Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Une approche d'inférence fondée sur la vraisemblance composite pondérée pour des modèles à deux niveaux issus de données d'enquête

par J.N.K. Rao, François Verret et Mike A. Hidiroglou

Janvier 2014

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 sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

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 « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Note de reconnaissance

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

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 291-311
Statistique Canada, N° 12-001-X au catalogue

Une approche d'inférence fondée sur la vraisemblance composite pondérée pour des modèles à deux niveaux issus de données d'enquête

J.N.K. Rao, François Verret et Mike A. Hidiroglou[1]

Résumé

Les modèles multiniveaux sont d'usage très répandu pour analyser les données d'enquête en faisant concorder la hiérarchie du plan de sondage avec la hiérarchie du modèle. Nous proposons une approche unifiée, basée sur une log-vraisemblance composite pondérée par les poids de sondage pour des modèles à deux niveaux, qui mène à des estimateurs des paramètres du modèle convergents sous le plan et sous le modèle, même si les tailles d'échantillon dans les grappes sont petites, à condition que le nombre de grappes échantillonnées soit grand. Cette méthode permet de traiter les modèles à deux niveaux linéaires ainsi que linéaires généralisés et requiert les probabilités d'inclusion de niveau 2 et de niveau 1, ainsi que les probabilités d'inclusion conjointe de niveau 1, où le niveau 2 représente une grappe et le niveau 1, un élément dans une grappe. Nous présentons aussi les résultats d'une étude en simulation qui donnent la preuve que la méthode proposée est supérieure aux méthodes existantes sous échantillonnage informatif.

Mots-clés : Vraisemblance composite; probabilités d'inclusion; échantillonnage informatif; modèles multiniveaux.

1 Introduction

Les données recueillies dans le cadre d'enquêtes socioéconomiques, sur la santé et autres à grande échelle sont utilisées abondamment à des fins analytiques, comme l'inférence sur les paramètres de modèles de régression linéaire et de régression logistique linéaire de populations. Ne pas tenir compte des caractéristiques du plan de sondage (comme la stratification, la mise en grappes et les probabilités de sélection inégales) peut donner lieu à des inférences incorrectes sur les paramètres du modèle, à cause du biais de sélection dans l'échantillon causé par l'échantillonnage informatif. Il est tentant d'étendre les modèles en incluant parmi les variables auxiliaires toutes les variables du plan de sondage qui définissent le processus de sélection à divers niveaux, puis d'ignorer le plan de sondage et d'appliquer des méthodes classiques au modèle étendu. Les principales difficultés de cette approche sont les suivantes (Pfeffermann et Sverchkov 2003) : 1) l'analyste pourrait ne pas connaître toutes les variables du plan ou ne pas avoir accès à toutes ces variables; 2) l'utilisation d'un trop grand nombre de variables du plan de sondage peut causer des difficultés d'inférence à partir du modèle étendu; 3) le modèle étendu pourrait ne plus présenter d'intérêt scientifique pour l'analyste. Par ailleurs, l'approche fondée sur le plan de sondage peut fournir des inférences par échantillonnage répété asymptotiquement valides sans modifier le modèle de l'analyste. Une approche unifiée, fondée sur des équations d'estimation pondérées par les poids de sondage conduit à des estimateurs convergents sous le plan des paramètres de « recensement », c'est-à-dire de population finie, qui à leur tour permettent d'estimer les paramètres associés du modèle. En outre, les méthodes de rééchantillonnage, comme le jackknife et le bootstrap pour données d'enquête, peuvent fournir des estimateurs de variance valides et des inférences connexes sur les paramètres de recensement. Les mêmes méthodes pourraient aussi être applicables à l'inférence sur les paramètres du modèle, dans de nombreux

1. J.N.K. Rao, École de mathématiques et de statistique, Université Carleton, Ottawa (Ontario), Canada, K1S 5B6. Courriel : jrao@math.carleton.ca; François Verret, Statistique Canada, 15 B, immeuble R.-H.-Coats, Ottawa (Ontario), Canada, K1A 0T6. Courriel : francois.verret@statcan.gc.ca; Mike A. Hidiroglou, Statistique Canada, 16 D, immeuble R.-H.-Coats, Ottawa (Ontario), Canada, K1A 0T6. Courriel : mike.hidiroglou@statcan.gc.ca.

cas d'enquêtes à grande échelle. Dans les autres cas, il est nécessaire d'estimer la variance sous le modèle des paramètres de recensement à partir de l'échantillon. L'estimateur de la variance totale est alors donné par la somme de cet estimateur et de l'estimateur de variance par rééchantillonnage. Beaumont et Charest (2010) ont étendu le bootstrap à l'estimation de la variance totale associée aux paramètres du modèle. Le lecteur est invité à consulter Rao et coll. (2010) pour un aperçu des méthodes d'inférence sur les paramètres de régression issus de données d'enquête complexes.

Dans le présent article, nous visons avant tout à faire des inférences fondées sur le plan de sondage sur les paramètres des composantes de la variance et sur les paramètres de régression de modèles multiniveaux en partant de données obtenues au moyen de plans d'échantillonnage à plusieurs degrés qui correspondent aux niveaux du modèle. Par exemple, dans une étude sur l'éducation menée auprès des élèves, les écoles (unités d'échantillonnage de premier degré) pourraient être sélectionnées avec probabilité proportionnelle à la taille de l'école et les élèves (unités d'échantillonnage de deuxième degré) pourraient être sélectionnés dans les écoles échantillonnées selon un plan d'échantillonnage aléatoire stratifié. De nouveau, ne pas tenir compte du plan de sondage et utiliser des méthodes classiques pour les modèles multiniveaux peut donner lieu à des inférences incorrectes en cas de biais de sélection dans l'échantillon. Dans l'approche fondée sur le plan de sondage, il est plus difficile d'estimer les paramètres des composantes de la variance du modèle que les paramètres de régression. Les travaux antérieurs sur les modèles multiniveaux pour données d'enquête sont résumés à la section 2. Notre objectif principal est de présenter une approche unifiée d'inférence pour des modèles multiniveaux provenant de données d'enquête, fondée sur une log-vraisemblance composite pondérée (section 4). La méthode proposée produit des inférences asymptotiquement valides sur les paramètres des composantes de la variance, même quand les tailles d'échantillon dans les grappes sont petites, à condition que le nombre de grappes échantillonnées soit grand, contrairement à certaines méthodes existantes résumées à la section 2. Les résultats d'une simulation limitée sont présentés à la section 5.

2 Modèles à deux niveaux : travaux antérieurs

2.1 Modèles à deux niveaux

Les modèles multiniveaux (ou modèles hiérarchiques) sont d'usage très répandu, notamment dans les domaines des sciences sociales, de l'éducation et de la santé, pour analyser les données d'enquête possédant une structure hiérarchique. Ici, nous nous concentrons sur les modèles à deux niveaux associés à l'échantillonnage à deux degrés de grappes (niveau 2) : un échantillon, s, d'unités de niveau 2, i, est sélectionné selon un plan spécifié, puis un échantillon, $s(i)$, d'éléments (ou unités de niveau 1), j, est sélectionné dans chacune des unités de niveau 2 échantillonnées i conformément à un autre plan spécifié. Nous supposons, en nous inspirant de la littérature sur les modèles multiniveaux pour données d'enquête, que le modèle concorde avec la hiérarchie du plan de sondage, comme dans l'exemple d'une enquête sur l'éducation réalisée auprès des élèves. Cependant, dans le cas de certaines enquêtes polyvalentes, la structure hiérarchique du plan de sondage pourrait être assez différente de la hiérarchie du modèle. Par exemple, l'Enquête longitudinale nationale auprès des enfants et des jeunes au Canada est réalisée selon un plan de sondage à plusieurs degrés où les degrés correspondent aux régions géographiques, aux

ménages dans une région et aux élèves dans un ménage, tandis qu'un modèle multiniveaux de l'éducation peut comprendre comme niveau les élèves, les classes, les écoles et les commissions scolaires (Rao et Roberts 1998). Puisque les grappes du plan de sondage recoupent les grappes du modèle pour ce genre d'enquête, il est difficile d'élaborer une méthode pondérée selon le plan de sondage appropriée d'inférence sur les paramètres du modèle qui permet de tenir compte de l'échantillonnage informatif des grappes et/ou des éléments dans les grappes échantillonnées. Sous échantillonnage informatif, le modèle supposé pour la population n'est pas nécessairement vérifié pour l'échantillon.

Soit N le nombre d'unités de niveau 2 dans la population et M_i, le nombre d'unités de niveau 1 dans l'unité i de niveau 2. Un modèle de superpopulation à deux niveaux est donné par

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} f\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right), \ \mathbf{v}_i \sim_{iid} f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right), i = 1,...,N \ ; j = 1,...,M_i, \tag{2.1}$$

où y_{ij} et $\mathbf{x}_{ij} = \left(x_{ij0},...,x_{ij,p-1}\right)^T$ sont la réponse et le vecteur de dimension p des valeurs des covariables associés à l'élément j dans la grappe i et $x_{ij0} = 1$, $\mathbf{v}_i$ désigne un effet aléatoire de niveau 2, et $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ désignent les paramètres associés aux deux degrés du modèle supposé. Ici $f\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right)$ et $f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right)$ sont les densités de probabilité spécifiées de y_{ij} sachant $\mathbf{x}_{ij}$ et $\mathbf{v}_i$, et de $\mathbf{v}_i$, respectivement. Notons, que, dans le modèle (2.1), les réponses y_{ij} d'une unité i donnée sont supposées être conditionnellement indépendantes sachant l'effet aléatoire $\mathbf{v}_i$, mais elles sont corrélées marginalement en raison de l'effet aléatoire $\mathbf{v}_i$ commun. La formulation du modèle (2.1) englobe à la fois les modèles à deux niveaux linéaires et les modèles à deux niveaux linéaires généralisés. Sous échantillonnage informatif des grappes et/ou des éléments dans les grappes échantillonnées, les méthodes classiques applicables aux modèles multiniveaux qui ne tiennent pas compte du plan de sondage et supposent que le modèle (2.1) est vérifié pour l'échantillon peuvent produire des estimateurs asymptotiquement biaisés des paramètres du modèle $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ (Pfeffermann et coll. 1998).

Cas particuliers

1) Un simple modèle de la moyenne à erreurs emboîtées souvent utilisé dans les études en simulation portant sur les modèles à deux niveaux est donné par

$$y_{ij} = \mu + v_i + e_{ij}, e_{ij} \sim_{iid} N\left(0, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right), \tag{2.2}$$

où $i = 1,...,N; j = 1,...,M_i$. Le modèle (2.2) peut être écrit sous la forme (2.1) comme

$$y_{ij} \mid v_i \sim_{ind} N\left(\mu + v_i, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right), \boldsymbol{\theta}_1 = \left(\mu, \sigma_e^2\right), \theta_2 = \sigma_v^2.$$

Marginalement, $y_{ij} \sim N\left(\mu, \sigma_v^2 + \sigma_e^2\right)$ mais y_{ij} et y_{ik} $(j \neq k)$ sont corrélées : $\text{corr}\left(y_{ij}, y_{ik}\right) = \rho = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2\right), \ j \neq k.$

2) Un modèle linéaire à deux niveaux, souvent utilisé en pratique, est donné par

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_i + e_{ij}, i = 1,...,N; j = 1,...,M_i, \tag{2.3}$$

où $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim_{iid} N_p(\mathbf{0}, \boldsymbol{\Sigma}_v), i = 1,...,N$ et $e_{ij} \sim_{iid} N(0, \sigma_e^2)$. Ce modèle peut également être exprimé sous la forme (2.1) comme

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2), \mathbf{v}_i \sim_{iid} N_p(0, \boldsymbol{\Sigma}_v) \tag{2.4}$$

où $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^T, \sigma_e^2)^T$ et $\boldsymbol{\theta}_2$ est le vecteur des $p(p+1)/2$ éléments distincts de $\boldsymbol{\Sigma}_v$. Marginalement, $y_{ij} \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta}, \mathbf{x}_{ij}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ij} + \sigma_e^2)$, mais y_{ij} et y_{ik} $(j \neq k)$ sont corrélées en raison de l'effet aléatoire commun $\mathbf{v}_i$. Cependant, dans le cas d'un modèle linéaire généralisé à deux niveaux, la loi marginale de y_{ij} ne donne généralement pas une expression analytique : par exemple, dans le cas d'un modèle linéaire logistique à deux niveaux pour réponses binaires.

2.2 Estimation ponctuelle

La log-vraisemblance de « recensement » ou de population sous le modèle à deux niveaux supposé (2.1) est donnée par

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log L_i(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N} l_i(\boldsymbol{\theta}) = l(\boldsymbol{\theta}), \tag{2.5}$$

où $\boldsymbol{\theta}$ est le vecteur comprenant les éléments $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$, et

$$L_i(\boldsymbol{\theta}) = \int \exp\left[\sum_{j=1}^{M_i} \log f\left(y_{ij} \middle| \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right)\right] f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right) d\mathbf{v}_i \tag{2.6}$$

voir Asparouhov (2006) et Rabe-Hesketh et Skrondal (2006). La fonction de score de recensement $\mathbf{U}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ satisfait $E_m\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$, où E_m désigne l'espérance sous le modèle. Le paramètre de recensement $\boldsymbol{\theta}_N$ est défini comme la solution unique de $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$ et $\boldsymbol{\theta}_N$ est convergent sous le modèle pour $\boldsymbol{\theta}$, où $\boldsymbol{\theta}_N$ est le vecteur des éléments $\boldsymbol{\theta}_{1N}$ et $\boldsymbol{\theta}_{2N}$.

Soit l'échantillon constitué de n grappes avec m_i éléments provenant de la grappe échantillonnée i. Soit π_i et $\pi_{j|i}$ les probabilités d'inclusion de niveau 2 et de niveau 1, respectivement, associées à la grappe i et à l'élément j dans la grappe i. Alors, les pondérations de niveau 2 et de niveau 1 sont données par $w_i = \pi_i^{-1}$ et $w_{j|i} = \pi_{j|i}^{-1}$, respectivement. Asparouhov (2006) et Rabe-Hesketh et Skrondal (2006) ont proposé une pseudo log-vraisemblance d'échantillon pondérée obtenue en remplaçant $\sum_{j=1}^{M_i}(.)$ dans (2.6) par $\sum_{j \in s_i} w_{j|i}(.)$ et $\sum_{i=1}^{N}(.)$ dans (2.5) par $\sum_{i \in s} w_i(.)$, où s désigne l'échantillon de grappes et $s(i)$ désigne l'échantillon d'éléments dans les grappes $i \in s$. Elle est donnée par

$$\tilde{l}_w(\boldsymbol{\theta}) = \sum_{i \in s} w_i \tilde{l}_{wi}(\boldsymbol{\theta}) \tag{2.7}$$

où $\tilde{l}_{wi}(\boldsymbol{\theta}) = \log \tilde{L}_{wi}(\boldsymbol{\theta})$ et

$$\tilde{L}_{wi}(\boldsymbol{\theta}) = \int \exp\left[\sum_{j \in s(i)} w_{j|i} \log f(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1) \right] f(\mathbf{v}_i \mid \boldsymbol{\theta}_2) d\mathbf{v}_i . \quad (2.8)$$

En maximisant la pseudo log-vraisemblance $\tilde{l}_w(\boldsymbol{\theta})$ donnée par (2.7), nous obtenons un estimateur du pseudo maximum de vraisemblance (PMV) $\tilde{\boldsymbol{\theta}}_w$. Les calculs sont exposés en détail dans Asparouhov (2006) et dans Rabe-Hesketh et Skrondal (2006). Dans le cas particulier des modèles linéaires à deux niveaux, Pfeffermann et coll. (1998) ont utilisé une méthode par les moindres carrés généralisés itérative proposée par Goldstein (1986). Notons que nous avons besoin des pondérations de niveau 1 et de niveau 2 pour calculer $\tilde{\boldsymbol{\theta}}_w$, contrairement au cas des modèles marginaux qui nécessitent seulement les pondérations non conditionnelles des éléments $w_{ij} = w_i w_{j|i}$.

La convergence sous le plan de sondage de l'estimateur PMV $\tilde{\boldsymbol{\theta}}_{2w}$ du paramètre de recensement $\boldsymbol{\theta}_{2N}$ ou la convergence sous le plan et sous le modèle de $\tilde{\boldsymbol{\theta}}_{2w}$ en tant qu'estimateur du paramètre du modèle $\boldsymbol{\theta}_2$ requiert que le nombre de grappes échantillonnées, n, ainsi que la taille d'échantillon dans les grappes, m_i, tendent vers l'infini, même dans le cas linéaire. En outre, le biais relatif des estimateurs sera important si les tailles d'échantillon m_i sont petites. Pour remédier à ce problème, plusieurs méthodes de rajustement des pondérations ont été proposées dans la littérature. En particulier, un facteur de mise à l'échelle k_{1i} est appliqué aux pondérations de niveau 1 $w_{j|i}$ dans (2.8) avant de maximiser la pseudo log-vraisemblance (2.7). Nous ne considérons ici que deux méthodes de rajustement des pondérations, désignées A et A1 (Asparouhov 2006). La méthode A utilise

$$k_{1i} = m_i \Big/ \sum_{j \in s(i)} w_{j|i} \quad (2.9)$$

Dans la méthode A1, k_{1i} est le même que dans la méthode A, mais les pondérations de niveau 2 w_i sont également rajustées au moyen du facteur $k_{2i} = 1/k_{1i}$ pour compenser le rajustement des pondérations de niveau 1. Asparouhov (2006) a mentionné l'utilisation d'un algorithme EM accéléré pour calculer l'estimateur PMV $\tilde{\boldsymbol{\theta}}_w$ avec Mplus 3 : www.statmodel.com : Muthén et Muthén, 1998-2005.

2.3 Estimation de la variance

En ce qui concerne l'estimation de la variance, Asparouhov (2006) a proposé un estimateur de variance « sandwich » par linéarisation de Taylor de $\tilde{\boldsymbol{\theta}}_w$, qui est donné par

$$v_L\left(\tilde{\boldsymbol{\theta}}_w\right) = \left(\tilde{\mathbf{I}}''_w\right)^{-1} \left[\sum_{i \in s} \left(k_{2i} w_i\right)^2 \tilde{\mathbf{I}}'_{wi} \left(\tilde{\mathbf{I}}'_{wi}\right)^T \right] \left(\tilde{\mathbf{I}}''_w\right)^{-1}, \quad (2.10)$$

où $\tilde{\mathbf{I}}'_w$ et $\tilde{\mathbf{I}}''_w$ désignent, respectivement, le vecteur des dérivées premières et la matrice des dérivées secondes de $\tilde{l}_w(\boldsymbol{\theta})$ évaluées à $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_w$, et $\tilde{\mathbf{I}}'_{wi}$ est la dérivée première de $\tilde{l}_{wi}(\boldsymbol{\theta})$ évaluée à $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_w$. Si la fraction d'échantillonnage de niveau 2 est faible, alors $v_L\left(\tilde{\boldsymbol{\theta}}_w\right)$ suit bien la variance de $\tilde{\boldsymbol{\theta}}_w$, mais non l'EQM de $\tilde{\boldsymbol{\theta}}_w$ si le biais relatif de $\tilde{\boldsymbol{\theta}}_w$ est grand.

Kovacevic et coll. (2006) ont étudié les estimateurs bootstrap de la variance de $\tilde{\boldsymbol{\theta}}_w$. Ils ont considéré deux options. L'option 1 consiste à utiliser les poids bootstrap de niveau 2 $w_i(b)$ basés sur la méthode de

Rao, Wu et Yue (1992) et à ne pas modifier les poids de niveau 1, c'est-à-dire $w_{j|i}(b) = w_{j|i}$, où $b = 1,...,B$ désigne les B échantillons bootstrap. L'option 2 consiste à appliquer la méthode du bootstrap de Rao, Wu et Yue (1992) au niveau 1 ainsi qu'au niveau 2, et à rajuster les poids bootstrap de niveau 1. En remplaçant les poids w_i et $w_{j|i}$ par $w_i(b)$ et $w_{j|i}(b)$ dans (2.7) et (2.8), on obtient les estimateurs bootstrap PMV $\tilde{\boldsymbol{\theta}}_w(b)$, $b = 1,...,B$ et l'estimateur bootstrap de la variance est donné par

$$v_{Boot}\left(\tilde{\boldsymbol{\theta}}_w\right) = \frac{1}{B}\sum_{b=1}^{B}\left[\tilde{\boldsymbol{\theta}}_w(b) - \tilde{\boldsymbol{\theta}}_w\right]\left[\tilde{\boldsymbol{\theta}}_w(b) - \tilde{\boldsymbol{\theta}}_w\right]^T. \tag{2.11}$$

Une étude en simulation de (2.11), fondée sur le simple modèle de la moyenne (2.2), a montré que l'option 1 peut donner lieu à une sous-estimation de la variance de $\tilde{\sigma}_{ew}^2$. L'option 2 a donné de meilleurs résultats que l'option 1. Grilli et Pratesi (2004) ont étudié une autre méthode bootstrap pour l'estimation de la variance.

3 Équations d'estimation pondérées par les poids de sondage

Aux sections 3 et 4, nous étudions les méthodes d'établissement des équations d'estimation pondérées par les poids de sondage pour les paramètres des modèles multiniveaux qui conduisent à des estimateurs convergents sous le plan et sous le modèle, même lorsque les tailles d'échantillon dans les grappes sont petites. Les méthodes proposées dépendent uniquement des probabilités d'inclusion d'ordre un π_i et $\pi_{j|i}$, et des probabilités d'inclusion conjointe $\pi_{jk|i}$ dans les grappes. À la section 3, nous présentons une approche simple, fondée sur les moments, des équations d'estimation pondérées, qui est applicable aux modèles de régression linéaires à erreurs emboîtées. À la section 4, nous proposons une méthode unifiée, fondée sur les log-vraisemblances composites pondérées. Cette méthode permet de traiter les modèles multiniveaux linéaires ainsi que linéaires généralisés, contrairement à la méthode fondée sur les moments, et elle aboutit à des estimateurs convergents sous le plan et sous le modèle. Elle ne dépend, elle aussi, que de π_i, $\pi_{j|i}$ et $\pi_{jk|i}$.

3.1 Estimation ponctuelle

Nous commençons par illustrer l'approche des équations d'estimation pondérées, en utilisant le simple modèle de la moyenne (2.2). Ici, nous voulons estimer $\boldsymbol{\theta} = \left(\mu, \sigma_v^2, \sigma_e^2\right)^T$ en partant d'un plan d'échantillonnage en grappes à deux degrés qui concorde avec la hiérarchie du modèle. Nous avons choisi pour cela les trois fonctions d'estimation (FE) suivantes :

$$u_1\left(y_{ij}, \boldsymbol{\theta}\right) = y_{ij} - \mu, \tag{3.1}$$

$$u_2\left(y_{ij}, \boldsymbol{\theta}\right) = \left(y_{ij} - \mu\right)^2 - \left(\sigma_v^2 + \sigma_e^2\right) \tag{3.2}$$

$$u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left[\left(y_{ij} - \mu\right) - \left(y_{ik} - \mu\right)\right]^2 - 2\sigma_e^2 = z_{ijk}{}^2 - 2\sigma_e^2, j \neq k, \tag{3.3}$$

où $z_{ijk} = y_{ij} - y_{ik}$. Les équations d'estimation de recensement correspondantes sont données par

$$U_1(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} u_1\left(y_{ij}, \boldsymbol{\theta}\right) = 0, U_2(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} u_2\left(y_{ij}, \boldsymbol{\theta}\right) = 0 \tag{3.4}$$

$$U_3(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j<k=1}^{M_i} u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = 0\,. \tag{3.5}$$

Le paramètre de recensement résultant, $\tilde{\boldsymbol{\theta}}_N$, est convergent sous le modèle pour $\boldsymbol{\theta}$ parce que les espérances sous le modèle des trois fonctions d'estimation (3.1) à (3.3) sont nulles. Il découle de (3.4) et (3.5) que les équations d'estimation pondérées par les poids de sondage (EEP) sont données par

$$\hat{U}_{w1}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} u_1\left(y_{ij}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w1i}(\boldsymbol{\theta}) = 0 \tag{3.6}$$

$$\hat{U}_{w2}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} u_2\left(y_{ij}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w2i}(\boldsymbol{\theta}) = 0 \tag{3.7}$$

$$\hat{U}_{w3}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w3i}(\boldsymbol{\theta}) = 0\,, \tag{3.8}$$

où $w_{jk|i} = \pi_{jk|i}^{-1}$. L'estimateur EEP, $\hat{\boldsymbol{\theta}}_w$, est obtenu en résolvant le système d'équations (3.6) à (3.8). Pour le modèle de la moyenne, nous obtenons les solutions explicites des EEP suivantes

$$\hat{\mu}_w = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij} y_{ij}\right) \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} \equiv \bar{y}_w \tag{3.9}$$

$$\hat{\sigma}_{vw}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij} - \bar{y}_w\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \tag{3.10}$$

$$\hat{\sigma}_{ew}^2 = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} z_{ijk}^2 \Big/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\right), \tag{3.11}$$

où $w_{ij} = w_i w_{j|i}$. Soulignons que la méthode des moments susmentionnés ne dépend pas de la loi de probabilité.

Nous notons que $\hat{U}_{wt}(\boldsymbol{\theta}), t = 1,2,3$ sont les fonctions d'estimation d'espérance nulle par rapport au plan de sondage et au modèle, c'est-à-dire $E_m E_p\left\{\hat{U}_{wt}(\boldsymbol{\theta})\right\} = 0$. En utilisant ce résultat, nous pouvons montrer que l'estimateur EEP $\hat{\boldsymbol{\theta}}_w = \left(\hat{\mu}_w, \hat{\sigma}_{vw}^2, \hat{\sigma}_{ew}^2\right)^T$ est convergent sous le plan et sous le modèle pour $\boldsymbol{\theta}$ à mesure que le nombre d'unités de niveau 2 dans l'échantillon, n, augmente, même si les tailles d'échantillon dans les grappes, m_i, sont petites. Cette propriété n'est pas nécessairement vérifiée pour les estimateurs présentés à la section 2. La méthode proposée nécessite toutefois les probabilités d'inclusion

conjointe dans les grappes $\pi_{jk|i}$. Ces probabilités sont obtenues facilement pour l'échantillonnage aléatoire simple ou stratifié dans les grappes, ou quand la fraction d'échantillonnage dans les grappes est faible. En outre, plusieurs bonnes approximations de $\pi_{jk|i}$ lorsque l'échantillonnage dans les grappes est effectué avec probabilités inégales sont disponibles, et ces approximations dépendent uniquement des probabilités d'inclusion marginales $\pi_{j|i}$ (Haziza, Mecatti et Rao 2008). L'estimateur EEP $\hat{\boldsymbol{\theta}}_w$ est également convergent sous le plan pour $\tilde{\boldsymbol{\theta}}_N$, en notant que $E_p\left\{\hat{U}_{wt}(\tilde{\boldsymbol{\theta}}_N)\right\} = 0$, $t = 1,2,3$.

Le choix des fonctions d'estimation (3.1) à (3.3) n'est pas forcément unique. Ainsi, nous pourrions remplacer l'équation précédente $u_2\left(y_{ij},\boldsymbol{\theta}\right)$ par $\tilde{u}_2\left(y_{ij},y_{ik},\boldsymbol{\theta}\right) = \left(y_{ij}-\mu\right)\left(y_{ik}-\mu\right) - \sigma_v^2$ dans (3.7) et garder (3.6) et (3.8). L'estimateur EEP résultant est également convergent sous le plan et sous le modèle pour $\boldsymbol{\theta}$ à mesure que le nombre d'unités de niveau 2 augmente. L'approche de la vraisemblance composite par paire pondérée décrite à la section 4 offre une méthode unifiée de génération des fonctions d'estimation.

Korn et Graubard (2003) ont utilisé pour le modèle de la moyenne une autre approche qui présente certaines similarités avec l'approche proposée. Sous leur approche, les « paramètres de recensement », S_e^2 et S_v^2, sont d'abord obtenus en supposant que le modèle est vérifié pour la population finie. Les estimateurs pondérés par les poids de sondage $\hat{S}_{ew}^2$ et $\hat{S}_{vw}^2$ des paramètres de recensement sont ensuite obtenus en supposant que M_i est connu pour les grappes échantillonnées. L'estimateur $\hat{S}_{ew}^2$ est donné par

$$\hat{S}_{ew}^2 = \left\{\frac{1}{2}\sum_{i\in s}\left(M_i-1\right)w_i\left[\sum_{j<k\in s(i)} w_{jk|i}\left(y_{ij}-y_{ik}\right)^2 \Big/ \sum_{j<k\in s(i)} w_{jk|i}\right]\right\}\left[\sum_{i\in s}\left(M_i-1\right)w_i\right]^{-1}, \qquad (3.12)$$

en supposant que $m_i > 1$ pour toutes les grappes échantillonnées. Notons que (3.12) nécessite les probabilités d'inclusion conjointe $\pi_{jk|i}$ comme la méthode proposée, mais qu'il induit un biais de ratio intra-grappe lorsque les tailles d'échantillon dans les grappes sont faibles, contrairement à notre méthode. L'expression pour $\hat{S}_{vw}^2$ est plus compliquée et nous invitons le lecteur à consulter Korn et Graubard (2003) pour la formule pertinente.

La méthode EEP peut être étendue facilement au modèle de régression linéaire à erreurs emboîtées

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij};\ e_{ij} \sim_{iid} N\left(0,\sigma_e^2\right),\ v_i \sim_{iid} N\left(0,\sigma_v^2\right). \qquad (3.13)$$

Dans ce cas, la fonction d'estimation (3.1) devient

$$u_1\left(y_{ij},\boldsymbol{\theta}\right) = \mathbf{x}_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right), \qquad (3.14)$$

La fonction d'estimation (3.2) devient

$$u_2\left(y_{ij},\boldsymbol{\theta}\right) = \left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2 - \left(\sigma_v^2+\sigma_e^2\right) \qquad (3.15)$$

et la fonction d'estimation (3.3) devient

$$u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T \boldsymbol{\beta}\right]^2 - 2\sigma_e^2, \; j \neq k, \tag{3.16}$$

où $\boldsymbol{\theta}$ est le vecteur des éléments $\boldsymbol{\beta}$, σ_v^2 et σ_e^2 et $z_{ijk} = y_{ij} - y_{ik}$. Les solutions explicites de $\hat{U}_{wt}(\boldsymbol{\theta}) = 0, t = 1,2,3$ correspondant aux équations (3.14) à (3.16) sont obtenues sous la forme

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\right)^{-1}\left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}y_{ij}\right), \tag{3.17}$$

$$\hat{\sigma}_{vw}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_w\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \tag{3.18}$$

et

$$\hat{\sigma}_{ew}^2 = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T\hat{\boldsymbol{\beta}}_w\right]^2 \Big/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\right). \tag{3.19}$$

3.2 Estimation de la variance

Un estimateur sandwich par linéarisation de Taylor de la variance de l'estimateur EEP $\hat{\boldsymbol{\theta}}_w$ peut être obtenu de manière analogue à l'estimateur de variance (2.10), à condition que la fraction d'échantillonnage de niveau 2 soit faible. Soit $\hat{\mathbf{U}}_w(\boldsymbol{\theta})$ le vecteur colonne dont les composantes sont $\hat{U}_{w1}(\boldsymbol{\theta})$, $\hat{U}_{w2}(\boldsymbol{\theta})$ et $\hat{U}_{w3}(\boldsymbol{\theta})$, et similairement $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$ le vecteur colonne dont les composantes sont $\hat{U}_{w1i}(\boldsymbol{\theta})$, $\hat{U}_{w2i}(\boldsymbol{\theta})$ et $\hat{U}_{w3i}(\boldsymbol{\theta})$. Alors, l'estimateur de variance par linéarisation est donné par

$$v_L\left(\hat{\boldsymbol{\theta}}_w\right) = \left(\hat{\mathbf{U}}'_w\right)^{-1}\left(\sum_{i\in s} w_i^2\,\hat{\mathbf{U}}_{wi}\hat{\mathbf{U}}_{wi}^T\right)\left[\left(\hat{\mathbf{U}}'_w\right)^{-1}\right]^T, \tag{3.20}$$

où $\hat{\mathbf{U}}_{wi}$ et $\hat{\mathbf{U}}'_w$ désignent $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$ évalué à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$, et la dérivée première $\hat{\mathbf{U}}'_w(\boldsymbol{\theta})$ évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$, respectivement. Les propriétés de l'estimateur de variance (3.20) sont étudiées par simulation à la section 5.2.

4 Log-vraisemblance composite pondérée : une approche unifiée

À la présente section, nous proposons une approche unifiée applicable aux modèles multiniveaux linéaires ainsi que linéaires généralisés. Cette approche est fondée sur le concept de la vraisemblance composite qui a acquis de la popularité dans la littérature ne portant pas sur les sondages pour traiter les données en grappes ou les données spatiales (voir par exemple, Lindsay 1988, Lele et Taper 2002 et Varin, Reid et Firth 2011). Une vraisemblance composite marginale par paire s'obtient en multipliant les contributions à la vraisemblance de toutes les paires distinctes dans les grappes. Notons que la vraisemblance composite est obtenue en prétendant que les sous-modèles sont indépendants. Lorsque le modèle de superpopulation est vérifié pour l'échantillon, nous pouvons obtenir les estimateurs des

paramètres en maximisant la vraisemblance composite par paire. Ici, nous étendons cette approche aux plans de sondage informatifs en obtenant des équations d'estimation pondérées qui requièrent seulement les poids marginaux w_i et $w_{j|i}$ et les poids par paire $w_{jk|i}$, comme à la section 3.

La log vraisemblance composite par paire de recensement est donnée par

$$l_C(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j<k=1}^{M_i} \log f\left(y_{ij},\, y_{ik}\middle|\boldsymbol{\theta}\right), \tag{4.1}$$

où $f\left(y_{ij},\, y_{ik}\middle|\boldsymbol{\theta}\right)$ est la densité de probabilité conjointe marginale de y_{ij} et y_{ik}. Nous estimons (4.1) par la log-vraisemblance composite par paire pondérée par les poids de sondage

$$l_{wC}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \log f\left(y_{ij},\, y_{ik}\middle|\boldsymbol{\theta}\right) \tag{4.2}$$

qui dépend seulement des probabilités d'inclusion de niveau 1 et de niveau 2 d'ordre 1 et de probabilités d'inclusion de niveau 1 d'ordre 2. Puis, nous résolvons les équations de score composite pondérées

$$\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta}) = \partial l_{wC}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}, \tag{4.3}$$

provenant de (4.2) pour obtenir un estimateur de la vraisemblance composite pondérée, $\hat{\boldsymbol{\theta}}_{wC}$, de $\boldsymbol{\theta}$. La méthode proposée est applicable aux modèles à deux niveaux linéaires et linéaires généralisés.

Nous notons que $\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})$, donné par (4.3), est un vecteur de fonctions d'estimation d'espérance nulle par rapport au plan et au modèle, c'est-à-dire $E_m E_p\left\{\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})\right\} = \mathbf{0}$. En utilisant ce résultat, on peut montrer que l'estimateur de la vraisemblance composite pondérée (VCP) $\hat{\boldsymbol{\theta}}_{wC}$ de $\boldsymbol{\theta}$ est convergent sous le modèle quand le nombre d'unités de niveau 2 dans l'échantillon, n, augmente, même si les tailles d'échantillon dans les grappes, m_i, sont petites. La preuve est exposée en détail dans Yi, Rao et Li (2012). Dans le contexte ne faisant pas appel au sondage, les preuves théoriques et empiriques que l'approche de la vraisemblance composite conduit à des estimateurs efficaces sont limitées (par exemple, Bellio et Varin 2005, Lindsay et coll. 2011). Notre étude en simulation (section 5) indique que l'approche de la vraisemblance composite pondérée donne de bons résultats en ce qui concerne l'efficacité, même si les tailles d'échantillon dans les grappes sont petites.

Dans le cas du modèle à erreurs emboîtées (3.13), en nous inspirant de Lele et Taper (2002), nous pouvons simplifier l'approche de la vraisemblance composite par paire en remplaçant la densité de probabilité bivariée $f\left(y_{ij}, y_{ik}\middle|\boldsymbol{\theta}\right)$ par les densités de probabilité univariées de y_{ij} et la différence $z_{ijk} = y_{ij} - y_{ik}$. Pour le modèle de la moyenne (2.2), nous avons $y_{ij} \sim N\left(\mu,\, \sigma_v^2 + \sigma_e^2\right)$ et $z_{ijk} \sim N\left(0,\, 2\sigma_e^2\right)$. En reparamétrisant $\boldsymbol{\theta} = \left(\mu, \sigma_v^2, \sigma_e^2\right)^T$ de manière que $\boldsymbol{\phi} = \left(\mu, \sigma^2, \sigma_e^2\right)^T$, où $\sigma^2 = \sigma_v^2 + \sigma_e^2$, nous voyons que les paramètres des deux densités de probabilité univariées sont distincts et que les log-vraisemblances composites correspondant à y_{ij} et z_{ijk} sont données par

$$l_{wCy}\left(\mu, \sigma^2\right) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} \log f\left(y_{ij}\middle|\mu, \sigma^2\right)$$

et

$$l_{wCz}\left(\sigma_e^2\right) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \log f\left(z_{ijk}\middle|\sigma_e^2\right).$$

Nous résolvons alors le système d'équations de score composite pondérées résultantes

$$\hat{U}_{wCy1}\left(\mu,\sigma^2\right) = \partial l_{wCy}\left(\mu,\sigma^2\right)\Big/\partial\mu = \sum_{i\in s} w_i \sum_{j\in s_i} w_{j|i}\left(y_{ij}-\mu\right)\Big/\sigma^2 = 0,$$

$$\hat{U}_{wCy2}\left(\mu,\sigma^2\right) = \partial l_{wCy}\left(\mu,\sigma^2\right)\Big/\partial\sigma^2 = \frac{1}{2}\sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\left[-\frac{1}{\sigma^2}+\frac{\left(y_{ij}-\mu\right)^2}{\sigma^4}\right] = 0$$

$$\hat{U}_{wCz}\left(\sigma_e^2\right) = \partial l_{wCz}\left(\sigma_e^2\right)\Big/\partial\sigma_e^2 = \frac{1}{2}\sum_{i} w_i \sum_{j<k\in s(i)} w_{jk|i}\left(-\frac{1}{\sigma_e^2}+\frac{z_{ijk}^2}{2\sigma_e^4}\right) = 0$$

pour obtenir les estimateurs de la vraisemblance composite pondérée (VCP) $\hat{\mu}_{wC}, \hat{\sigma}_{vwC}^2$ et $\hat{\sigma}_{ewC}^2$. Les estimateurs VCP sont identiques aux estimateurs (3.9) à (3.11) obtenus par l'approche des équations d'estimation pondérées de la section 3.

Nous nous penchons maintenant sur le modèle de régression linéaire à erreurs emboîtées (3.13). Mentionnons pour commencer que $y_{ij} \sim N\left(\mathbf{x}_{ij}^T\boldsymbol{\beta},\, \sigma^2\right)$, où $\sigma^2 = \sigma_v^2 + \sigma_e^2$, et $z_{ijk} = y_{ij} - y_{ik} \sim N\left\{\left(\mathbf{x}_{ij}-\mathbf{x}_{ik}\right)^T\boldsymbol{\beta},\, 2\sigma_e^2\right\}$. Il s'ensuit que les équations de score composite pondérées sont données par

$$\begin{aligned}\hat{\mathbf{U}}_{wCy1}(\boldsymbol{\beta},\sigma^2) &= \partial l_{wCy}\left(\boldsymbol{\beta},\sigma^2\right)\Big/\partial\boldsymbol{\beta} \\ &= \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\mathbf{x}_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right) = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{U}}_{wCy2}(\boldsymbol{\beta},\sigma^2) &= \partial l_{wCy}\left(\boldsymbol{\beta},\sigma^2\right)\Big/\partial\sigma^2 \\ &= -\frac{1}{2}\sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\left[\frac{1}{\sigma^2}-\frac{\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2}{\sigma^4}\right] = 0\end{aligned}$$

et

$$\begin{aligned}\hat{\mathbf{U}}_{wCz}(\sigma_e^2) &= \partial l_{wCz}\left(\sigma_e^2\right)\Big/\partial\sigma_e^2 \\ &= -\frac{1}{2}\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left\{\frac{1}{\sigma_e^2}-\frac{\left[z_{ijk}-\left(\mathbf{x}_{ij}-\mathbf{x}_{ik}\right)^T\boldsymbol{\beta}\right]^2}{2\sigma_e^4}\right\} = 0.\end{aligned}$$

Les estimateurs VCP résultants de $\boldsymbol{\beta}$, σ_v^2 et σ_e^2 sont donnés par

$$\hat{\boldsymbol{\beta}}_{wC} = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\right)^{-1}\left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}y_{ij}\right),$$

$$\hat{\sigma}_{wC}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{wC}\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij},$$

et

$$\hat{\sigma}^2_{ewC} = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T \hat{\boldsymbol{\beta}}_{wC} \right]^2 \Bigg/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \right).$$

L'estimateur de σ_v^2 est donné par $\hat{\sigma}^2_{vwC} = \hat{\sigma}^2_{wC} - \hat{\sigma}^2_{ewC}$. De nouveau, les estimateurs VCP $\hat{\boldsymbol{\beta}}_{WC}$, $\hat{\sigma}^2_{vWC}$ et $\hat{\sigma}^2_{eWC}$ sont identiques aux estimateurs (3.17) à (3.19) obtenus par l'approche des équations d'estimation pondérées de la section 3.

L'approche de la vraisemblance composite susmentionnée, fondée sur y_{ij} et $z_{ijk} = y_{ij} - y_{ik}$, n'est pas applicable au modèle à deux niveaux linéaire donné par (2.4), parce que le vecteur de paramètres, $\boldsymbol{\theta}$, n'est pas identifiable sous la vraisemblance composite obtenue à partir des y_{ij} et z_{ijk}. Nous devons faire appel à la méthode par paire pour traiter le modèle (2.4).

Marginalement, $\left(y_{ij}, y_{ik}\right)^T$ suit une loi normale bivariée de moyennes $\mathbf{x}_{ij}^T\boldsymbol{\beta}$ et $\mathbf{x}_{ik}^T\boldsymbol{\beta}$ et de matrice de covariance 2×2

$$\boldsymbol{\Sigma}_{i(jk)} = \begin{bmatrix} \sigma_e^2 + \mathbf{x}_{ij}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ij} & \mathbf{x}_{ij}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ik} \\ \mathbf{x}_{ik}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ij} & \sigma_e^2 + \mathbf{x}_{ik}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ik} \end{bmatrix}.$$

Maintenant, il découle de (4.3) que les équations de score composite pondérées sont données par

$$\boldsymbol{\beta}: \qquad \hat{\mathbf{U}}_{wC\boldsymbol{\beta}} = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\mathbf{X}_{i(jk)}^T\boldsymbol{\Sigma}_{i(jk)}^{-1}\left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T\boldsymbol{\beta}\right) = \mathbf{0} \tag{4.4}$$

et

$$\begin{aligned} \boldsymbol{\tau}: \qquad \hat{\mathbf{U}}_{wCl} = \frac{1}{2}\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\Bigg[& \left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{i(jk)}^{-1}\frac{\partial\boldsymbol{\Sigma}_{i(jk)}}{\partial\tau_l}\boldsymbol{\Sigma}_{i(jk)}^{-1}\left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T\boldsymbol{\beta}\right) \\ & -\mathrm{tr}\left(\boldsymbol{\Sigma}_{i(jk)}^{-1}\frac{\partial\boldsymbol{\Sigma}_{i(jk)}}{\partial\tau_l}\right)\Bigg] = \mathbf{0}, \qquad l = 1,\ldots, p(p+1)/2 + 1 = P \end{aligned} \tag{4.5}$$

où $\mathbf{X}_{i(jk)}$ est la matrice de dimensions $2\times p$ contenant les lignes $\mathbf{x}_{ij}^T$ et $\mathbf{x}_{ik}^T$, $\mathbf{y}_{i(jk)} = \left(y_{ij}, y_{ik}\right)^T$, et $\boldsymbol{\tau}$ est le vecteur de dimension P contenant les éléments $\tau_1 = \sigma_e^2$ et les $p(p+1)/2$ éléments distincts de $\boldsymbol{\Sigma}_v$ désignés par $\tau_2,\ldots,\tau_P$. Nous pouvons résoudre les équations de score composite pondérées (4.4) et (4.5) itérativement en utilisant la méthode de Newton-Raphson ou une autre méthode itérative pour obtenir les estimateurs VCP $\hat{\boldsymbol{\beta}}_{wC}$ et $\hat{\boldsymbol{\tau}}_{wC}$.

Dans le cas particulier du modèle de régression linéaire à erreurs emboîtées (3.13), les équations de score de recensement, fondées sur la log-vraisemblance de recensement complète $l(\boldsymbol{\theta})$ donnée par (2.5), peuvent s'écrire sous une forme explicite. Les équations de score pondérées d'échantillon correspondantes ne dépendent que des poids de niveau 1 $w_{j|i}$ et $w_{jk|i}$ et des poids de niveau 2 w_i, comme les équations de score composite pondérées (voir l'annexe). Les estimateurs résultants sont convergents sous le modèle pour $\boldsymbol{\theta}$, contrairement aux estimateurs fondés sur la pseudo log-vraisemblance pondérée $l_w(\boldsymbol{\theta})$ donnés par (2.7) et (2.8). Cependant, pour des modèles plus complexes, comme les modèles à deux niveaux avec pentes aléatoires, les équations de score pondérées d'échantillon dépendront des probabilités d'inclusion de niveau 1 d'ordres 3 et 4, contrairement aux équations de score composite pondérées (4.3) qui ne

dépendent que des probabilités d'inclusion de niveau 1 d'ordres 1 et 2, même pour les modèles multiniveaux complexes. Par conséquent, nous n'avons pas inclus l'approche des équations de score pondérées fondée sur la log-vraisemblance de recensement complète dans l'étude en simulation.

5 Étude en simulation

Nous avons réalisé une petite étude en simulation de la performance des estimateurs EEP proposés sous le simple modèle de la moyenne à erreurs emboîtées, en utilisant $\mu = 0{,}5$, $\sigma_v^2 = 0{,}5$ et $\sigma_e^2 = 2{,}0$. La population est constituée de $N = 1\,000$ grappes, chacune contenant $M_i = M = 100$ éléments. Nous avons utilisé un plan d'échantillonnage à deux degrés avec $n = 50$ grappes échantillonnées et $m_i = m = 5$ éléments sélectionnés dans chaque grappe échantillonnées. Les grappes ont été sélectionnées par échantillonnage aléatoire simple, et les éléments dans les grappes, par la méthode d'échantillonnage avec probabilité proportionnelle à la taille (PPT) de Rao-Sampford (Rao 1965 et Sampford 1967) avec des mesures de taille spécifiées z_{ij}. Nous avons choisi les mesures de taille de manière à refléter les divers niveaux d'informativité de l'échantillonnage.

À l'instar d'Asparouhov (2006), nous prenons en considération la sélection invariante ainsi que non invariante. Pour la sélection invariante, la mesure de taille z_{ij} dépend seulement des erreurs de niveau 1 et ne varie pas d'une grappe à l'autre. En particulier, nous posons

$$z_{ij} = \left(1 + \exp\left\{-0.5\left[e_{ij}/\alpha + e_{ij}^{*}\left(1 - \alpha^{-2}\right)^{1/2}\right]\right\}\right)^{-1}, \tag{5.1}$$

où e_{ij}^{*} est indépendante de e_{ij} mais de même loi, $N\left(0, \sigma_e^2 = 2.0\right)$. Pour la sélection non invariante, la mesure de taille z_{ij} dépend des erreurs de niveau 1 ainsi que de niveau 2 et n'est donc pas invariante d'une grappe à l'autre. En particulier, dans (3.7), nous remplaçons e_{ij} et e_{ij}^{*} par $v_i + e_{ij}$ et $v_i^{*} + e_{ij}^{*}$, respectivement, où v_i^{*} est indépendant de v_i mais de même loi $N\left(0, \sigma_v^2 = 0{,}5\right)$. Nous considérons quatre valeurs de α dans (5.1) : $\alpha = 1, 2, 3, \infty$, où $\alpha = \infty$ correspond à l'échantillonnage non informatif dans chaque grappe, $\alpha = 1$ correspond à l'échantillonnage le plus informatif, et l'informativité diminue quand α augmente.

Nous avons utilisé l'approche fondée sur le plan de sondage et le modèle (*pm*) pour simuler $R = 1\,000$ échantillons pour chaque valeur spécifiée de α et séparément pour les sélections invariante et non invariante. Sous cette approche, nous avons généré une population correspondant à $N = 1000$ et $M_i = M = 100$ à partir du modèle, puis nous avons sélectionné un échantillon à deux degrés d'éléments comme il est spécifié plus haut. Le processus en deux étapes a été répété $R = 1000$ fois pour simuler 1000 échantillons.

5.1 Performance des estimateurs

Partant de chaque échantillon, nous avons calculé les estimations de μ, σ_v^2 et σ_e^2 en utilisant le maximum de vraisemblance restreint (REML), les méthodes d'ajustement des pondérations A et A1, la

méthode EEP proposée et la méthode de rechange de Korn et Graubard (appelée KG). Les biais et les variances des estimateurs ont été calculés en se servant des 1000 estimations. La performance des divers estimateurs est évaluée en utilisant deux mesures, à savoir le ratio du biais = RB = (biais)/ (racine carrée de la variance) et la racine carrée de l'erreur quadratique moyenne relative = REQMR = (racine carrée de l'EQM)/ (valeur réelle du paramètre). Les tableaux 5.1, 5.2 et 5.3 donnent, respectivement, les valeurs du RB des estimateurs de μ, σ_v^2 et σ_e^2. Les valeurs de la REQMR des estimateurs de μ, σ_v^2 et σ_e^2 sont présentées aux tableaux 5.4, 5.5 et 5.6, respectivement.

Tableau 5.1
Ratio du biais (%) des estimateurs de μ

α	Sélection invariante			Sélection non invariante		
	REML	**A**	**A1/EEP/KG**	**REML**	**A**	**A1/EEP/KG**
1	346,5	80,2	2,2	370,9	83,9	3,0
2	167,7	40,1	0,3	172,3	45,3	6,1
3	114,3	30,7	4,5	114,9	30,8	4,8
∞	2,0	2,5	2,1	-1,5	-2,4	-2,2

Le tableau 5.1 donne le ratio du biais (%) des estimateurs de μ fondés sur le REML, les méthodes d'ajustement des pondérations A et A1, et les méthodes KG et EEP. Notons que dans le cas de μ, les estimateurs A1, KG et EEP (VCP) sont identiques. Les résultats du tableau 5.1 montrent que le RB est le même pour les sélections invariante et non invariante, et que le RB des méthodes REML et A diminue lorsque α augmente. En outre, la méthode REML produit un biais plus important sous échantillonnage informatif, même pour $\alpha = 3$; par exemple, le RB pour la méthode REML varie de 114 % à 346 % sous sélection invariante. La méthode A conduit aussi à un RB important sous échantillonnage informatif; par exemple le RB pour la méthode A varie de 30,8 % à 83,9 % sous sélection non invariante. Par ailleurs, le RB des méthodes EEP, A1 et KG ne dépend pas de α et est faible ($|RB| < 6\%$). Sous échantillonnage non informatif, la méthode REML donne d'aussi bons résultats que prévus ($|RB| < 3\%$).

Pour l'estimation de σ_v^2, commençons par noter que la proportion de fois que l'estimation de σ_v^2 est négative est nulle dans les simulations pour les quatre valeurs de α et pour toutes les méthodes d'estimation (REML, A, A1, EEP et KG). Le tableau 5.2 donne les valeurs du RB pour les estimateurs de σ_v^2. Il montre que le RB de la méthode REML n'est pas affecté par α sous sélection invariante, mais qu'il l'est sous sélection non invariante. Dans ce dernier cas, le REML donne lieu à une sous-estimation importante pour $\alpha = 1$ (RB = -49 %), mais |RB| diminue à mesure que α augmente. Le tableau 5.2 montre aussi que les méthodes A et A1 ne donnent pas d'aussi bons résultats sous échantillonnage informatif (RB variant de 16 % à 60 %). La méthode KG n'a pas donné de bons résultats pour $\alpha = 1$ (RB = 33 % sous sélection invariante et RB = 24 % sous sélection non invariante). Par ailleurs, la méthode EEP donne de bons résultats pour toutes les valeurs de α (RB variant de -4 % à -13 %) mais la sous-estimation est systématique pour les diverses valeurs de α.

Le tableau 5.3 donne les valeurs du RB des estimateurs de σ_e^2. Il montre que ces valeurs sont les mêmes pour les sélections invariante et non invariante, comme dans le cas de μ. Les méthodes REML et KG donnent lieu à une sous-estimation importante quand $\alpha = 1$ (RB = -107 % pour la méthode REML et RB = -71 % pour la méthode KG sous sélection invariante), mais |RB| diminue quand α augmente et devient négligeable pour $\alpha = \infty$. La performance des estimateurs A et A1 est médiocre pour toutes les valeurs de α, y compris $\alpha = \infty$. Par ailleurs, la méthode EEP donne de bons résultats pour toutes les valeurs de α avec |RB| < 8 %. Il semble que l'instabilité introduite par le facteur d'échelle (2.9) pourrait avoir contribué à la grande valeur de |RB| pour les méthodes A et A1 même sous échantillonnage non informatif ($\alpha = \infty$).

Tableau 5.2
Ratio du biais (%) des estimateurs de σ_v^2

α	REML	A	A1	EEP	KG
			Sélection invariante		
1	0,6	59,5	59,3	-8,5	33,2
2	0,5	24,5	26,3	-10,0	8,0
3	-3,4	16,1	18,2	-13,6	0,4
∞	-0,1	14,8	17,1	-8,9	0,6
			Sélection non invariante		
1	-49,0	50,1	58,9	-4,4	24,0
2	-10,9	24,6	28,7	-7,0	7,1
3	-4,0	20,0	22,7	-7,8	4,6
∞	-1,3	12,8	13,9	-13,3	-1,6

Tableau 5.3
Ratio du biais (%) des estimateurs de σ_e^2

α	REML	A	A1	EEP	KG
			Sélection invariante		
1	-106,9	-118,4	-66,9	2,4	-71,2
2	-22,7	-43,6	-34,3	2,1	-16,5
3	-9,4	-31,7	-28,4	2,9	-6,5
∞	-0,4	-21,8	-23,8	0,3	0,4
			Sélection non invariante		
1	-115,3	-131,3	-79,6	-6,9	-82,6
2	-30,4	-51,1	-43,3	-7,6	-23,9
3	-12,5	-34,9	-32,2	-2,3	-10,3
∞	1,1	-20,2	-21,8	2,6	1,6

Tableau 5.4
Racine carrée de l'erreur quadratique moyenne relative (%) des estimateurs de μ

α	Sélection invariante			Sélection non invariante		
	REML	A	A1/EEP/KG	REML	A	A1/EEP/KG
1	93,3	35,9	29,4	92,5	35,4	29,2
2	51,6	29,3	27,8	52,8	30,4	28,9
3	40,5	28,2	27,5	40,8	28,7	28,1
∞	25,8	26,1	26,5	26,6	27,3	27,7

Racine carrée de l'erreur quadratique moyenne relative

Le tableau 5.4 montre que les valeurs de la REQMR (%) des estimations de μ sont les mêmes pour les sélections invariante et non invariante, et que la REQMR des méthodes REML et A diminue quand α augmente. Sous échantillonnage informatif avec $\alpha = 1$, la REQMR pour la méthode REML est grande comparativement à celle de la méthode EEP (A1 et KG) lorsque le RB est grand. Par exemple, la REQMR = 93 % pour la méthode REML comparativement à REQMR = 29 % pour la méthode EEP. Comme prévu, la méthode REML donne la plus petite REQMR sous échantillonnage non informatif, mais l'augmentation de la REQMR pour les autres méthodes est assez faible. En outre, la REQMR de la méthode EEP (A1 et KG) dépend de α.

Tableau 5.5
Racine carrée de l'erreur quadratique moyenne relative (%) des estimateurs de σ_v^2

α	REML	A	A1	EEP	KG
	Sélection invariante				
1	36,5	47,3	51,1	43,6	43,8
2	37,1	39,7	41,1	40,5	39,5
3	36,3	37,3	38,7	39,5	37,8
∞	35,8	36,9	38,1	38,7	37,2
	Sélection non invariante				
1	36,7	44,6	52,6	43,4	41,5
2	35,6	37,9	40,4	39,3	37,7
3	37,0	38,7	40,4	40,2	38,8
∞	36,6	37,2	38,0	39,0	37,8

Pour ce qui est de la REQMR des estimateurs de σ_v^2, le tableau 5.5 montre que la méthode REML donne de bons résultats pour toutes les valeurs de α sous sélection invariante parce que le RB est petit dans ce cas. Nous constatons aussi que les méthodes KG et EEP ont une REQMR comparable pour toutes les valeurs de α. Le tableau 5.5 révèle aussi que les méthodes A et A1 produisent une REQMR un peu plus grande pour $\alpha = 1$: 51 % pour A1 et 47 % pour A sous sélection invariante comparativement à 44 % pour la méthode EEP.

Tableau 5.6
Racine carrée de l'erreur quadratique moyenne relative (%) des estimateurs de σ_e^2

α	REML	A	A1	EEP	KG
			Sélection invariante		
1	13,5	14,5	12,8	13,9	12,9
2	9,7	10,4	10,4	11,0	10,0
3	9,5	10,0	10,1	10,7	9,8
∞	10,1	10,3	10,5	11,1	10,3
			Sélection non invariante		
1	13,7	14,8	12,9	13,2	13,0
2	10,0	10,9	10,9	11,3	10,3
3	9,7	10,4	10,7	11,2	10,2
∞	10,3	10,6	10,8	11,4	10,7

Le tableau 5.6 donne les valeurs de la REQMR pour les estimateurs de σ_e^2 et nous constatons que les valeurs sont similaires pour les sélections invariante et non invariante. Le tableau montre aussi que les valeurs de la REQMR sont comparables pour les méthodes EEP, A, Al et KG, même si, en ce qui concerne le ratio du biais, les méthodes A, Al et KG donnent de moins bons résultats que la méthode EEP. Cette situation est due au fait que la variance est plus grande pour EEP que pour les autres méthodes. Par exemple, dans le cas de la sélection invariante et $\alpha = 1$, nous obtenons les variances qui suivent pour les méthodes EEP, KG et REML: 0,0771, 0,0438 et 0,0339 avec des ratios du biais (%) correspondants provenant du tableau 5.3 : 2,4, -71,2 et -106,9.

5.2 Performance de l'estimateur de variance

Nous présentons maintenant certains résultats de simulation concernant le biais relatif de l'estimateur de variance par linéarisation (3.12) de l'estimateur EEP (VCP) $\hat{\boldsymbol{\theta}}_w$. Nous avons d'abord répété le processus en deux étapes $R_1 = 2000$ fois et calculé $v_L^{(r)}\left(\hat{\boldsymbol{\theta}}_w\right)$ en partant de chaque échantillon à deux

degrés $r = 1,..., 2000$. Les moyennes des éléments diagonaux de $E\left\{v_L\left(\hat{\boldsymbol{\theta}}_w\right)\right\} \approx v_L\left(\hat{\boldsymbol{\theta}}_w\right) = R_1^{-1}\sum_{r=1}^{R_1} v_L^{(r)}\left(\hat{\boldsymbol{\theta}}_w\right)$ sont désignées par $\overline{v}_L\left(\hat{\mu}_w\right)$, $\overline{v}_L\left(\hat{\sigma}_{vw}^2\right)$ et $\overline{v}_L\left(\hat{\sigma}_{ew}^2\right)$, respectivement. Nous avons ensuite généré $R_2 = 10000$ échantillons indépendants et calculé l'erreur quadratique moyenne (EQM) empirique des trois estimateurs $\hat{\mu}_w$, $\hat{\sigma}_{vw}^2$ et $\hat{\sigma}_{ew}^2$. Nous avons $\mathrm{EQM}\left(\hat{\mu}_w\right) \approx R_2^{-1}\sum_{r=1}^{R_2}\left(\hat{\mu}_w^{(r)} - \mu\right)^2$, où $\hat{\mu}_w^{(r)}$ est l'estimation de μ d'après le r^{e} échantillon simulé, et les expressions similaires pour $\mathrm{EQM}\left(\hat{\sigma}_{vw}^2\right)$ et $\mathrm{EQM}\left(\hat{\sigma}_{ew}^2\right)$.

Le biais relatif de $v_L\left(\hat{\mu}_w\right)$ est calculé comme

$$\mathrm{BR}\left\{v_L\left(\hat{\mu}_w\right)\right\} = \left[\overline{v}_L\left(\hat{\mu}_w\right)/\mathrm{EQM}\left(\hat{\mu}_w\right)\right] - 1$$

et $\mathrm{BR}\left\{v_L\left(\hat{\sigma}_{vw}^2\right)\right\}$ et $\mathrm{BR}\left\{v_L\left(\hat{\sigma}_{ew}^2\right)\right\}$ ont été calculés de la même façon. Le tableau 5.7 donne les valeurs du BR pour les trois estimateurs de variance sous sélections invariante et non invariante et $\alpha = 1, 2, 3, \infty$. L'examen du tableau 5.7 montre clairement que l'estimateur de variance par linéarisation donne de bons résultats pour toutes les combinaisons, avec $|\mathrm{BR}| < 10\%$.

Tableau 5.7
Biais relatif (%) des estimateurs de variance

α	$v_L(\hat{\mu}_w)$	$v_L(\hat{\sigma}_{vw}^2)$	$v_L(\hat{\sigma}_{ew}^2)$
	Sélection invariante		
1	-3,0	-6,2	-7,5
2	-5,2	-4,5	-3,1
3	-1,3	-3,8	-1,8
∞	-0,9	-2,5	-2,0
	Sélection non invariante		
1	-3,8	-8,3	-4,2
2	-4,5	-5,8	-7,3
3	-4,3	-4,6	-5,7
∞	-2,4	-2,7	-2,9

6 Conclusion

Dans le présent article, nous avons proposé une approche unifiée fondée sur la vraisemblance composite pondérée (VCP) pour les modèles à deux niveaux pour faire des inférences sur des données d'enquête complexes. Les méthodes VCP proposées sont asymptotiquement valides même quand les tailles d'échantillon dans les grappes échantillonnées (unités de niveau 1) sont petites, contrairement à

certaines méthodes existantes, mais il est nécessaire de connaître les probabilités d'inclusion conjointe dans les grappes échantillonnées. Souvent, il est possible de traiter l'échantillonnage dans les grappes comme étant effectué avec remise, en raison des petites fractions d'échantillonnage dans les grappes. En outre, d'excellentes approximations des probabilités d'inclusion conjointe, qui ne dépendent que des probabilités d'inclusion marginales, sont disponibles lorsque les fractions d'échantillonnage ne sont pas petites (Haziza et coll. 2008). Nous prévoyons examiner l'exactitude de ce genre d'approximations dans le cadre d'une future étude. Des études en simulation de la performance des estimateurs VCP (4.5) et (4.6) pour les modèles à deux niveaux (2.3) fondées sur la méthode par paire seront également réalisées.

Les méthodes fondées sur la vraisemblance composite sont utilisées principalement lorsque la vraisemblance complète est complexe. Notre développement dans le contexte des sondages donne la preuve que la méthode fondée sur la vraisemblance complète avec pondérations n'est pas faisable pour des modèles multiniveaux, tandis que la méthode fondée sur la vraisemblance composite pondérée facilite l'obtention d'inférences valides, même si les tailles d'échantillon de grappe sont petites.

7 Remerciements

Nous remercions deux examinateurs et le rédacteur associé de leurs suggestions et commentaires constructifs.

Annexe

Équations de score pondérées : modèle de régression linéaire à erreurs emboîtées

Pour le modèle de régression linéaire à erreurs emboîtées (2.3), une forme explicite de la log-vraisemblance complète de recensement s'obtient en utilisant la forme explicite de la matrice de covariance $\mathbf{V}_i$ de $\mathbf{y}_i = \left(y_{i1},...,y_{iM_i}\right)^T$. Nous avons $\mathbf{V}_i^{-1} = \sigma_e^{-2}\left[\mathbf{I}_i - \sigma_v^2 / \lambda_i \mathbf{1}_i \mathbf{1}_i^T\right]$, où $\lambda_i = \sigma_e^2 + M_i\sigma_v^2$, $\mathbf{I}_i$ est la matrice identité de dimensions $M_i \times M_i$ et $\mathbf{1}_i$ est le vecteur unité de dimension $M_i \times 1$. En utilisant l'expression pour $\mathbf{V}_i^{-1}$, les équations de score de recensement s'obtiennent sous la forme

$$\boldsymbol{\beta}: \quad \left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mathbf{x}_{ij}y_{ij} - \sigma_v^2\sum_{i=1}^{N}\lambda_i^{-1}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\mathbf{x}_{ij}y_{ik}\right)\right] - \left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \sigma_v^2\sum_{i=1}^{N}\lambda_i^{-1}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\mathbf{x}_{ij}\mathbf{x}_{ik}^T\right)\right]\boldsymbol{\beta} = 0 \quad \text{(A.1)}$$

$$\sigma_v^2: \quad \sum_{i=1}^{N}\lambda_i^{-2}\left[\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik} - \mathbf{x}_{ik}^T\boldsymbol{\beta}\right)\right] - \sum_{i=1}^{N}\lambda_i^{-1}M_i = 0 \quad \text{(A.2)}$$

$$\sigma_e^2: \quad \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2+\sum_{i=1}^{N}\left(M_i\sigma_v^4\lambda_i^{-2}-2\sigma_v^2\lambda_i^{-1}\right)\sum_{j,k=1}^{M_i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik}-\mathbf{x}_{ik}^T\boldsymbol{\beta}\right) - \sigma_e^2\sum_{i=1}^{N}\left(1-\sigma_v^2\lambda_i^{-1}\right)M_i=0 \tag{A.3}$$

Partant de (A.1), nous obtenons les équations de score pondérées

$$\boldsymbol{\beta}: \quad \sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\mathbf{x}_{ij}y_{ij}-\sigma_v^2\sum_{i\in s}w_i\lambda_i^{-1}\left(\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\mathbf{x}_{ij}y_{ik}\right) -\left[\sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T-\sigma_v^2\sum_{i\in s}w_i\lambda_i^{-1}\left(\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\mathbf{x}_{ij}\mathbf{x}_{ik}\right)\right]\boldsymbol{\beta}=0 \tag{A.4}$$

où $w_{jj|i}=w_{j|i}$. Notons que les tailles de grappe M_i pour $i\in s$ sont supposées connues. On ne doit pas remplacer M_i par son estimation $\sum_{j\in s(i)}w_{j|i}$, parce que celle-ci comprend un biais de ratio dû aux petites tailles d'échantillon dans les grappes. L'équation d'estimation (A.4) est sans biais sous le plan pour l'équation de recensement (A.1).

Passons maintenant à l'équation de score pondérée pour σ_v^2, nous obtenons en partant de (A.2)

$$\sigma_v^2: \quad \sum_{i\in s}w_i\lambda_i^{-2}\left[\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik}-\mathbf{x}_{ik}^T\boldsymbol{\beta}\right)\right]-\sum_{i\in s}w_i\lambda_i^{-1}\sum_{j\in s(i)}w_{j|i}=0 \tag{A.5}$$

L'équation d'estimation (A.5) est sans biais pour (A.2). Enfin, l'équation de score pondérée pour σ_e^2 s'obtient à partir de (A.3) sous la forme

$$\sigma_e^2: \quad \sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\left(y_{ij}-x_{ij}^T\beta\right)^2+\sum_{i\in s}w_i\left(M_i\sigma_v^4\lambda_i^{-2}-2\sigma_v^2\lambda_i^{-1}\right)\sum_{j,k\in s(i)}w_{jk|i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik}-\mathbf{x}_{ik}^T\boldsymbol{\beta}\right) -\sigma_e^2\sum_{i\in s}w_i\left(1-\sigma_v^2\lambda_i^{-1}\right)\sum_{j\in s(i)}w_{j|i}=0 \tag{A.6}$$

Il découle des équations (A.4) à (A.6) que les équations de score pondérées dépendent uniquement des pondérations d'ordre 1 w_i et $w_{j|i}$ et des pondérations d'ordre 2 $w_{jk|i}$ dans le cas particulier d'un modèle de régression linéaire à erreurs emboîtées.

Bibliographie

Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, 35, 439-460.

Beaumont, J.-F., et Charest, A.-S. (2010). Bootstrap variance estimation with survey data when estimating model parameters. Document non publié (fourni par les auteurs).

Bellio, R., et Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 3, 217-227.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Grilli, L., et Pratesi, M. (2004). Estimation pondérée dans le cadre de modèles multiniveaux ordinaux et binaires sous un plan d'échantillonnage informatif. *Techniques d'enquête*, 30, 103-114.

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*, 66, 91-108.

Korn, E.L., et Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society B*, 65, 175-190.

Kovacevic, M.S., Rong, H. et You, Y. (2006). Bootstrapping for variance estimation in multi-level models fitted to survey data. *Proceedings of ASA Section on Survey Research Methods*, American Statistical Association, 3260-3269.

Lele, S., et Taper, M.L. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 103, 117-125.

Lindsay, B.G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, (Ed. N.U. Prabhu), Providence: American Mathematical Society, 221-239.

Lindsay, B.G., Yi, G.Y. et Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71-105.

Muthén, L.K., et Muthén, B.O. (1998-2005). *Mplus User's Guide. 3rd ed. Los Angeles, CA: Muthén & Muthén*.

Pfeffermann, D., et Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Dans *Analysis of Survey Data*, (Éds. R. Chambers et C.J. Skinner), Wiley, Chichester, 175-196.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society B*, 60, 23-56.

Rabe-Hesketh, S., et Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society A*, 169, 805-827.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

Rao, J.N.K., et Roberts, G. (1998). Discussion on the papers by Firth and Bennett and Pfeffermann *et al*. *Journal of the Royal Statistical Society B*, 60, 50-51.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

Rao, J.N.K., Hidiroglou, M., Yung, W. et Kovacevic, M. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of the Indian Society of Agricultural Statistics*, 64, 129-135.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Varin, C., Reid, N. et Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.

Yi, G.Y., Rao, J.N.K. et Li, H. (2012). A weighted composite likelihood approach for analysis of survey data under two level models. Disponible sur demande auprès de jrao@math.carleton.ca.