} I_{2k}$, la variable indicatrice I_2 désignant l'inclusion d'une unité de population dans s_2 conditionnellement à la sélection de l'échantillon s_1. Nous pouvons maintenant écrire $\boldsymbol{\mathcal{X}}'\mathbf{w}^* = \boldsymbol{\mathcal{X}}_U'\mathbf{w}_U^*$ et $\boldsymbol{\Psi}'\mathbf{w}^* = \boldsymbol{\Psi}_U'\mathbf{w}_U^*$, où $\boldsymbol{\mathcal{X}}_U$ et $\boldsymbol{\Psi}_U$ sont les contreparties de population de $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{\Psi}$, respectivement; toutes les sous-matrices dans $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{\Psi}$ sont élargies au niveau de la population et ont N lignes. Ensuite, en indiquant $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}'\mathbf{w}^*$ et $\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}}'\mathbf{w}^*$, nous obtenons :

$$\mathcal{B} = \text{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\left[\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1} = \boldsymbol{\Psi}_U'\,\text{Var}(\mathbf{w}_U^*)\,\boldsymbol{\mathcal{X}}_U\left[\boldsymbol{\mathcal{X}}_U'\,\text{Var}(\mathbf{w}_U^*)\,\boldsymbol{\mathcal{X}}_U\right]^{-1}. \tag{3.7}$$

Nous obtenons par la suite une expression plus analytique utile de $\mathcal{B}$ en utilisant le lemme suivant. La démonstration se trouve en annexe.

Lemme 1

$$\text{Var}(\mathbf{w}_U^*) = \begin{pmatrix}\text{Var}(\mathbf{w}_{1U}) & \text{Var}(\mathbf{w}_{1U})\\ \text{Var}(\mathbf{w}_{1U}) & \text{Var}(\mathbf{w}_U)\end{pmatrix}, \tag{3.8}$$

où $\text{Var}(\mathbf{w}_{1U}) = \left\{(\pi_{1kl} - \pi_{1k}\pi_{1l})/\pi_{1k}\pi_{1l}\right\}$, $\text{Var}(\mathbf{w}_U) = \left\{(\pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l})/\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\right\}$.

Au moyen de (3.7) et (3.8), il est facile de démontrer que (3.4) s'exprime comme suit :

$$\mathcal{B} = \begin{bmatrix}-\text{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1} & \text{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}\\ \mathbf{I} & \text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}\end{bmatrix}. \tag{3.9}$$

Dans cette représentation de $\mathcal{B}$, la propriété $\text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}}) = \text{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ est implicite, découlant de (3.8), ce qui implique que $\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \text{Var}(\tilde{\mathbf{t}}_{\mathbf{x}}) - \text{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$, et la propriété $\text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \text{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})$, ce qui

implique $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$ (cette covariance étant le bloc hors diagonale de $\boldsymbol{\mathcal{X}}_U' \mathrm{Var}(\mathbf{w}_U^*) \boldsymbol{\mathcal{X}}_U$). Ensuite, (3.2) peut s'exprimer explicitement comme suit :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^B &= \tilde{\mathbf{t}}_{\mathbf{y}} - \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^B &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.10}$$

Compte tenu de la propriété $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, il s'ensuit immédiatement que :

$$\begin{aligned} \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{y}}^B) &= \mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{y}}) - \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \right]^{-1} \mathrm{Cov}'(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad - \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} \mathrm{Cov}'(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}}^B) &= \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}}) - \mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} \mathrm{Cov}'(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.11}$$

Remarque 3.1. Chaque composante ou combinaison linéaire de composantes de $\hat{\mathbf{t}}_{\mathbf{y}}^B$ est le MELSB pour le total correspondant. De plus, comme le montre (3.11), l'efficacité de $\hat{\mathbf{t}}_{\mathbf{y}}^B$, par rapport à $\tilde{\mathbf{t}}_{\mathbf{y}}$, dépend de la force de la corrélation de $\mathbf{y}$ avec $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ ainsi que de la différence de taille de l'échantillon (et peut-être du plan de sondage) pour les échantillons s_1 et s_2.

Remarque 3.2. En raison de la propriété d'orthogonalité $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, le coefficient de l'un ou l'autre des termes $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ dans (3.10) ne changerait pas si l'autre était égal à $\mathbf{0}$ dans (3.2). Par exemple, le MELSB pour $\mathbf{t}_{\mathbf{y}}$ basé sur $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}})$ serait $\hat{\mathbf{t}}_{\mathbf{y}}^B$ comme dans (3.10), mais sans le dernier terme. Ceci se résout facilement comme un cas particulier de la configuration complète $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$. Cette propriété d'orthogonalité explique la réduction additive de la variance observée dans la première équation de (3.11).

Remarque 3.3. Le MELSB $\hat{\mathbf{t}}_{\mathbf{x}}^B$ dans (3.10) peut également être produit au moyen de la configuration réduite $(\hat{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ dans (3.1). Le même meilleur estimateur linéaire, pour une seule variable cible, a été calculé différemment dans un contexte d'échantillonnage général à une phase par Fuller et Isaki (1981) et Montanari (1987). En particulier, pour la variable auxiliaire $\mathbf{x}_1$, nous avons $\hat{\mathbf{t}}_{\mathbf{x}_1}^B = \mathbf{t}_{\mathbf{x}_1}$. Ensuite, il est facile de vérifier que le MELSB dans (3.1) peut par contre être calculé en deux étapes de meilleure estimation linéaire sans biais au moyen de la configuration $(\tilde{\mathbf{t}}_{\mathbf{y}}^B, \hat{\mathbf{t}}_{\mathbf{x}}^B, \tilde{\mathbf{t}}_{\mathbf{x}}^B)$, où $\tilde{\mathbf{t}}_{\mathbf{y}}^B = \tilde{\mathbf{t}}_{\mathbf{y}} + \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ et $\tilde{\mathbf{t}}_{\mathbf{x}}^B = \tilde{\mathbf{t}}_{\mathbf{x}} + \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ sont les MELSB produits respectivement par les configurations à une phase $(\tilde{\mathbf{t}}_{\mathbf{y}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ et $(\tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$. De fastidieux calculs algébriques permettraient de démontrer que $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_2}, \tilde{\mathbf{t}}_{\mathbf{x}_2}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}, \mathbf{t}_{\mathbf{x}_1} - \tilde{\mathbf{t}}_{\mathbf{x}_1})$ est une autre configuration du MELSB, qui est plus explicite et équivaut à celle de (3.1). Cela prouve que la configuration compacte de (3.1) fournit l'estimation linéaire la plus efficace de $\mathbf{t}_{\mathbf{y}}$ et $\mathbf{t}_{\mathbf{x}}$ au moyen de toutes les estimations pertinentes disponibles.

3.2 Le meilleur estimateur linéaire sans biais à deux phases comme estimateur par calage

En utilisant la notation conduisant à (3.7), et la configuration $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = (\hat{\mathbf{t}}_{\mathbf{y}}^{B'}, \hat{\mathbf{t}}_{\mathbf{x}}^{B'})'$ et $\boldsymbol{\Delta} = \mathrm{Var}(\mathbf{w}_U^*)$, nous pouvons exprimer le MELSB dans (3.3) comme étant $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \boldsymbol{\mathcal{B}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, où $\boldsymbol{\mathcal{B}} = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}$ et $\mathbf{t}_{\boldsymbol{\mathcal{X}}} = (\mathbf{0}', \mathbf{t}_{\mathbf{x}_1}')'$, ou sous une forme plus évocatrice :

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} \;=\; \boldsymbol{\Psi}_U' \Big[\mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \big(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \big) \Big]. \tag{3.12}$$

D'après (3.12), il semble que $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B}$ a la forme d'un estimateur par calage, qui comporte un vecteur de population de poids calés $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \big(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \big)$ et un vecteur de totaux de calage $\mathbf{t}_{\boldsymbol{\mathcal{X}}}$. Cette constatation est mise en forme dans le théorème suivant et démontrée en annexe.

Théorème 1. *Le vecteur* $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \big(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \big)$ *minimise la distance des moindres carrés généralisés* $(\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*)$ *soumise aux contraintes* $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, *c'est-à-dire* $\mathbf{X}_U' \mathbf{c}_{1U} = \mathbf{X}_U' \mathbf{c}_U$ *et* $\mathbf{X}_{1U}' \mathbf{c}_{1U} = \mathbf{t}_{\mathbf{x}_1}$, *où* $(\mathbf{c}_{1U}, \mathbf{c}_U)$ *correspond à* $(\mathbf{w}_{1U}, \mathbf{w}_U)$.

Le théorème 1 montre que la meilleure estimation linéaire sans biais au moyen de la configuration $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ est essentiellement une procédure de calage par laquelle les deux estimations $\hat{\mathbf{t}}_{\mathbf{x}}$ et $\tilde{\mathbf{t}}_{\mathbf{x}}$ de $\mathbf{t}_{\mathbf{x}}$ sont calées l'une sur l'autre, c'est-à-dire qu'elles sont alignées, et que l'estimation est calée sur le total $\hat{\mathbf{t}}_{\mathbf{x}_1}$. Nous pouvons maintenant exprimer formellement le MELSB $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B}$ comme un estimateur par calage $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = \boldsymbol{\Psi}_U' \mathbf{c}_U^*$, ses deux composantes étant données dans les formes linéaires simples $\hat{\mathbf{t}}_{\mathbf{y}}^{B} = \mathbf{Y}_U' \mathbf{c}_U$ et $\hat{\mathbf{t}}_{\mathbf{x}}^{B} = \mathbf{X}_U' \mathbf{c}_U$.

L'autre solution de construction en deux étapes du MELSB mentionnée dans la remarque 3.3 ci-dessus peut également être réalisée au moyen d'une procédure de calage en deux étapes comprenant $\mathbf{w}_U^*$ dans les deux étapes. En effet, en partitionnant $\boldsymbol{\mathcal{X}}_U$ par ses deux sous-matrices de colonnes en tant que $\boldsymbol{\mathcal{X}}_U = (\boldsymbol{\mathcal{X}}_{12U}, \boldsymbol{\mathcal{X}}_{1U})$ et en relevant que $\boldsymbol{\mathcal{X}}_{12U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U} = \mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, nous pouvons facilement décomposer le vecteur $\mathbf{c}_U^*$ comme suit :

$$\begin{aligned} \mathbf{c}_U^* \;&= \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{12U} (\boldsymbol{\mathcal{X}}_{12U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{12U})^{-1} \big(\mathbf{0} - \boldsymbol{\mathcal{X}}_{12U}' \mathbf{w}_U^* \big) \\ &\quad + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U} (\boldsymbol{\mathcal{X}}_{1U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U})^{-1} \big(\mathbf{t}_{\mathbf{x}_1} - \boldsymbol{\mathcal{X}}_{1U}' \mathbf{w}_U^* \big). \end{aligned} \tag{3.13}$$

Dans le membre de droite de (3.13), la somme des premier et deuxième termes résulte d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_{12U}' \mathbf{c}_U^* = \mathbf{X}_U' \mathbf{c}_{1U} - \mathbf{X}_U' \mathbf{c}_U = \mathbf{0}$ seulement, tandis que la somme des premier et troisième termes résulte d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_{1U}' \mathbf{c}_U^* = \mathbf{t}_{\mathbf{x}_1}$ seulement.

Maintenant, si nous posons $\boldsymbol{\Delta}_1 = \mathrm{Var}(\mathbf{w}_{1U})$ et $\boldsymbol{\Delta}_2 = \mathrm{Var}(\mathbf{w}_U)$, ces variances étant précisées dans (3.8), il découle facilement de (3.13) que les estimateurs par calage optimaux $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{x}}^{B}$ dans (3.10) peuvent s'exprimer sous une forme explicite, qui sera reprise plus tard :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{B} \;&= \tilde{\mathbf{t}}_{\mathbf{y}} + \big[\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U - \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_U \big] \big[\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U - \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_U \big]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^{B} \;&= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.14}$$

4. Estimation linéaire optimale dans un échantillonnage à deux phases

4.1 L'estimateur optimal à deux phases

La matrice $\mathcal{B}$ de (3.7) comprend les variances et les covariances qui doivent être estimées. Compte tenu de $\mathrm{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U$, et en reprenant (3.8), les estimations sans biais évidentes sont $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$ et $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$, où la matrice $(n_1 + n_2) \times (n_1 + n_2)$ $\hat{\boldsymbol{\Delta}} = \widehat{\mathrm{Var}}(\mathbf{w}_U^*)$ comporte des blocs diagonaux $\hat{\boldsymbol{\Delta}}_1 = \{(\pi_{1kl} - \pi_{1k}\pi_{1l})/(\pi_{1k}\pi_{1l}\pi_{1kl})\}$, $\hat{\boldsymbol{\Delta}}_2 = \{(\pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l})/(\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\pi_{1kl}\pi_{2kl})\}$, et des blocs non diagonaux $\hat{\boldsymbol{\Delta}}_{12}$, $\hat{\boldsymbol{\Delta}}_{21} = \hat{\boldsymbol{\Delta}}_{12}'$ avec $\hat{\boldsymbol{\Delta}}_{12} = \{(\pi_{1kl} - \pi_{1k}\pi_{1l})/(\pi_{1k}\pi_{1l}\pi_{1kl}\pi_{2l})\}$, et où $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Psi}$ sont les matrices d'échantillon de (3.5).

Nous obtenons alors, comme éléments des matrices $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ et $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, les estimations sans biais des toutes les variances et covariances de (3.9), c'est-à-dire $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$, $\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2$, $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{X}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2$. Cependant, la matrice $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ comprend aussi les éléments $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_{12} \mathbf{X}_2$ et $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1 - \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_{12} \mathbf{X}_2$, qui manifestement ne retiennent pas les propriétés $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, respectivement. Les estimations sans biais pour les variances et les covariances de (3.9) pourraient être utilisées directement, mais l'estimation de la forme simple $\mathcal{B}$ dans (3.9) ne pourrait pas être exprimée comme étant $\boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1}$; donc l'estimateur qui en résulte ne conservait pas la forme de calage du MELSB dans (3.12). Cette complication est contournée par la reformulation suivante. Redéfinissons $\mathbf{w}^*$, $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{\Psi}$ comme étant :

$$\mathbf{w}^* = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w} \\ \mathbf{w}_1 \end{pmatrix}, \qquad \boldsymbol{\mathcal{X}} = \begin{pmatrix} -\mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{11} \end{pmatrix}, \qquad \boldsymbol{\Psi} = \begin{pmatrix} -\mathbf{Y}_1 & -\mathbf{X}_1 \\ \mathbf{Y}_2 & \mathbf{X}_2 \\ \mathbf{Y}_1 & \mathbf{X}_1 \end{pmatrix}, \tag{4.1}$$

où les matrices d'échantillon $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_{11}$ et $\mathbf{Y}_2$ sont les mêmes que précédemment et $\mathbf{Y}_1$ est la matrice de $\mathbf{y}$ pour l'échantillon s_1 avec les valeurs fictives $\mathbf{y}_k$ pour $k \notin s_2$. Il apparaît clairement que $\boldsymbol{\mathcal{X}}' \mathbf{w}^*$ et $\boldsymbol{\Psi}' \mathbf{w}^*$ sont exactement comme dans (3.6). Ensuite, en ayant comme précédemment $\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}}' \mathbf{w}^*$ et $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}' \mathbf{w}^*$, nous obtenons encore $\mathcal{B} = \mathrm{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1}$, où $\mathrm{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}_U' \mathrm{Var}(\mathbf{w}_U^*) \boldsymbol{\mathcal{X}}_U$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}_U' \mathrm{Var}(\mathbf{w}_U^*) \boldsymbol{\mathcal{X}}$, comme dans (3.7), mais $\mathbf{w}_U^*$, $\boldsymbol{\mathcal{X}}_U$ et $\boldsymbol{\Psi}_U$ étant les contreparties de population des valeurs redéfinies $\mathbf{w}^*$, $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Psi}$. Une extension du lemme 1 à la variable redéfinie $\mathbf{w}^*$ donne :

$$\mathrm{Var}(\mathbf{w}_U^*) = \begin{pmatrix} \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) \\ \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_U) & \mathrm{Var}(\mathbf{w}_{1U}) \\ \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) \end{pmatrix},$$

où $\mathrm{Var}(\mathbf{w}_{1U})$ et $\mathrm{Var}(\mathbf{w}_U)$ sont les mêmes que dans le lemme 1. Il est maintenant facile de vérifier qu'encore une fois $\mathcal{B}$ peut être exprimé analytiquement comme dans (3.9) et que les deux composantes

du MELSB sont identiques à celles données par (3.10). Plus important encore, il découle de cette forme spéciale de $\text{Var}(\mathbf{w}_U^*)$ que nous avons de nouveau $\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U$ et $\text{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U$, où maintenant $\boldsymbol{\Delta} = \text{diag}(-\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_1)$ et $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ comme cela a été défini précédemment. Ainsi, nous obtenons encore le MELSB sous la forme de calage de (3.12), et la décomposition orthogonale retenue du vecteur des poids calés dans (3.13) mène facilement à l'expression (3.14). La propriété d'orthogonalité $\boldsymbol{\mathcal{X}}_{12U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U} = \mathbf{0}$ est induite par la structure diagonale par blocs de la variable $\boldsymbol{\mathcal{X}}_U$ redéfinie, plutôt que par la structure spéciale de la matrice initiale $\boldsymbol{\Delta}$ utilisée dans (3.12).

Pour ce qui est du MELSB reconstruit, nous avons maintenant les estimations sans biais $\widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$ et $\widehat{\text{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$, où $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Psi}$ sont les matrices d'échantillon dans (4.1), et $\hat{\boldsymbol{\Delta}} = \text{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$ avec $\hat{\boldsymbol{\Delta}}_1$, $\hat{\boldsymbol{\Delta}}_2$ tels qu'ils sont définis au début de la section. À partir de là, nous recalculons facilement les estimations sans biais des variances et covariances de (3.9), mais deux des éléments de la matrice d'échantillon $\boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$ qui comprennent $\mathbf{Y}_1$, à savoir $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$ et $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$, nécessitent une attention particulière. Les valeurs fictives (non observées) $\mathbf{y}_k$ pour $k \notin s_2$, nécessaires à l'élargissement de $\mathbf{Y}_1$ à la matrice de population $\mathbf{Y}_U$ dans le MELSB reconstruit, sont fixées égales à zéro, et les valeurs $\mathbf{y}_k$ pour $k \in s_2$ sont alors nécessairement pondérées par $1/\pi_{2k}$. Ensuite $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$ et $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$ se réduisent à $\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1$ et $\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, qui sont respectivement les estimations sans biais $\widehat{\text{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}})$ et $\widehat{\text{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})$. La variable $\boldsymbol{\mathcal{B}}$ estimée dans (3.9) est maintenant donnée par :

$$\hat{\boldsymbol{\mathcal{B}}} = \begin{bmatrix} \left[\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1\right] \left[\mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1\right]^{-1} & \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} \\ \mathbf{I} & \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} \end{bmatrix}.$$

Le MELSB $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \boldsymbol{\mathcal{B}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ comportant une variable $\boldsymbol{\mathcal{B}}$ estimée sera appelé « estimateur linéaire sans biais optimal », « estimateur optimal » en abrégé, et sera représenté par $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ et ses deux composantes seront données par :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^O &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1\right] \left[\mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \qquad (4.2) \\ \hat{\mathbf{t}}_{\mathbf{x}}^O &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned}$$

Il s'agit de la version d'échantillon des MELSB dans (3.14), avec des coefficients estimés. En particulier, $\hat{\mathbf{t}}_{\mathbf{x}}^O$ est l'estimateur optimal à une phase habituel de $\mathbf{t}_{\mathbf{x}}$ reposant sur $\mathbf{x}_1$ comme variable auxiliaire et les données de l'échantillon complet de première phase s_1; voir Montanari (1987) et Rao (1994).

Remarque 4.1. Quand n_2 est très proche de n_1, l'estimateur optimal $\hat{\mathbf{t}}_{\mathbf{y}}^O$ peut être assez instable en raison de la quasi-singularité de la matrice inversée dans le coefficient de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$, et ainsi devenir très inefficace; voir cependant la remarque 6.1 ultérieure concernant les plans à deux phases dans lesquels cette instabilité ne pose pas de problème. En général, il ne s'agit pas d'une configuration réaliste dans un échantillonnage à deux phases, où n_2 est généralement beaucoup plus petit que n_1.

Après la construction de $\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_1$ et $\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_{11}$ en tant que deux des estimations dans $\hat{\boldsymbol{\mathcal{B}}}$, il apparaît que ces deux formes bilinéaires peuvent s'exprimer autrement sous la forme $\breve{\mathbf{Y}}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1$ et $\breve{\mathbf{Y}}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}$, respectivement, où $\breve{\mathbf{Y}}_1$ est une version pondérée de $\mathbf{Y}_1$ dans laquelle $\breve{\mathbf{y}}_k = \mathbf{y}_k / \pi_{2k}$ si $k \in s_2$ et $\breve{\mathbf{y}}_k = 0$ si $k \notin s_2$. Alors $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}'\mathbf{w}^* = \breve{\boldsymbol{\Psi}}'\mathbf{w}^*$, où $\breve{\boldsymbol{\Psi}}$ est $\boldsymbol{\Psi}$ dans (4.1), avec $\breve{\mathbf{Y}}_1$ à la place de $\mathbf{Y}_1$, et $\hat{\boldsymbol{\mathcal{B}}}$ peut être formulé de façon compacte comme étant $\hat{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{\Psi}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}$, où $\hat{\boldsymbol{\Delta}} = \text{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$. À partir de maintenant, $\hat{\boldsymbol{\Delta}}$ désignera la matrice $\text{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$.

Comme c'est le cas dans Montanari (1987) et Rao (1994) pour l'estimateur optimal à une phase, pour les grands échantillons s_1 et s_2, l'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ est une approximation du MELSB $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B$, et il est donc approximativement sans biais. De plus, la variance de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$ est une approximation de celle de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B$, qui se résout facilement pour donner $\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B) = \text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}) - \text{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\left[\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1}\text{Cov}'(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, c'est-à-dire la forme compacte de (3.11). Ensuite, au moyen des estimations $\widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}})$ et $\widehat{\text{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, calculées auparavant, nous obtenons la variance approximative estimée de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$ comme étant $\widehat{\text{VA}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O) = \widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}) - \widehat{\text{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\left[\widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1}\widehat{\text{Cov}}'(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$. À partir de là, nous calculons les expressions pratiques pour le calcul $\widehat{\text{VA}}(\hat{\mathbf{t}}_{\text{y}}^O) = \mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{Y}_2 - \breve{\boldsymbol{\Psi}}_1'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\breve{\boldsymbol{\Psi}}_1$, où $\breve{\boldsymbol{\Psi}}_1$ est la première sous-matrice de colonnes de $\breve{\boldsymbol{\Psi}}$, et $\widehat{\text{VA}}(\hat{\mathbf{t}}_{\text{x}}^O) = \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\left[\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\right]^{-1}\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1$.

4.2 L'estimateur optimal à deux phases comme estimateur par calage

L'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, avec $\hat{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{\Psi}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}$, prend la forme

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \breve{\boldsymbol{\Psi}}'\left[\mathbf{w}^* + \hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*)\right],$$

d'un estimateur par calage, avec un vecteur de totaux de calage $\mathbf{t}_{\boldsymbol{\mathcal{X}}}$ et un vecteur d'échantillon de poids calés $\mathbf{c}^* = \mathbf{w}^* + \hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}\left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*\right)$ satisfaisant à $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$. Cela est établi formellement par le théorème suivant, dont la démonstration est omise, car elle est semblable à celle du théorème 1.

Théorème 2. *Le vecteur* $\mathbf{c}^* = \mathbf{w}^* + \hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*)$ *minimise la distance des moindres carrés généralisés* $(\mathbf{c}^* - \mathbf{w}^*)'\hat{\boldsymbol{\Delta}}^{-1}(\mathbf{c}^* - \mathbf{w}^*)$ *soumise aux contraintes* $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, *c'est-à-dire* $\mathbf{X}_1'\mathbf{c}_1 = \mathbf{X}_2'\mathbf{c}$ *et* $\mathbf{X}_{11}'\mathbf{c}_1 = \mathbf{t}_{\mathbf{x}_1}$, *où* $(\mathbf{c}_1, \mathbf{c})$ *correspond à* $(\mathbf{w}_1, \mathbf{w})$.

Le vecteur d'échantillon $\mathbf{c}^*$ admet la même décomposition orthogonale que sa contrepartie de population $\mathbf{c}_U^*$ dans (3.13). Nous pouvons maintenant exprimer formellement l'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$ en tant qu'estimateur par calage $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \breve{\boldsymbol{\Psi}}'\mathbf{c}^*$ qui, compte tenu de $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, est généré par le calage simultané des deux estimations $\hat{\mathbf{t}}_{\mathbf{x}}$ et $\tilde{\mathbf{t}}_{\mathbf{x}}$ de $\mathbf{t}_{\mathbf{x}}$ l'une sur l'autre, et de l'estimation $\hat{\mathbf{t}}_{\mathbf{x}_1}$ sur le total $\mathbf{t}_{\mathbf{x}_1}$.

Sous une forme développée, le vecteur $\mathbf{c}^*$ est :

$$\mathbf{c}^* = \begin{pmatrix}\mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\end{pmatrix} = \begin{pmatrix}\mathbf{w}_1 + \hat{\boldsymbol{\Delta}}_1\mathbf{X}_1[\mathbf{X}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\\ \mathbf{w} + \hat{\boldsymbol{\Delta}}_2\mathbf{X}_2[\mathbf{X}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\\ \mathbf{w}_1 + \hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}(\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11})^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})\end{pmatrix}. \tag{4.3}$$

Ensuite, au moyen de la partition $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{12}, \boldsymbol{\mathcal{X}}_1)$, où $\boldsymbol{\mathcal{X}}_{12}$ et $\boldsymbol{\mathcal{X}}_1$ sont les deux sous-matrices de colonnes orthogonales de $\boldsymbol{\mathcal{X}}$ figurant dans (4.1), les deux contraintes s'expriment par $\boldsymbol{\mathcal{X}}_{12}' \mathbf{c}^* = \mathbf{X}_2' \mathbf{c}_2 - \mathbf{X}_1' \mathbf{c}_1 = \mathbf{0}$ et $\boldsymbol{\mathcal{X}}_1' \mathbf{c}^* = \mathbf{X}_{11}' \mathbf{c}_3 = \mathbf{t}_{\mathbf{x}_1}$. Il découle aussi de (4.3) que $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \breve{\boldsymbol{\Psi}}' \mathbf{c}^*$ implique (4.2). Concernant les deux composantes de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$, nous observons que $\hat{\mathbf{t}}_{\mathbf{x}}^O = -\mathbf{X}_1' \mathbf{c}_1 + \mathbf{X}_2' \mathbf{c}_2 + \mathbf{X}_1' \mathbf{c}_3 = \mathbf{X}_1' \mathbf{c}_3$ et que

$$\hat{\mathbf{t}}_{\mathbf{y}}^O = \breve{\mathbf{Y}}_1' (\mathbf{c}_3 - \mathbf{c}_1) + \mathbf{Y}_2' \mathbf{c}_2 = \sum\nolimits_{s_2} \left[(c_{3k} - c_{1k}) / \pi_{2k} + c_{2k} \right] \mathbf{y}_k .$$

L'expression explicite de $\hat{\mathbf{t}}_{\mathbf{y}}^O$, en termes d'unités d'échantillonnage, est :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^O \quad &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\sum\nolimits_{s_2} \sum\nolimits_{s_2} \hat{\Delta}_{2kl} \, \mathbf{y}_k \, \mathbf{x}_l' - \sum\nolimits_{s_2} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \breve{\mathbf{y}}_k \, \mathbf{x}_l' \right] \times \\ &\quad \left[\sum\nolimits_{s_2} \sum\nolimits_{s_2} \hat{\Delta}_{2kl} \, \mathbf{x}_k \, \mathbf{x}_l' - \sum\nolimits_{s_1} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \mathbf{x}_k \, \mathbf{x}_l' \right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \left(\sum\nolimits_{s_2} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \breve{\mathbf{y}}_k \, \mathbf{x}_{1l}' \right) \left(\sum\nolimits_{s_1} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \mathbf{x}_{1k} \, \mathbf{x}_{1l}' \right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}), \end{aligned} \tag{4.4}$$

où $\hat{\Delta}_{1kl}$ et $\hat{\Delta}_{2kl}$ sont les kl^{e} éléments de $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$, respectivement. La formule (4.4) est simplifiée dans certains plans à deux phases employés dans d'importantes enquêtes à grande échelle. Des exemples en sont donnés dans Hidiroglou et Särndal (1998) et Turmelle et Beaucage (2013). Plus précisément, c'est le cas quand un échantillonnage indépendant (Poisson, ou Poisson stratifié) est utilisé dans l'une des deux phases, c'est-à-dire quand $\pi_{1kl} = \pi_{1k} \pi_{1l}$ ou $\pi_{2kl} = \pi_{2k} \pi_{2l}$. La simplification est considérable quand les deux phases ont un échantillonnage indépendant. Alors, $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ sont toutes deux diagonales, avec des éléments diagonaux $\hat{\Delta}_{1kk} = (1/\pi_{1k})((1/\pi_{1k}) - 1)$ et $\hat{\Delta}_{2kk} = (1/\pi_{1k}\pi_{2k})((1/\pi_{1k}\pi_{2k}) - 1)$, respectivement, et (4.4) comporte uniquement des sommations simples. D'autres plans de sondage à deux phases dans lesquels (4.4) comporte seulement des sommations simples, bien que $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ ne soient pas des diagonales, comportent un échantillonnage aléatoire simple ou un échantillonnage aléatoire simple stratifié dans l'une ou l'autre des phases. Voir un exemple d'enquête avec ce type de plan d'échantillonnage à deux phases dans Hidiroglou (2001). Toutefois, en général, l'estimateur optimal peut se révéler peu commode, car il nécessite l'utilisation des probabilités d'inclusion conjointes de la première phase et de la deuxième phase π_{1kl} et π_{2kl}, qui ne sont pas connues pour certains plans de sondage complexes. Même quand ces probabilités conjointes sont connues, mais que les matrices $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ ne sont pas diagonales, le coefficient estimé $\hat{\boldsymbol{\mathcal{B}}}$ et, par conséquent, l'estimateur optimal peuvent être instables en cas de très petits échantillons, surtout si la dimension du vecteur auxiliaire $\mathbf{x}$ est grande. Il est possible de surmonter ces difficultés, avec une certaine perte d'optimalité, en utilisant des approximations simples des variances et des covariances dans $\hat{\boldsymbol{\mathcal{B}}}$; pour ce qui est des estimations de la variance approximatives fondées uniquement sur des probabilités d'inclusion du premier ordre, voir par exemple Haziza, Mecatti et Rao (2008) et les références qui s'y trouvent. Une approximation de $\hat{\boldsymbol{\mathcal{B}}}$ très pratique pour le calcul permettant d'obtenir un estimateur à deux phases qui appartient à la classe des estimateurs par la régression généralisée est décrite dans la section suivante.

5. Un estimateur par la régression généralisée en deux phases

Une variante de $\hat{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{\Psi}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1}$, efficace pour ce qui est du calcul, mais qui n'est généralement pas optimale, est le coefficient de régression généralisée (GREG) $\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}} = \boldsymbol{\Psi}' \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}})^{-1}$, où $\boldsymbol{\Psi}$ est comme dans (4.1) avec $\mathbf{y}_k = 0$ si $k \notin s_2$, et $\boldsymbol{\Lambda}$ est la matrice de « pondération » $\mathrm{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_1)$, étant donné que $\boldsymbol{\Lambda}_1 = \mathrm{diag}\{w_{1k} / q_{1k}\}$ et $\boldsymbol{\Lambda}_2 = \mathrm{diag}\{w_k / q_{2k}\}$, et que q_{1k}, q_{2k} sont des constantes positives. Cela donne l'estimateur par la régression généralisée (GREG) :

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}} (\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}} \mathbf{t}_{\boldsymbol{\mathcal{X}}} + (\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime})' \mathbf{w}^{*}. \tag{5.1}$$

Notons que $\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}$ est optimal dans le sens des moindres carrés, c'est-à-dire qu'il minimise la distance quadratique $(\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime})' \boldsymbol{\Lambda} (\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime})$, impliquant les résidus $\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime}$ dans $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$, alors que le coefficient $\hat{\boldsymbol{\mathcal{B}}}$ minimise $(\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\boldsymbol{\mathcal{B}}})' \hat{\boldsymbol{\Delta}} (\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\boldsymbol{\mathcal{B}}})'$, la variance approximative estimée de l'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$. En ce sens, $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$ est une approximation de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$. Les deux composants de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$, similaires dans leur structure aux composantes de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$ dans (4.2), sont :

$$\begin{aligned}
\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 + \mathbf{Y}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_1\right] \left[\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 + \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_1\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\
&\quad + \mathbf{Y}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\
\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}} &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}).
\end{aligned} \tag{5.2}$$

L'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}}$ est l'estimateur GREG standard à une phase basé sur s_1 et la variable auxiliaire $\mathbf{x}_1$. L'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$, dont les deux termes de régression orthogonale sont indiqués dans (5.2), est exprimé explicitement en termes d'unités d'échantillon comme suit :

$$\begin{aligned}
\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\sum\nolimits_{s_2} (\Lambda_{1k} + \Lambda_{2k})\, \mathbf{y}_k\, \mathbf{x}_k'\right] \left[\sum\nolimits_{s_2} \Lambda_{2k}\, \mathbf{x}_k\, \mathbf{x}_k' + \sum\nolimits_{s_1} \Lambda_{1k}\, \mathbf{x}_k\, \mathbf{x}_k'\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\
&\quad + \left(\sum\nolimits_{s_2} \Lambda_{1k}\, \mathbf{y}_k\, \mathbf{x}_{1k}'\right) \left(\sum\nolimits_{s_1} \Lambda_{1k}\, \mathbf{x}_{1k}\, \mathbf{x}_{1k}'\right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}),
\end{aligned}$$

où $\Lambda_{1k} = w_{1k} / q_{1k}$ et $\Lambda_{2k} = w_k / q_{2k}$ sont le k^{e} élément de $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$, respectivement. Il faut spécifier les constantes q_{ik} comme étant $q_{ik} = n_i$ pour tenir compte de la différentielle dans la taille d'échantillon de s_i; voir dans Merkouris (2004) une justification dans le contexte d'un calage d'échantillons combinés. Un ajustement équivalent des poids dans Λ_{1k} et Λ_{2k} peut être effectué par la multiplication de w_{1k} dans Λ_{1k} par $\phi = n_2 / n_1$. Les valeurs de q_{ik} qui convertissent l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ en l'estimateur optimal $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ peuvent être spécifiées pour les plans d'échantillonnage à deux phases pour lesquels une estimation optimale est possible, comme dans le contexte similaire de l'échantillonnage matriciel (Merkouris, 2015). Pour l'exemple simple comprenant un échantillonnage de Poisson dans les deux phases, cette spécification est $q_{1k} = \pi_{1k} / (1 - \pi_{1k})$ et $q_{2k} = \pi_{1k}\, \pi_{2k} / (1 - \pi_{1k}\, \pi_{2k})$, rendant $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ identiques à $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$.

Le vecteur des poids calés associés à l'estimateur GREG $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$ est $\mathbf{c}^{\mathrm{GR}} = \mathbf{w}^{*} + \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}})^{-1} (\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}' \mathbf{w}^{*})$. Il a la même forme que $\mathbf{c}^{*}$ dans (4.3), mais avec $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ à la place de $-\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$, et il minimise la distance des moindres carrés généralisés $(\mathbf{c}^{\mathrm{GR}} - \mathbf{w}^{*})' \boldsymbol{\Lambda}^{-1} (\mathbf{c}^{\mathrm{GR}} - \mathbf{w}^{*})$ soumise

aux contraintes $\boldsymbol{\mathcal{X}}'\mathbf{c}^{\mathrm{GR}} = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$. La partition $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{12}, \boldsymbol{\mathcal{X}}_1)$, définie après (4.3), permet la décomposition orthogonale du vecteur $\mathbf{c}^*$:

$$\begin{aligned}\mathbf{c}^{\mathrm{GR}} &= \mathbf{w}^* + \boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12}(\boldsymbol{\mathcal{X}}_{12}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}_{12}'\mathbf{w}^*) \\ &\quad + \boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1(\boldsymbol{\mathcal{X}}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1)^{-1}(\mathbf{t}_{\mathbf{x}_1} - \boldsymbol{\mathcal{X}}_1'\mathbf{w}^*).\end{aligned} \tag{5.3}$$

Dans le membre de droite de (5.3), la somme des premier et deuxième termes résulterait d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_{12}'\mathbf{c}^{\mathrm{GR}} = \mathbf{0}$ seulement, tandis que la somme du premier et du troisième terme résulterait d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_1'\mathbf{c}^{\mathrm{GR}} = \mathbf{t}_{\mathbf{x}_1}$ seulement. En pratique, cela signifie que le vecteur $\mathbf{c}^*$ pourrait être formé en concaténant les vecteurs de poids générés par deux calages distincts, c'est-à-dire le calage de $(\mathbf{w}_1', \mathbf{w}')'$ au moyen de $(-\mathbf{X}_1', \mathbf{X}_2')'$ suivi du calage de $\mathbf{w}_1$ au moyen de $\mathbf{X}_{11}$. Cependant, la procédure de calage en une étape générant $\mathbf{c}^{\mathrm{GR}}$ est plus pratique.

À partir de sa linéarisation en séries de Taylor, l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ dans (5.1) est approximativement (pour les grands échantillons) sans biais. De plus, si l'on désigne par $\mathbf{e}$ la matrice des résidus d'échantillon $\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}'}$, la variance approximative estimée de $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} = \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}\mathbf{t}_{\boldsymbol{\mathcal{X}}} + \mathbf{e}'\mathbf{w}^*$ est la variance estimée de $\mathbf{e}'\mathbf{w}^*$, c'est-à-dire $\widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}) = \widehat{\mathrm{Var}}(\mathbf{e}'\mathbf{w}^*) = \mathbf{e}'\hat{\boldsymbol{\Delta}}\mathbf{e}$, tandis que la variance estimée de l'estimateur de HT $\tilde{\mathbf{t}}_{\mathbf{y}}$ est $\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) = \widehat{\mathrm{Var}}(\boldsymbol{\Psi}_1'\mathbf{w}^*) = \mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{Y}_2$, où $\boldsymbol{\Psi}_1$ est la première sous-matrice de colonnes de $\boldsymbol{\Psi}$.

En utilisant maintenant la forme de calage $\boldsymbol{\Psi}_1'\mathbf{c}^{\mathrm{GR}}$ de $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ et la décomposition orthogonale (5.3) de $\mathbf{c}^{\mathrm{GR}}$, nous obtenons facilement la décomposition $\mathbf{e} = \boldsymbol{\Psi}_1 - \boldsymbol{\mathcal{X}}_{12}\hat{\boldsymbol{\beta}}_{\mathbf{x}}' - \boldsymbol{\mathcal{X}}_1\hat{\boldsymbol{\beta}}_{\mathbf{x}_1}'$, où $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \boldsymbol{\Psi}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12}(\boldsymbol{\mathcal{X}}_{12}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12})^{-1}$ et $\hat{\boldsymbol{\beta}}_{\mathbf{x}_1} = \boldsymbol{\Psi}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1(\boldsymbol{\mathcal{X}}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1)^{-1}$ sont les coefficients de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$, respectivement. Mentionnons que $\boldsymbol{\Psi}_1 - \boldsymbol{\mathcal{X}}_{12}\hat{\boldsymbol{\beta}}_{\mathbf{x}}'$ est la matrice des résidus dans l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}} = \tilde{\mathbf{t}}_{\mathbf{y}} + \hat{\boldsymbol{\beta}}_{\mathbf{x}}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})$ qui résultent du calage comprenant seulement $\boldsymbol{\mathcal{X}}_{12}$, et $\boldsymbol{\Psi}_1 - \boldsymbol{\mathcal{X}}_1\hat{\boldsymbol{\beta}}_{\mathbf{x}_1}'$ est la matrice des résidus dans l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}_1} = \tilde{\mathbf{t}}_{\mathbf{y}} + \hat{\boldsymbol{\beta}}_{\mathbf{x}_1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ qui résulte du calage comprenant seulement $\boldsymbol{\mathcal{X}}_1$. Ensuite, au moyen de l'orthogonalité de $\boldsymbol{\mathcal{X}}_1$ et $\boldsymbol{\mathcal{X}}_{12}$, nous démontrons sans difficulté que :

$$\widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) = \widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) + \widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}_1}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}),$$

ce qui implique que la réduction de la variance due à l'utilisation des deux variables auxiliaires $\mathbf{x}_1$ et $\mathbf{x}$ dans la procédure de régression (ou de calage) est additive. Ainsi, si l'on reprend la remarque 3.2, l'estimateur par la régression généralisée retient cette propriété d'additivité du MELSB de $\mathbf{t}_{\mathbf{y}}$.

6. Comparaisons avec les méthodes existantes

Hidiroglou (2001) décrit une méthode antérieure de l'estimation linéaire optimale dans les plans de sondage à deux phases, qui comprend le type standard d'information auxiliaire examiné dans les sections 2 à 5. La formulation commence par poser le postulat d'une forme de régression pour l'estimateur de $\mathbf{t}_{\mathbf{y}}$, pour un y univarié , qui est identique à la forme de l'estimateur $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ à la première ligne de (3.2);

ensuite, les deux coefficients inconnus sont déterminés de façon à minimiser la variance de cet estimateur. Dans l'estimation des deux coefficients, les identités $\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}}) - \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$ n'ont pas été prises en compte dans les premier et deuxième coefficients, respectivement, et les variances et covariances dans les deux coefficients comportant des estimateurs de la première phase ont été estimées au moyen de l'échantillon de la deuxième phase seulement, faisant ainsi abstraction des renseignements pertinents provenant de la plus grande partie de l'échantillon de la première phase. L'estimateur qui en résulte ne s'est pas révélé être un estimateur par calage. De fait, cette version de l'estimateur optimal ne peut pas être construite comme un estimateur par calage. Comme variante pouvant s'appliquer à cela, Hidiroglou (2001) a considéré un estimateur GREG dont les deux coefficients (de $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ et $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$) peuvent être justifiés soit en supposant des modèles de régression différents pour chaque phase, soit par deux calages successifs. Ce même estimateur GREG avait été proposé précédemment par Hidiroglou et Särndal (1998), mais sans référence à l'estimation optimale. Dans l'approche par calage d'Hidiroglou et de Särndal (1998), l'estimateur $\hat{\mathbf{t}}_{\mathbf{x}_1}$ est d'abord calé sur son total $\mathbf{t}_{\mathbf{x}_1}$, au moyen de s_1, et l'estimateur GREG $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}|\mathbf{x}_1}$ de $\mathbf{t}_{\mathbf{x}}$ est ensuite généré au moyen des poids calés représentés par $\tilde{w}_{1k}$. Alors, le poids total de $k \in s_2$ est formé comme étant $\tilde{w}_k = \tilde{w}_{1k}\, w_{2k}$. Dans un deuxième calage comprenant s_2 et $\tilde{w}_k$, l'estimateur $\tilde{\mathbf{t}}_{\mathbf{x}}$ est calé sur $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}|\mathbf{x}_1}$. Les poids calés de s_2 qui en résultent servent ensuite à générer l'estimateur GREG de $\mathbf{t}_{\mathbf{y}}$, désigné ici par $\hat{\mathbf{t}}^{\mathrm{HS}}_{\mathbf{y}}$.

Estevao et Särndal (2002, 2009) ont proposé une version plus simple de l'estimateur $\hat{\mathbf{t}}^{\mathrm{HS}}_{\mathbf{y}}$, dans laquelle les poids de sondage globaux $w_k = w_{1k}\, w_{2k}$ pour $k \in s_2$ sont utilisés dans le deuxième calage. Au moyen de la notation actuelle, cet estimateur, représenté ici par $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$, peut être exprimé sous forme de régression comme suit :

$$\begin{aligned}\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 \left(\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2\right)^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 \left(\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2\right)^{-1} \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}).\end{aligned} \tag{6.1}$$

Ici, les matrices de pondération standards $\boldsymbol{\Lambda}_1 = \mathrm{diag}\{w_{1k}\}$ et $\boldsymbol{\Lambda}_2 = \mathrm{diag}\{w_k\}$ sont utilisées. Estevao et Särndal (2009) ont démontré que cet estimateur est asymptotiquement équivalent à l'estimateur $\hat{\mathbf{t}}^{\mathrm{HS}}_{\mathbf{y}}$. Pour l'estimateur $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$ figurant dans (6.1), Estevao et Särndal (2002) fournissent deux représentations de régression linéaire correspondant aux deux étapes de calage. En remplaçant $\mathbf{y}$ par $\mathbf{x}$ dans (6.1), nous obtenons $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{x}}$, qui est identique à $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}}$ dans (5.2).

À titre de comparaison, l'estimateur par la régression proposé dans la section 5 est motivé par la structure de calage à une seule étape de l'estimateur optimal à deux phases, dont il est une approximation utilisée à des fins de commodité. Comparativement aux estimateurs par la régression concurrents évalués dans la présente section, il tire son efficacité statistique et computationnelle d'une procédure de calage à une seule étape comprenant les échantillons combinés de la première phase et de la deuxième phase, et dans laquelle les totaux estimés de la première phase et de la deuxième phase sont calés les uns sur les autres. Par conséquent, les coefficients de régression des termes $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ et $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ intègrent l'information de l'échantillon complet s_1, comme dans l'estimateur optimal, et pour cette raison, ils sont des

estimations plus stables de leurs contreparties de population. Une comparaison empirique de l'estimateur par la régression proposé avec les estimateurs par la régression concurrents est présentée dans l'étude par simulations à la section 7.

Le remplacement de $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ par $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ dans (6.1) convertit le coefficient de l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{\mathrm{GR}}$ généré par le calage de la première étape en coefficient $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})[\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}_1})]^{-1}$ de l'estimateur par la régression optimal à une seule phase $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{O}$, et le coefficient de l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$ généré par le calage de la deuxième étape en coefficient $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})[\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{x}})]^{-1}$. Ce dernier coefficient peut être considéré comme pseudo-optimal puisque $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{O}$ est traité comme une constante dans le calage de la deuxième étape, ce qui génère un estimateur pseudo-optimal $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$. Ensuite, si à la place des matrices d'échantillon $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ dans (6.1), nous utilisons les matrices de population $\boldsymbol{\Delta}_1$ et $\boldsymbol{\Delta}_2$, nous construisons le pseudo-MELSB :

$$\begin{aligned}\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \left(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}\right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}),\end{aligned} \tag{6.2}$$

où les coefficients de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ sont respectivement $\mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})\left[\mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}$ et $\mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})\left[\mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}$. Ainsi, l'estimateur GREG (6.1) peut être considéré comme une approximation de $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$, qui est l'estimateur $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$ avec des coefficients estimés (par analogie avec la relation entre l'estimateur optimal $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ et le MELSB $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$, dans (4.2) et (3.14)). L'estimateur pseudo-MELSB $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{PSB}}$, obtenu à partir de (6.2) en remplaçant $\mathbf{y}$ par $\mathbf{x}$, est identique au MELSB $\hat{\mathbf{t}}_{\mathbf{x}}^{B}$, dans (3.14). Par ailleurs, les estimateurs $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$ sont identiques seulement à condition de respecter la proposition suivante (voir la démonstration en annexe).

Proposition 1. Les estimateurs $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$ sont identiques si et seulement si $\boldsymbol{\Delta}_1 = \delta\boldsymbol{\Delta}_2$ pour une constante δ.

Remarque 6.1. La condition de la proposition 1 se vérifie si le même plan avec probabilités d'inclusion égales est utilisé dans les deux phases; la constante δ est alors une fonction des probabilités d'inclusion de l'échantillon. Les plans à deux phases qui satisfont à cette condition sont l'échantillonnage aléatoire simple et l'échantillonnage de Bernoulli dans les deux phases, ainsi que leurs versions de stratification avec stratification identique et répartition proportionnelle de l'échantillon dans les deux phases. Cette condition a une importance pratique : pour ces plans, les contreparties d'échantillon de $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$, c'est-à-dire $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$, seront presque identiques en cas de grands échantillons. De plus, $\boldsymbol{\Delta}_1 = \delta\boldsymbol{\Delta}_2$ laisse supposer que le signe moins dans le coefficient de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ dans (3.14) et (4.2) pourrait devenir un signe plus, $1-\delta$ étant exclus, ce qui annulerait le problème de singularité mentionné dans la remarque 4.1.

Remarque 6.2. On peut vérifier simplement que $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$ est un estimateur par calage, construit par une procédure de calage en deux étapes (comme pour le $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$ estimateur). De plus, comme $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$, l'estimateur $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$ est formé au moyen des poids calés de l'échantillon de la deuxième phase seulement.

Enfin, il faut mentionner que Chen et Kim (2014) ont proposé un « estimateur par la régression optimal par rapport au plan », sans la propriété de calage, pour une application et une configuration de variable auxiliaire particulières. De plus, Wu et Luan (2003) ont proposé un estimateur par calage qui est optimal dans un cadre assisté par modèle et qui a $\mathbf{x}_2$ comme seul vecteur auxiliaire.

7. Étude par simulations

Nous avons réalisé une étude par simulations pour évaluer les performances des estimateurs à deux phases du total t_y proposés, pour les variables scalaires y, x_1 et x_2, et les comparer avec les estimateurs par régression concurrents examinés ci-dessus. Les distributions de ces variables ont été spécifiées comme suit. La distribution de x_1 est une loi log-normale comportant les paramètres de moyenne et de variance $(\mu_{x_1} = 4, \sigma^2_{x_1} = 4)$. La distribution de x_2 est spécifiée par le modèle linéaire $x_2 = 5 + x_1 + \epsilon$, où $\epsilon \sim N(0, \sigma^2_\epsilon)$, et la distribution de y est spécifiée par le modèle linéaire $y = 10 + 2x_1 + 3x_2 + \eta$, où $\eta \sim N(0, \sigma^2_\eta)$. La valeur de σ^2_ϵ détermine la relation linéaire entre x_2 et x_1, telle qu'elle est définie par le carré du coefficient de corrélation de la population $r^2_{x_1, x_2}$, et la valeur de σ^2_η détermine la relation linéaire de y avec x_1 et x_2, telle qu'elle est définie par le coefficient de détermination $r^2 = [r^2_{y,x_1} + r^2_{y,x_2} - 2r_{y,x_1} r_{y,x_2} r_{x_1,x_2}] / (1 - r^2_{x_1,x_2})$.

Trois valeurs, 0, 0,25 et 0,75 ont été précisées pour $r^2_{x_1,x_2}$, et deux valeurs, 0,25 et 0,75, pour r^2, ce qui donne six combinaisons de valeurs $(r^2_{x_1,x_2}, r^2)$. Pour la valeur $r^2_{x_1,x_2} = 0$, en particulier, la distribution log-normale bivariée pour (x_1, x_2) avec les paramètres $(\mu_{x_1} = 4, \sigma^2_{x_1} = 4)$, $(\mu_{x_2} = 9, \sigma^2_{x_2} = 9)$ et une corrélation nulle a été utilisée. Les valeurs requises de σ^2_ϵ et σ^2_η sont facilement déterminées, tandis que les valeurs pour r^2_{y,x_1} et r^2_{y,x_2} sont implicitement précisées. Pour chacune de ces six combinaisons, nous avons simulé une population de taille $N =$ 50 000 en générant des valeurs à partir des distributions des composantes du vecteur (y, x_1, x_2). Quatre combinaisons de tailles d'échantillon de la première phase et de la deuxième phase (n_1, n_2) comportant une valeur n_1 fixe et une valeur n_2 variable ont été spécifiées, à savoir (3 000; 2 000), (3 000; 1 500), (3 000; 1 000), (3 000; 500), ce qui a créé un total de 24 configurations de simulation.

Trois plans de sondage à deux phases différents ont été examinés. Un échantillonnage aléatoire simple (EAS) sans remise a d'abord été utilisé dans les deux phases. Pour ce plan de sondage, désigné par (EAS, EAS), le MELSB $\hat{t}^B_y$ dans (3.10) et sa variance exacte dans (3.11) peuvent être calculés. En s'appuyant sur le fait que dans un EAS, la corrélation des estimateurs de HT pour deux totaux est identique au coefficient de corrélation des variables associées, des calculs algébriques fastidieux mais simples donnent la différence relative des variances (DRV) des estimateurs $\hat{t}^B_y$ et $\tilde{t}_y$ comme étant :

$$\frac{\mathrm{Var}(\tilde{t}_y) - \mathrm{Var}(\hat{t}^B_y)}{\mathrm{Var}(\tilde{t}_y)} = \frac{N(n_1 - n_2)}{n_1(N - n_2)} r^2 + \frac{n_2(N - n_1)}{n_1(N - n_2)} r^2_{y,x_1}.$$

La DRV en pourcentage est la mesure de l'efficacité du MELSB $\hat{t}^B_y$ par rapport à l'estimateur de HT $\tilde{t}_y$. Cette efficacité maximale exacte servira à mesurer le degré de correspondance de l'approximation du

MELSB par l'estimateur optimal, pour les différentes tailles d'échantillon, ainsi que l'efficacité des autres estimateurs concurrents par rapport à l'estimateur de HT. Nous constatons que quand n_2 tend vers zéro, la DRV tend vers r^2 et que quand n_2 tend vers n_1, la DRV tend vers r^2_{y,x_1} (l'efficacité du MELSB basé sur s_1 et x_1). Le deuxième plan à deux phases, représenté par (EASS, EAS), était un échantillonnage aléatoire simple stratifié (EASS) et un EAS à la première et à la deuxième phase, respectivement. Les populations simulées étaient stratifiées selon la taille de la variable y, avec trois strates de tailles $N_1 = 30\,000$, $N_2 = 15\,000$, $N_3 = 5\,000$ et une répartition proportionnelle de l'échantillon s_1 aux trois strates, ce qui donne des probabilités d'inclusion égales dans chacune des deux phases. Pour ce plan également, il est possible de calculer le MELSB $\hat{t}_y^B$ et sa variance exacte. Le troisième plan à deux phases, représenté par (EAS, ESPPT), comportait un EAS dans la première phase et un échantillonnage systématique avec probabilité proportionnelle à la taille (ESPPT) dans la deuxième phase, où la mesure de taille utilisée était la transformation simple $z_2 = 15 + 0{,}5x_2$ de la variable x_2; si x_2 était utilisé comme taille, nous obtiendrions $\hat{t}_{x_2} = \tilde{t}_{x_2}$. Dans ce cas, il est impossible de calculer le MELSB $\hat{t}_y^B$ (et l'estimateur optimal $\hat{t}_y^O$), en raison des probabilités inconnues π_{2kl}. Toutefois, les estimateurs GREG peuvent être calculés.

Pour chacun de ces trois plans à deux phases, et pour les 24 configurations de simulation, l'échantillonnage a été répété 30 000 fois; à chacune des répétitions, nous avons calculé les estimateurs $\tilde{t}_y$, $\hat{t}_y^O$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ et $\hat{t}_y^{\mathrm{HS}}$ pour obtenir leur biais empirique et leur variance. Dans tous ces cas, la simulation a montré que le biais de tous les estimateurs était négligeable, y compris pour les tailles de sous-échantillons n_2 les plus petites. C'est pourquoi leur comparaison est fondée sur leurs variances par rapport à la variance de référence de l'estimateur de HT $\tilde{t}_y$. Plus précisément, l'efficacité de chacun des estimateurs concurrents $\hat{t}_y^O$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ et $\hat{t}_y^{\mathrm{HS}}$ est évaluée par la différence relative en pourcentage entre sa variance empirique et la variance empirique de l'estimateur $\tilde{t}_y$; par exemple, pour $\hat{t}_y^{\mathrm{GR}}$ la différence relative est $\left[\mathrm{Var}(\tilde{t}_y) - \mathrm{Var}(\hat{t}_y^{\mathrm{GR}})\right] \big/ \mathrm{Var}(\tilde{t}_y)$. La différence relative montre la réduction de la variance de l'estimateur particulier par rapport à la variance de l'estimateur de base $\tilde{t}_y$.

Dans le plan (EAS, EAS), l'efficacité exacte du MELSB $\hat{t}_y^B$ par rapport à l'estimateur de HT $\tilde{t}_y$ augmente à mesure que n_2 diminue et que nous passons à des valeurs plus élevées de (r^2_{y,x_2}, r^2), ce qui corrobore l'information fournie dans la remarque 3.1; voir la colonne 2 du tableau 7.1. Il est aussi confirmé que l'efficacité de $\hat{t}_y^B$ tend vers r^2 quand n_2 diminue, plus rapidement pour une valeur de $r^2_{x_1,x_2}$ plus élevée. Une approximation proche de cette efficacité maximale est donnée par l'efficacité empirique de $\hat{t}_y^O$, même pour les tailles de sous-échantillons les plus petites n_2; voir la colonne 3 du tableau 7.1. Pour le plan (EASS, EAS), l'efficacité exacte de $\hat{t}_y^B$, indiquée dans la colonne 6 du tableau 7.1, présente une tendance semblable à celle observée dans le plan (EAS, EAS). Une approximation proche de cette efficacité est donnée par l'efficacité empirique de $\hat{t}_y^O$; voir la colonne 7 du tableau 7.1. Dans les deux plans (EAS, EAS) et (EASS, EAS), l'approximation de $\hat{t}_y^B$ par $\hat{t}_y^O$ est un peu plus faible dans certaines configurations comprenant la plus grande valeur de n_2, pour la raison donnée dans la remarque 4.1.

Bien que l'estimateur $\hat{t}_y^O$ puisse être calculé dans les plans (EAS, EAS) et (EASS, EAS), les performances des estimateurs par calage (GREG) $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$, plus pratiques et d'application générale, présentent un grand intérêt. Pour (EAS, EAS), les efficacités empiriques de ces estimateurs sont indiquées

dans les colonnes 4 et 5 du tableau 7.1. Le signe négatif indique une perte d'efficacité par rapport à l'estimateur de HT. L'efficacité de $\hat{t}_y^{\mathrm{GR}}$ est une approximation proche de l'efficacité de $\hat{t}_y^{O}$, sauf pour les quatre configurations indiquées par $r^2_{x_1,x_2} = 0$, $r^2 = 0{,}25$; 0,75 et $n_2 = 2\,000$; 1 500; plus particulièrement, quand $n_2 = 2\,000$ l'estimateur $\hat{t}_y^{\mathrm{GR}}$ est légèrement moins efficace que l'estimateur $\tilde{t}_y$. En revanche, l'estimateur $\hat{t}_y^{\mathrm{ES}}$ est moins efficace que l'estimateur $\tilde{t}_y$ dans six configurations, quand $r^2_{x_1,x_2} = 0$, $r^2 =$ 0,25; 0,75 et $n_2 = 2\,000$; 1 500; 1 000 et beaucoup moins efficace quand $n_2 = 2\,000$; 1 500. Le point important dans les colonnes 4 et 5 est que l'estimateur $\hat{t}_y^{\mathrm{GR}}$ est nettement plus efficace que l'estimateur $\hat{t}_y^{\mathrm{ES}}$ dans toutes les configurations, et encore plus pour les valeurs les plus élevées de n_2 et les valeurs les plus élevées de (r^2_{y,x_2}, r^2); cela indique que $\hat{t}_y^{\mathrm{GR}}$ est plus efficace lorsqu'il s'agit d'utiliser l'information tirée du complément de s_2 et d'exploiter les corrélations élevées de y avec x_1 et x_2. Étant donné que l'efficacité de l'estimateur $\hat{t}_y^{\mathrm{HS}}$ était pratiquement identique à celle de $\hat{t}_y^{\mathrm{ES}}$ dans les trois plans, elle n'est pas indiquée dans le tableau 7.1. Pour (EASS, EAS), les efficacités empiriques des estimateurs par calage $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ sont indiquées dans les colonnes 8 et 9 du tableau 7.1. Il convient de mentionner que les corrélations à l'intérieur des strates sont beaucoup plus faibles que les corrélations pour l'ensemble de la population (voir le tableau 7.1). De plus, l'estimateur de HT $\tilde{t}_y$ est très efficace en raison de la stratification, surtout pour les plus grandes valeurs de n_2. L'estimateur $\hat{t}_y^{\mathrm{GR}}$ est moins efficace que l'estimateur $\tilde{t}_y$ dans 3 des 24 configurations (qui comprennent $n_2 = 2\,000$), alors que pour les autres, son efficacité augmente considérablement à mesure que n_2 diminue, se rapprochant de l'efficacité de $\hat{t}_y^{O}$. L'estimateur $\hat{t}_y^{\mathrm{ES}}$ est moins efficace que l'estimateur $\tilde{t}_y$ dans 12 configurations. L'estimateur $\hat{t}_y^{\mathrm{GR}}$ est beaucoup plus efficace que l'estimateur $\hat{t}_y^{\mathrm{ES}}$ dans toutes les configurations, et encore plus pour les valeurs les plus élevées de n_2 et à mesure que nous passons de $r^2 = 0{,}25$ à $r^2 = 0{,}75$, et considérablement plus que dans le plan d'échantillonnage (EAS, EAS).

Tableau 7.1
Pourcentage d'efficacité de $\hat{t}_y^{B}$, $\hat{t}_y^{O}$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$

	(EAS, EAS)				(EASS, EAS)				(EAS, ESPPT)	
n_2	$\hat{t}_y^{B}$	$\hat{t}_y^{O}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{B}$	$\hat{t}_y^{O}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$
$\sigma^2_\eta = 292{,}41$; $r^2_{x_1,x_2} = 0{,}00$; $r^2_{y,x_1} = 0{,}04$; $r^2_{y,x_2} = 0{,}21$; $r^2 = 0{,}25$										
2 000	11,54	10,09	-1,66	-26,88	16,74	13,69	-23,49	-79,94	-5,51	-30,38
1 500	15,00	13,84	10,91	-13,61	20,01	17,80	7,22	-38,46	4,57	-20,69
1 000	18,41	17,68	17,79	-0,71	22,22	21,20	19,34	-11,76	9,69	-10,22
500	21,74	20,62	20,77	11,29	23,81	22,34	22,75	7,84	11,24	0,67
$\sigma^2_\eta = 32{,}15$; $r^2_{x_1,x_2} = 0{,}00$; $r^2_{y,x_1} = 0{,}13$; $r^2_{y,x_2} = 0{,}62$; $r^2 = 0{,}75$										
2 000	34,35	31,21	-4,23	-80,22	52,31	48,02	-66,06	-232,15	-21,68	-107,41
1 500	44,84	42,34	33,09	-41,49	61,53	59,02	24,66	-108,30	15,95	-74,74
1 000	55,10	53,80	53,83	-2,25	67,58	65,48	59,17	-30,84	36,49	-39,03
500	65,16	63,87	63,90	34,31	71,85	70,53	70,41	27,83	45,55	2,00
$\sigma^2_\epsilon = 12{,}11$; $\sigma^2_\eta = 632{,}52$; $r^2_{x_1,x_2} = 0{,}25$; $r^2_{y,x_1} = 0{,}12$; $r^2_{y,x_2} = 0{,}24$; $r^2 = 0{,}25$										
2 000	16,70	16,89	16,69	9,88	17,14	15,56	2,08	-23,18	10,24	3,03
1 500	18,85	19,02	19,52	13,57	20,26	19,45	16,46	-3,21	13,83	7,08
1 000	20,97	20,79	20,52	16,50	22,38	21,07	21,04	6,84	13,03	8,04
500	23,04	22,28	21,67	19,64	23,91	22,84	23,41	16,40	11,57	9,38

EAS = échantillonnage aléatoire simple; EASS = échantillonnage aléatoire simple stratifié;
ESPPT = échantillonnage systématique avec probabilité proportionnelle à la taille.

Tableau 7.1(suite)
Pourcentage d'efficacité de $\hat{t}_y^B$, $\hat{t}_y^O$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$

	(EAS, EAS)				(EASS, EAS)				(EAS, ESPPT)	
n_2	$\hat{t}_y^B$	$\hat{t}_y^O$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^B$	$\hat{t}_y^O$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$
σ_ϵ^2 =12,11; σ_η^2 = 70,68; $r^2_{x_1,x_2}$ = 0,25; r^2_{y,x_1} = 0,36; r^2_{y,x_2} = 0,71; r^2 = 0,75										
2 000	49,70	48,33	46,89	25,33	53,11	50,78	20,11	-38,15	35,49	11,64
1 500	56,23	55,48	56,46	38,08	61,86	60,63	53,81	8,36	46,33	22,98
1 000	62,63	62,20	61,35	48,69	67,71	66,38	66,10	34,90	47,91	30,19
500	68,90	68,09	65,58	59,65	71,90	70,99	71,00	55,27	47,68	38,93
σ_ϵ^2 =1,33; σ_η^2 = 340,40; $r^2_{x_1,x_2}$ = 0,75; r^2_{y,x_1} = 0,22; r^2_{y,x_2} = 0,24; r^2 = 0,25										
2 000	23,36	23,67	23,01	10,81	18,09	15,47	-6,07	-46,54	16,62	3,38
1 500	23,78	23,63	24,19	13,85	20,83	19,60	14,52	-17,17	19,77	7,95
1 000	24,20	23,83	23,15	16,97	22,68	21,67	21,58	0,86	17,04	10,02
500	24,61	23,54	22,24	19,92	24,01	22,39	22,98	13,24	14,52	11,77
σ_ϵ^2 =1,33; σ_η^2 = 37,82; $r^2_{x_1,x_2}$ = 0,75; r^2_{y,x_1} = 0,67; r^2_{y,x_2} = 0,72; r^2 = 0,75										
2 000	69,84	67,98	65,10	26,96	60,26	56,75	32,34	-27,50	56,65	13,24
1 500	71,17	69,57	70,70	38,91	66,25	64,49	59,73	14,39	65,49	26,11
1 000	72,47	71,26	69,17	49,62	70,17	68,80	68,69	40,44	61,00	35,28
500	73,74	72,19	67,58	60,66	72,94	71,12	71,10	56,90	54,67	44,54

EAS = échantillonnage aléatoire simple; EASS = échantillonnage aléatoire simple stratifié;
ESPPT = échantillonnage systématique avec probabilité proportionnelle à la taille.

Pour (EAS, ESPPT), les efficacités empiriques des estimateurs par calage $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ sont présentées dans les colonnes 10 et 11 du tableau 7.1. La tendance de ces efficacités est très semblable à celle du plan (EAS, EAS). C'est particulièrement le cas pour l'efficacité de $\hat{t}_y^{\mathrm{GR}}$ par rapport à $\hat{t}_y^{\mathrm{ES}}$, qui n'est pas présentée dans le tableau 7.1, mais qui peut se calculer facilement au moyen des efficacités de $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$ indiquées. L'estimateur de HT $\tilde{t}_y$ lui-même est plus efficace dans ce plan de sondage à deux phases, ce qui explique que l'efficacité des deux estimateurs par calage $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$ est quelque peu inférieure à celle observée dans les plans (EAS, EAS) et (EASS, EAS).

Toute l'étude par simulations a été répétée dans la population simulée pour le vecteur (y, x_1, x_2) généré à partir d'une loi log-normale trivariée comportant les structures de corrélation spécifiées. Pour les trois plans (EAS, EAS), (EAS, ESPPT) et (EASS, EAS), les résultats (non présentés ici) étaient très semblables à ceux basés sur le modèle linéaire pour y utilisé précédemment.

Il est intéressant d'examiner la configuration des variables auxiliaires dans laquelle la variable scalaire x_1 est augmentée pour s'établir à $(1, x_1)$, avec des totaux connus (N, t_{x_1}). Ensuite, dans le plan de sondage (EAS, EAS), dans lequel la construction du MELSB $\hat{t}_y^B$ et de l'estimateur optimal $\hat{t}_y^O$ est réalisable, l'utilisation de la configuration complète $(1, x_1, x_2)$ dans le calage donne le même $\hat{t}_y^B$ et pratiquement le même $\hat{t}_y^B$ que l'utilisation de (x_1, x_2). Cela convertirait également l'estimateur par la régression $\hat{t}_y^{\mathrm{GR}}$ en $\hat{t}_y^O$ (quand nous utilisons le même ajustement $1/\pi_{2k}$ de $\mathbf{Y}_1$ comme dans $\hat{t}_y^O$), et l'estimateur par la régression $\hat{t}_y^{\mathrm{ES}}$ en estimateur pseudo-optimal $\hat{t}_y^{\mathrm{PSO}}$ (défini à la section 6). Ces propriétés sont calculées à partir d'une théorie connue, voir par exemple Merkouris (2004, 2015), plus directement pour $\hat{t}_y^{\mathrm{ES}}$ et le second terme de la régression de $\hat{t}_y^{\mathrm{GR}}$ et l'estimateur optimal $\hat{t}_y^O$, indépendamment de toute relation fonctionnelle précise de y avec $(1, x_1, x_2)$. Les trois estimateurs fondés sur l'échantillon

présenteraient alors un comportement empirique pratiquement identique. Cela est conforme à la proposition 1, qui donne la condition (satisfaite par des plans spécifiques, notamment (EAS, EAS)) selon laquelle l'estimateur par la régression pseudo-optimal $\hat{t}_y^{\mathrm{PSO}}$ est asymptotiquement équivalent à l'estimateur optimal proposé $\hat{t}_y^O$. Des calculs expérimentaux ont confirmé cette équivalence. Dans le plan de sondage (EASS, EAS), l'utilisation de $(1, x_1, x_2)$ donne les mêmes $\hat{t}_y^B$ et $\hat{t}_y^O$ que l'utilisation de (x_1, x_2), et convertit les estimateurs $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ respectivement en estimateurs $\hat{t}_y^O$ et $\hat{t}_y^{\mathrm{PSO}}$. Cependant, selon la proposition 1, l'équivalence des deux derniers estimateurs, et par conséquent de $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$, ne se vérifie pas dans ce plan de sondage.

8. Discussion

La méthode décrite d'estimation optimale et par la régression pour un échantillonnage à deux phases comprend un calage en une étape des poids des échantillons combinés de la première et de la deuxième phase. il est ainsi possible d'obtenir, au moyen d'un seul ensemble de poids calés qui comprend toute l'information disponible des deux phases, une estimation grandement améliorée de la valeur totale d'une variable cible, comme le démontre l'étude par simulations. Ces poids pourraient servir à calculer d'autres statistiques pondérées, y compris les moyennes, les ratios, les quantiles et les coefficients de régression. Le cadre de la méthode est suffisamment général pour englober des plans complexes à plusieurs degrés et une stratification différente aux deux phases, ainsi que différents types de variables auxiliaires connues au niveau de la population ou de l'échantillon; Estevao et Särndal (2002) définissent ainsi 10 cas différents d'information auxiliaire. De plus, la méthode peut être étendue aux plans de sondage à phases multiples au moyen d'une configuration de calage appropriée.

L'estimation d'un total pour tout domaine (sous-population) d'intérêt $U_d \subset U$ peut facilement être réalisée au moyen des poids calés et en additionnant les valeurs d'échantillon pondérées de la variable d'intérêt sur U_d. Pour que l'estimateur de domaine qui en résulte soit l'estimateur linéaire optimal, il faut combiner linéairement les estimations pour un domaine de $\mathbf{t}_{\mathbf{y}}$, $\mathbf{t}_{\mathbf{x}_1}$ et $\mathbf{t}_{\mathbf{x}_2}$ en effectuant un calage optimal au niveau du domaine avec les totaux de calage de domaine et avec la modification appropriée de la matrice $\boldsymbol{\mathcal{X}}$. Plusieurs options de calage, concernant l'utilisation de l'information auxiliaire disponible au niveau de la population, du domaine et de l'échantillon à deux phases, peuvent être prises en compte pour déterminer l'estimation la plus efficace des totaux de domaine dans une application donnée. Les travaux de Merkouris (2010) à ce propos seraient utiles dans ce contexte.

Les variances approximatives estimées de l'estimateur optimal à deux phases et de l'estimateur par la régression à deux phases, fondées sur la linéarisation en séries de Taylor, ont été données aux sections 4.1 et 5, respectivement. Pour l'estimateur par la régression à deux phases, les méthodes de rééchantillonnage d'estimation de la variance, comme la méthode du jackknife ou la méthode bootstrap, pourraient être appliquées autrement ou elles seraient la seule possibilité quand les probabilités d'inclusion de la première ou de la deuxième phase ne sont pas connues. Une littérature abondante traite de ces méthodes de

rééchantillonnage appliquées aux estimateurs par la régression existants dans un échantillonnage à deux phases. À ces fins, la fonction de calage en une seule étape de la méthode d'estimation par la régression proposée peut se révéler utile. Cette question, qui dépasse la portée du présent article, mériterait d'être approfondie.

Remerciements

L'auteur remercie le rédacteur en chef, le rédacteur associé et les deux arbitres pour leurs commentaires et suggestions qui ont mené à une amélioration importante de l'article.

Annexe

Démonstration du lemme 1

La matrice symétrique $\text{Var}(\mathbf{w}_U^*)$ a la forme de (3.8), mais elle a $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U)$ comme bloc hors diagonale. Le kl^{e} élément de la matrice $\text{Var}(\mathbf{w}_{1U})$ est :

$$\text{Cov}(w_{1U_k}, w_{1U_l}) = [E(I_{1k}\, I_{1l}) - E(I_{1k})\, E(I_{1l})] / \pi_{1k}\, \pi_{1l} = (\pi_{1kl} - \pi_{1k}\pi_{1l}) / \pi_{1k}\, \pi_{1l}\,.$$

Le kl^{e} élément de la matrice $\text{Var}(\mathbf{w}_U)$ est :

$$\begin{aligned}\text{Cov}(w_{U_k}, w_{U_l}) &= [E(I_{1k}\, I_{2k}\, I_{1l}\, I_{2l}) - E(I_{1k}\, I_{2k})\, E(I_{1l}\, I_{2l})] / \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l} \\ &= [E_1(I_{1k}\, I_{1l}\, E_2(I_{2k}\, I_{2l})) - E_1(I_{1k}\, E_2(I_{2k}))\, E_1(I_{1l}\, E_2(I_{2l}))] / \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l} \\ &= [\pi_{1kl}\, \pi_{2kl} - \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l}] / \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l}\,,\end{aligned}$$

où E_1 et E_2 désignent l'espérance pour la première et la deuxième phase de l'échantillonnage, respectivement. À partir d'arguments similaires, il s'ensuit que le kl^{e} élément de la matrice $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U)$ est :

$$\text{Cov}(w_{1U_k}, w_{U_l}) = [E(I_{1k}\, I_{1l}\, I_{2l}) - E(I_{1k})\, E(I_{1l}\, I_{2l})] / \pi_{1k}\, \pi_{1l}\, \pi_{2l} = (\pi_{1kl} - \pi_{1k}\, \pi_{1l}) / \pi_{1k}\, \pi_{1l}\,.$$

Cela montre que $\text{Cov}(w_{1U_k}, w_{U_l}) = \text{Cov}(w_{1U_k}, w_{1U_l})$ et donc que $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U) = \text{Var}(\mathbf{w}_{1U})$, ce qui termine la démonstration.

Démonstration du théorème 1

La matrice $\boldsymbol{\Delta} = \text{Var}(\mathbf{w}_U^*)$ est non singulière si et seulement si $\text{Var}(\mathbf{w}_U) - \text{Var}(\mathbf{w}_{1U})$ est non singulier. Cela découle d'un résultat général sur les inverses des matrices partitionnées (voir Harville, 2008, page 98). Toutefois, $\text{Var}(\mathbf{w}_U) - \text{Var}(\mathbf{w}_{1U}) = \text{Var}(\mathbf{w}_{1U} - \mathbf{w}_U)$, parce que $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U) = \text{Var}(\mathbf{w}_{1U})$, et par conséquent $\text{Var}(\mathbf{w}_U) - \text{Var}(\mathbf{w}_{1U})$ est non singulier, étant une matrice variance-covariance. Alors, pour trouver le vecteur $\mathbf{c}_U^*$ qui minimise $(\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*)$ soumis aux contraintes $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, nous considérons la fonction $\mathbf{F} = (\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*) - \boldsymbol{\lambda}' \boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^*$ où $\boldsymbol{\lambda}$ est un vecteur des multiplicateurs de Lagrange. Nous obtenons alors le système d'équations :

$$\frac{\partial \mathbf{F}}{\partial \mathbf{c}_U^*} = 2\boldsymbol{\Delta}^{-1}(\mathbf{c}_U^* - \mathbf{w}_U^*) - \boldsymbol{\mathcal{X}}_U \boldsymbol{\lambda} = \mathbf{0}$$
$$\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* - \mathbf{t}_{\boldsymbol{\mathcal{X}}} = \mathbf{0}.$$

En multipliant la première équation par $\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta}$, en utilisant $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$ et en résolvant $\boldsymbol{\lambda}$, nous obtenons $\boldsymbol{\lambda} = 2(\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}\left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^*\right)$. Si nous l'insérons dans la première équation et que nous résolvons $\mathbf{c}_U^*$, nous obtenons $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta}\boldsymbol{\mathcal{X}}_U(\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}\left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^*\right)$.

Démonstration de la proposition 1

De toute évidence, les coefficients de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ dans (3.14) et (6.2) sont identiques si $\boldsymbol{\Delta}_1 = \delta \boldsymbol{\Delta}_2$. Ensuite, en utilisant la partition $\mathbf{X}_U = (\mathbf{X}_{1U}, \mathbf{X}_{2U})$, nous exprimons le coefficient de $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ dans (6.2) comme suit. Nous obtenons d'abord :

$$\begin{aligned}\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} &= \mathbf{Y}_U' \boldsymbol{\Delta}_2 (\mathbf{X}_{1U}, \mathbf{X}_{2U}) \begin{pmatrix} \mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} & \mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \\ \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} & \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \\ \mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \end{pmatrix} \\ &= \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} \left[A_{11}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{12}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})\right] \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \left[A_{21}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{22}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})\right],\end{aligned}$$

où A_{11}, A_{12}, A_{21}, A_{22} sont calculés algébriquement à partir des inverses des matrices partitionnées. En particulier,

$$A_{11} = (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} - A_{12}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})(\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1}$$

et $A_{21} = -A_{22}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U}) \times (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1}$. Alors,

$$\begin{aligned} A_{11}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{12}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) &= (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} + A_{12}\mathbf{B} \\ A_{21}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{22}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) &= A_{22}\mathbf{B}, \end{aligned}$$

où

$$\mathbf{B} = \mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} - \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}).$$

Il est alors facile de vérifier que si $\boldsymbol{\Delta}_1 = \delta \boldsymbol{\Delta}_2$, nous avons $(\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} = \delta \mathbf{I}$ et $\mathbf{B} = \mathbf{0}$. Il s'ensuit que $\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} = \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}$, et que les coefficients de $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ dans (3.14) et (6.2) sont aussi identiques.

Bibliographie

Australian Bureau of Statistics (2004). Estimation for the household income and expenditure survey. Document de recherche 1352.0.55.063.

Beaumont, J.-F., Beliveau, A. et Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3, 524-542.

Brick, J.M., et Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752.

Chen, S., et Kim, J.K. (2014). Two-phase sampling experiment for propensity score estimation in self-selected samples. *The Annals of Applied Statistics*, 3, 1492-1515.

Chipperfield, J.O., et Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 227-244.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., et Särndal, C.-E. (2009). Un nouveau visage pour l'échantillonnage à deux phases avec estimateurs par calage. *Techniques d'enquête*, 35, 1, 3-16. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10880-fra.pdf.

Fuller, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 2, 177-190. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1990002/article/14537-fra.pdf.

Fuller, W.A. (1998). Replication variance estimation for two-phase sampling. *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A., et Isaki, C.T. (1981). Survey design under superpopulation models. Dans *Current Topics in Survey Sampling*, (Éds., D. Krewski, J.N.K. Rao et R. Platek), New York: Academic Press, 199-226.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, 169, 439-457.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Haziza, D., Mecatti, F. et Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *METRON-International Journal of Statistics*, vol LXVI, 91-108.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 2, 157-169. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001002/article/6091-fra.pdf.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 1, 11-20. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1998001/article/3905-fra.pdf.

Hidiroglou, M.A., Rao, J.N.K. et Haziza, D. (2008). Variance estimation in two-phase sampling. *Australian and New Zealand Journal of Statistics*, 51, 127-141.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Ser. B*, 42, 221-226.

Kim, J.K., et Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kim, J.K., et Yu, C.L. (2011). Estimation de la variance par répliques sous échantillonnage à deux phases. *Techniques d'enquête*, 37, 1, 73-81. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11448-fra.pdf.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 311-320.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for more efficient small domain estimation. *Journal of the Royal Statistical Society, Ser. B*, 72, 27-48.

Merkouris, T. (2015). Une méthode d'estimation efficace pour l'échantillonnage matriciel. *Techniques d'enquête*, 41, 1, 249-276. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14174-fra.pdf.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model-Assisted Survey Sampling*. New York: Springer.

Turmelle, C., et Beaucage, Y. (2013). Le Programme intégré de la statistique des entreprises : utilisation d'un plan de sondage à deux phases pour produire des estimations fiables. *Recueil : Symposium 2013, Produire des estimations fiables à partir de bases imparfaites.*

Wolter, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Wu, C., et Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 2, 119-131.