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

Techniques d'enquête

Estimateurs de la variance robustes pour estimateurs par la régression généralisée dans des échantillons en grappes

par Timothy L. Kennel et Richard Valliant

Date de diffusion : le 17 décembre 2019

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 STATCAN.infostats-infostats.STATCAN@canada.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

Programme des services de dépôt

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

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.

Techniques d'enquête, décembre 2019
Vol. 45, nº 3, pp. 455-481
Statistique Canada, nº 12-001-X au catalogue

Estimateurs de la variance robustes pour estimateurs par la régression généralisée dans des échantillons en grappes

Timothy L. Kennel et Richard Valliant[1]

Résumé

Les estimateurs de la variance par linéarisation classiques de l'estimateur par la régression généralisée sont souvent trop petits, ce qui entraîne des intervalles de confiance ne donnant pas le taux de couverture souhaité. Pour remédier à ce problème, on peut apporter des ajustements à la matrice chapeau dans l'échantillonnage à deux degrés. Nous présentons la théorie de plusieurs nouveaux estimateurs de la variance et les comparons aux estimateurs classiques dans une série de simulations. Les estimateurs proposés corrigent les biais négatifs et améliorent les taux de couverture de l'intervalle de confiance dans diverses situations correspondant à celles rencontrées en pratique.

Mots-clés : Estimateur de la variance jackknife; ajustement de matrice chapeau; ajustement d'effets de levier; modèle de superpopulation; échantillon à deux degrés; estimateur de la variance sandwich.

1 Introduction

L'estimation par la régression généralisée (GREG) est une technique courante utilisée pour caler les estimations, réduire les erreurs d'échantillonnage et corriger les erreurs non dues à l'échantillonnage. Les enquêtes officielles auprès des ménages utilisent souvent la régression généralisée pour caler des estimations fondées sur les données d'échantillon aux contrôles de population, assurer des estimations convergentes des caractéristiques démographiques dans l'ensemble des enquêtes, et réduire les erreurs de non-réponse et de sous-dénombrement. L'estimation par la régression généralisée est également fréquemment utilisée parce qu'elle tire sa puissance des données auxiliaires, ce qui entraîne des erreurs d'échantillonnage plus petites que celles d'autres estimateurs fondés sur le plan.

Les techniques utilisées couramment pour estimer les erreurs d'échantillonnage des estimateurs calés à partir d'échantillons complexes soit nécessitent d'importantes ressources de calcul, soit tendent à sous-estimer les erreurs d'échantillonnage véritables, surtout en cas de taille d'échantillon petite ou moyenne. Deux des techniques courantes d'estimation de la variance d'échantillonnage des estimateurs GREG sont la linéarisation et le rééchantillonnage. Les estimateurs par la linéarisation (Särndal, Swensson et Wretman, 1989) ne convergent peut-être pas assez rapidement vers l'erreur d'échantillonnage véritable pour produire des résultats exacts dans des échantillons petits ou moyens. Särndal, Swensson et Wretman (1992, page 176) constatent que « pour des statistiques complexes comme un estimateur de variance de covariance ou de coefficient de corrélation pour la population, il se peut qu'il faille des échantillons assez grands pour que le biais soit négligeable ». Par ailleurs, d'autres techniques de rééchantillonnage comme le jackknife et le bootstrap qui produisent généralement des estimations de la variance plus importantes peuvent exiger d'importantes ressources de calcul.

1. Timothy L. Kennel est chef adjoint de la Division des méthodes statistiques à la Decennial Statistical Studies Division du U.S. Census Bureau (Bureau du recensement des États-Unis). Courriel : timothy.l.kennel@census.gov; Richard Valliant est chercheur émérite au Survey Research Center, à l'Institute for Social Research de l'Université du Michigan, Ann Arbor, MI. Courriel : valliant@umich.edu.

Les estimateurs sandwich ajustés pour les effets de levier constituent une autre méthode d'estimation des erreurs d'échantillonnage fondées sur le plan de sondage ayant également des justifications fondées sur un modèle. Royall et Cumberland (1978) ont appliqué cette méthode pour concevoir des estimateurs de la variance de prédiction d'estimateurs des totaux de population finie. À partir d'un cadre fondé sur un modèle, Long et Ervin (2000) et MacKinnon et White (1985) ont montré comment l'estimateur sandwich pouvait servir à l'estimation de la variance pour des estimateurs de paramètres de régression, même lorsque la composante variance du modèle de travail était spécifiée de façon erronée. Valliant (2002) a adopté cette méthode pour estimer la variance fondée sur le plan d'estimateurs GREG à un degré d'échantillonnage. Le présent article prolonge le travail de Valliant en l'appliquant aux plans d'échantillonnage en grappes.

À la section 2, nous présentons l'estimateur GREG et plusieurs estimateurs de la variance de substitution. Tous les calculs sont en annexe. À la section 3, nous montrons les performances des nouveaux estimateurs de la variance dans plusieurs simulations. À la section 4, nous synthétisons nos constatations par une conclusion.

2 Résultats théoriques

Supposons une population ayant $i = 1, 2, \ldots, M$ grappes. Dans la grappe i, il y a N_i éléments de sorte qu'il y a $N = \sum_{i=1}^{M} N_i$ éléments dans la population. L'univers des grappes est exprimé par U et l'univers des éléments dans la grappe i est U_i. La variable d'analyse y_{ik} est associée à l'élément k de la grappe i. La population totale de y est $t_{Uy} = \sum_{i=1}^{M} \sum_{k=1}^{N_i} y_{ik}$. Chaque élément de population a également un vecteur p de variables auxiliaires, $\mathbf{x}_{ik}$, qui peut être utilisé dans l'estimation. On sélectionne un échantillon à deux degrés sans remise aux premier et deuxième degrés. La probabilité de sélection de la grappe i est π_i, et $\pi_{k|i}$ est la probabilité de sélection conditionnelle de l'élément k dans la grappe i. La probabilité globale de sélection de l'élément ik est $\pi_{ik} = \pi_i \pi_{k|i}$. Soit s l'ensemble de grappes d'échantillon et s_i l'ensemble d'éléments d'échantillon dans la grappe i. Le nombre de grappes d'échantillon est m tandis que le nombre d'éléments d'échantillon sélectionnés de la grappe d'échantillon i est n_i. La taille de l'échantillon total des éléments est $n = \sum_{i \in s} n_i$.

Dans le modèle de travail, supposons que $\mathbf{Y}_U$, le vecteur N des variables d'analyse, suit le modèle linéaire suivant :

$$\begin{aligned} E_\xi(\mathbf{Y}_U) &= \mathbf{X}\boldsymbol{\beta} \\ \operatorname{cov}_\xi(\mathbf{Y}_U) &= \boldsymbol{\Psi} \end{aligned} \tag{2.1}$$

où l'indice ξ désigne une espérance par rapport à un modèle; $\mathbf{X} = \left[\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_M^\top\right]^\top$ est la matrice $N \times p$ des variables auxiliaires et $\mathbf{X}_i$ est la matrice $N_i \times p$ des variables auxiliaires pour les éléments N_i dans la grappe i; et $\boldsymbol{\beta}$ est un vecteur de paramètre de longueur p. On suppose que les éléments des grappes sont corrélés tandis que les éléments des différentes grappes sont indépendants selon le modèle. Ainsi, la matrice de covariance $\boldsymbol{\Psi}$ est une matrice diagonale par blocs $N \times N$ avec des matrices diagonales

$\boldsymbol{\Psi}_i = [\psi_{ik}]_{N_i \times N_i}$. Une des principales caractéristiques des estimateurs de la variance que nous proposons est qu'il n'est pas nécessaire de connaître la forme particulière de ψ_{ik} pour construire les estimateurs de la variance. Les estimateurs de la variance proposés seront convergents, quelle que soit la forme de $\boldsymbol{\Psi}$.

Särndal et coll. (1992, chapitre 8) examinent trois estimateurs GREG différents pouvant être utilisés dans les échantillons en grappes. Tous trois dépendent des données disponibles. Considérons leur cas B, qui se produit lorsque des données au niveau de l'unité sont disponibles pour l'échantillon complet et que des totaux de contrôle sont disponibles pour la population. Dans ce cas, l'estimateur GREG est

$$\begin{aligned} \hat{t}_y^{gr} &= \hat{t}_{y\pi} + \hat{\mathbf{B}}^\top \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right) \\ &= \mathbf{g}^\top \boldsymbol{\Pi}^{-1} \mathbf{y}_s \end{aligned} \tag{2.2}$$

où $\mathbf{y}_s$ est le vecteur n des y pour les éléments d'échantillon, $\hat{t}_{y\pi}$ est l'estimateur π du total des y, $\mathbf{t}_{Ux}$ est le vecteur p des totaux de population des x, $\hat{\mathbf{t}}_{x\pi}$ est l'estimateur π de $\mathbf{t}_{Ux}$, et (si $\boldsymbol{\Psi}$ est connu) $\hat{\mathbf{B}} = \mathbf{A}^{-1}\mathbf{X}_s^\top \boldsymbol{\Psi}_s^{-1} \boldsymbol{\Pi}^{-1} \mathbf{y}_s$ avec $\mathbf{A} = \mathbf{X}_s^\top \boldsymbol{\Psi}_s^{-1} \boldsymbol{\Pi}^{-1} \mathbf{X}_s$, $\mathbf{X}_s$ la matrice des variables auxiliaires de l'échantillon, et $\boldsymbol{\Pi} = \text{diag}[\pi_{ik}]\,(i \in s, k \in s_i)$; $\boldsymbol{\Psi}_s$ est la partie de $\boldsymbol{\Psi}$ associée aux éléments d'échantillon; et $\mathbf{g}^\top = \mathbf{1}_n^\top + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1}\mathbf{X}_s^\top \boldsymbol{\Psi}_s^{-1}$ où $\mathbf{1}_n$ est un vecteur de n valeurs 1.

La composante du poids g de la grappe d'échantillon i est $\mathbf{g}_i^\top = \mathbf{1}_{n_i}^\top + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1}\mathbf{X}_{si}^\top \boldsymbol{\Psi}_{si}^{-1}$, $\mathbf{X}_{si}^\top = \left[\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right]$ étant la matrice $p \times n_i$ des variables auxiliaires pour les éléments d'échantillon dans la grappe d'échantillon i, $\boldsymbol{\Psi}_{si}$ est la partie $n_i \times n_i$ de $\boldsymbol{\Psi}_i$ pour les éléments d'échantillon dans la grappe d'échantillon i, et $\mathbf{1}_{n_i}$ est un vecteur de n_i valeurs 1. Puisque $\boldsymbol{\Psi}$ est généralement inconnu, une valeur de substitution $\mathbf{Q}$ peut être utilisée pour $\boldsymbol{\Psi}_s^{-1}$; $\mathbf{Q} = \mathbf{I}$ est un choix courant. Plus bas, nous supposons qu'une valeur générale $\mathbf{Q}$ est utilisée dans l'estimation par la régression généralisée plutôt que $\boldsymbol{\Psi}_s^{-1}$.

2.1 Estimateurs de la variance actuels

Särndal et coll. (1992, résultat 8.9.1) présentent un estimateur de la variance par rapport au plan $\hat{t}_y^{gr}$, qui comporte des probabilités de sélection conjointe des grappes et des éléments des grappes. En cas d'échantillonnage de Poisson, aux deux degrés, leur estimateur est

$$\upsilon_g = \sum_{i \in s} \frac{(1-\pi_i)}{\pi_i^2} \left(\hat{t}_{e,i}^{\,g}\right)^2 + \sum_{i \in s} \frac{1}{\pi_i} \sum_{k \in s_i} \frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2} g_{ik}^2 e_{ik}^2 \tag{2.3}$$

où $\hat{t}_{e,i}^{\,g} = \sum_{s_i} g_{ik} e_{ik} \big/ \pi_{k|i}$, g_{ik} est la composante k^{e} du vecteur $\mathbf{g}_i$, et $e_{ik} = y_{ik} - \mathbf{x}_{ik}^\top \hat{\mathbf{B}}$. Les calculs pour cet estimateur sont plus simples que la formule générale qui utilise des probabilités de sélection conjointe et peut avoir des performances satisfaisantes en cas de plans ptπ où l'on peut obtenir une approximation de la variance des estimateurs par des formules qui supposent une indépendance entre les sélections.

Voici un estimateur approprié si l'échantillonnage au premier degré est sélectionné avec remise :

$$\upsilon_{wr} = \frac{m}{m-1} \sum_{i \in s} \left(e_{1i} - \overline{e}_1\right)^2 \tag{2.4}$$

avec $e_{1i} = \sum_{k \in s_i} e_{ik} \big/ \pi_{ik}$ et $\overline{e}_1 = m^{-1} \sum_{i \in s} e_{1i}$. L'estimateur par linéarisation jackknife est (Yung et Rao, 1996)

$$\upsilon_{JL} = \frac{m-1}{m} \sum_{i \in s} (e_{2i} - \overline{e}_2)^2 \tag{2.5}$$

où $e_{2i} = \sum_{k \in s_i} g_{ik} e_{ik} \big/ \pi_{ik}$ et $\overline{e}_2 = m^{-1} \sum_{i \in s} e_{2i}$, g_{ik} étant la composante k^{e} du vecteur $\mathbf{g}_i$.

La méthode jackknife est une autre technique courante d'estimation de la variance. Krewski et Rao (1981) présentent plusieurs façons asymptotiquement équivalentes d'exprimer le jackknife. La forme suivante de l'estimateur jackknife constitue un point de départ pratique pour les calculs qui suivent :

$$\upsilon_{\text{Jack}} = \frac{m-1}{m} \sum_{i \in s} \left(\hat{t}_{y(i)}^{gr} - \hat{t}_{y(\cdot)}^{gr} \right)^2 \tag{2.6}$$

où $\hat{t}_{y(i)}^{gr}$ est la valeur de l'estimateur GREG après suppression de la grappe i et $\hat{t}_{y(\cdot)}^{gr}$ est la moyenne de toutes les estimations $\hat{t}_{y(i)}^{gr}$. L'utilisation de (2.6) peut exiger d'importantes ressources de calcul, car il faut calculer m estimations différentes de $\hat{t}_{y(i)}^{gr}$. Les estimateurs υ_{Jack}, υ_{wr} et υ_{JL} sont tous convergents par rapport au plan de sondage dans les conditions de Krewski et Rao (1981) et de Yung et Rao (1996). L'une de leurs principales conditions était que les grappes devaient être sélectionnées avec remise. Cette hypothèse simplifie les calculs théoriques, mais elle est utilisée seulement par souci de commodité. En effet, de nombreuses études empiriques ont démontré que les résultats théoriques étaient de bons prédicteurs de la performance des estimateurs dans les plans sans remise, tant que la fraction de sondage au premier degré est petite.

2.2 Nouveaux estimateurs de la variance

Nous utilisons le cadre fondé sur un modèle pour construire de nouveaux estimateurs de la variance. En premier lieu, nous calculons la variance fondée sur le modèle de $\hat{t}_y^{gr}$. Supposons que le modèle (2.1) se vérifie et que l'échantillonnage est ignorable, en ce sens que la probabilité qu'une unité soit dans l'échantillon donné $\mathbf{Y}_U$ et $\mathbf{X}$ dépend seulement de $\mathbf{X}$ (voir par exemple la discussion dans Valliant, Dorfman et Royall, 2000, section 2.6.2 et les références supplémentaires qui y sont citées). Ensuite, nous construisons des estimateurs de la variance du modèle, au moyen d'ajustements de la matrice chapeau pour tenir compte de l'hétérogénéité dans les données. Nous évaluons les propriétés fondées sur le plan de sondage des nouveaux estimateurs de la variance dans une simulation.

Pour calculer la variance du modèle de $\hat{t}_y^{gr}$, soit $\mathbf{y}_i$ le vecteur de population des variables d'analyse pour la grappe i, et $\mathbf{y}_{si}$ le vecteur des éléments d'échantillon. Comme le montre l'annexe A.2, sous le modèle (2.1), la variance fondée sur le modèle de $\hat{t}_y^{gr}$ est :

$$\begin{aligned} \text{var}_\xi \left(\hat{t}_y^{gr} - t_{Uy} \right) &= \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2 \sum_{i \in s} \left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \text{cov}_\xi (\mathbf{y}_{si}, \mathbf{y}_i) \, \mathbf{1}_{N_i} \right] + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N \\ &= L_1 - 2L_2 + L_3 \end{aligned}$$

où $\text{var}_\xi(\mathbf{y}_{si}) = \boldsymbol{\Psi}_{si}$, la partie de $\boldsymbol{\Psi}$ associée à des éléments dans s_i, et $\mathbf{1}_{N_i}$ et $\mathbf{1}_N$ sont des vecteurs de N_i et N 1.

La variance de l'erreur fondée sur le modèle de $\hat{t}_y^{gr}$ nécessite de connaître $\boldsymbol{\Psi}$ pour toute la population. En l'absence de solides hypothèses établissant un lien entre les structures de covariance de l'échantillon et hors de l'échantillon, les composantes de $\boldsymbol{\Psi}$ associées aux valeurs non échantillonnées ne peuvent pas être estimées à partir de l'échantillon. Cependant, comme le montre l'annexe A.2, dans certaines conditions raisonnables, les ordres des termes sont $L_1 = O\left(M^2/m\right)$ et $L_2 = L_3 = O(M)$, de sorte que L_1 domine la variance à mesure que le nombre de grappes d'échantillon et de population augmente. Ainsi,

$$\text{av}_\xi\left(\hat{t}_y^{gr} - t_{Uy}\right) = \sum_{i\in s} \mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1} \boldsymbol{\Psi}_{si} \boldsymbol{\Pi}_i^{-1} \mathbf{g}_i \tag{2.7}$$

où av_ξ désigne la variance du modèle asymptotique selon les hypothèses de l'annexe A.1. On peut former un estimateur robuste du deuxième membre de (2.7) même si $\boldsymbol{\Psi}_{si}$ est inconnu. En revanche, si le nombre de grappes de population augmente au même taux que les grappes d'échantillon (c'est-à-dire que $f = m/M$ converge vers une constante non nulle), alors L_1, L_2 et L_3 peuvent tous contribuer de façon importante à la variance asymptotique. Dans le présent article, nous examinerons uniquement l'estimation de L_1.

À moins que la vraie matrice de variance de $\mathbf{y}_s$ soit connue, il faut estimer $\boldsymbol{\Psi}_i$. À l'annexe A.3, nous montrons que dans les grands échantillons $\text{var}_\xi(\mathbf{e}_i) \approx \boldsymbol{\Psi}_i$, où $\mathbf{e}_i = \mathbf{y}_{si} - \hat{\mathbf{y}}_{si}$, avec $\hat{\mathbf{y}}_{si} = \mathbf{X}_{si}\hat{\mathbf{B}}$ et $\mathbf{X}_{si}$ étant la matrice $n_i \times p$ des variables auxiliaires pour les éléments d'échantillon dans la grappe d'échantillon i. Si on substitue $\mathbf{e}_i\mathbf{e}_i^\top$ à $\boldsymbol{\Psi}_{si}$ dans (2.7), on obtient l'estimateur sandwich

$$\upsilon_R = \sum_{i\in s} \mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \boldsymbol{\Pi}_i^{-1} \mathbf{g}_i. \tag{2.8}$$

D'après les résultats présentés à l'annexe A.3, υ_R est approximativement sans biais pour $\text{av}_\xi\left(\hat{t}_y^{gr} - t_{Uy}\right)$ dans les grands échantillons. Cet estimateur sandwich est aussi étroitement lié à l'estimateur par grappe ultime fondé sur le plan de sondage pour un plan dans lequel les grappes sont sélectionnées avec remise, qui est, à son tour, semblable à υ_g et υ_{JL} avec un échantillonnage avec remise. Par conséquent, υ_R possède des propriétés souhaitables fondées à la fois sur le plan et sur le modèle.

Dans les échantillons de taille petite à moyenne, υ_R présente un biais par rapport au modèle et sous-estime souvent la variance véritable. On peut ajuster la matrice chapeau pour le corriger. Comme on le montre l'annexe A.3,

$$\text{E}_\xi\left(\mathbf{e}_i\mathbf{e}_i^\top\right) = \text{var}_\xi(\mathbf{e}_i) = \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_{si}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i;\, i,\, j\in s} \mathbf{H}_{ij}\boldsymbol{\Psi}_{sj}\mathbf{H}_{ij}^\top \tag{2.9}$$

où $\mathbf{H}_{ij} = \mathbf{X}_{si}^\top \mathbf{A}^{-1}\mathbf{X}_{sj}\mathbf{Q}_j\boldsymbol{\Pi}_j^{-1}$ $(i,\ j = 1, \ldots, m)$, $\mathbf{Q}_j$ et $\boldsymbol{\Pi}_j$ étant les parties $n_j \times n_j$ de $\mathbf{Q}$ et $\boldsymbol{\Pi}$ étant associé à la grappe d'échantillon j. Comme dans Li et Valliant (2009) et Valliant (2002), on peut recueillir $\mathbf{H}_{ij}$ dans une matrice chapeau pondérée selon l'enquête :

$$\begin{aligned}\mathbf{H} &= \mathbf{X}_s\mathbf{A}^{-1}\mathbf{X}_s^{\top}\mathbf{Q}\boldsymbol{\Pi}^{-1} \\ &= \begin{bmatrix} \mathbf{X}_{s1}\mathbf{A}^{-1}\mathbf{X}_{s1}^{\top}\mathbf{Q}_1\boldsymbol{\Pi}_1^{-1} & \cdots & \mathbf{X}_{s1}\mathbf{A}^{-1}\mathbf{X}_{sm}^{\top}\mathbf{Q}_m\boldsymbol{\Pi}_m^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{sm}\mathbf{A}^{-1}\mathbf{X}_{s1}^{\top}\mathbf{Q}_1\boldsymbol{\Pi}_1^{-1} & \cdots & \mathbf{X}_{sm}\mathbf{A}^{-1}\mathbf{X}_{sm}^{\top}\mathbf{Q}_m\boldsymbol{\Pi}_m^{-1} \end{bmatrix}. \end{aligned} \tag{2.10}$$

Selon les hypothèses de l'annexe A.1, $\mathbf{H} = O\left(m^{-1}\right)$, ce qui permet de conclure que $\text{var}_{\xi}\left(\mathbf{e}_i\right) \approx \boldsymbol{\Psi}_{si}$. Les sous-matrices diagonales $\mathbf{H}_{ii}$ sont des matrices analogues aux effets de levier dans un échantillonnage à un degré. Dans une régression des moindres carrés ordinaires, le vecteur des valeurs prédites peut s'écrire $\hat{\mathbf{y}} = \mathbf{H}_{\text{MCO}}\mathbf{y}$ avec $\mathbf{H}_{\text{MCO}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$. Les effets de levier sont des diagonales de la matrice chapeau, $\mathbf{H}_{\text{MCO}}$, qui peuvent servir à corriger un petit biais d'échantillon dans $e_i^2 = \left(y_i - \hat{y}_i\right)^2$ comme estimateur de $\text{var}_{\xi}\left(y_i\right)$. Nous utilisons $\mathbf{H}_{ii}$ de façon analogue ci-dessous.

Pour tenir compte du fait que $\mathbf{e}_i\mathbf{e}_i^{\top}$ présente un biais par rapport au modèle pour les échantillons petits à moyens, nous apportons des ajustements de type levier à $\mathbf{e}_i\mathbf{e}_i^{\top}$. Si $\mathbf{Q} = \mathbf{I}$ et que l'échantillon est autopondéré (c'est-à-dire $\boldsymbol{\Pi} = c\mathbf{I}$ pour certains $0 < c < 1$), alors $\text{var}_{\xi}\left(\mathbf{e}_i\right) = \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_{si}$ (voir l'annexe A.3). Si on résout $\boldsymbol{\Psi}_{si}$ et le substitue dans (2.8), on obtient l'estimateur de la variance :

$$\upsilon_D = \sum_{i\in s}\mathbf{g}_i^{\top}\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^{\top}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i \tag{2.11}$$

qui, dans ce cas particulier, est aussi approximativement sans biais étant donné que $\mathbf{H}_{ii} = O\left(m^{-1}\right)$. Une des caractéristiques indésirables de υ_D est qu'il peut être négatif ou avoir des contributions négatives de certaines grappes si $\upsilon_{Di} = \mathbf{g}_i^{\top}\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^{\top}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i < 0$. Pour ces grappes, le remplacement de υ_{Di} par $\upsilon_{Ri} = \mathbf{g}_i^{\top}\boldsymbol{\Pi}_i^{-1}\mathbf{e}_i\mathbf{e}_i^{\top}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i$ permet d'obtenir un estimateur de la variance positif. Cet ajustement est utilisé dans la simulation de la section 3.

Aux annexes A.4 et A.5, nous montrons que l'estimateur de la variance jackknife peut être écrit exactement comme suit :

$$\upsilon_{\text{Jack}} = \frac{m-1}{m}\left[\sum_{i\in s}\left(D_i - \bar{D}\right)^2 - 2\sum_{i\in s}\left(D_i - \bar{D}\right)F_i + \sum_{i\in s}F_i^{\,2}\right] \tag{2.12}$$

où

$$\begin{aligned} F_i &= \left(G_i - \bar{G}\right) - \frac{1}{n}\left(K_i - \bar{K}\right) \\ D_i &= \mathbf{g}_i^{\top}\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i \\ K_i &= \left(\mathbf{1}_N^{\top}\mathbf{X}_U - m\mathbf{1}_{n_i}^{\top}\boldsymbol{\Pi}_i^{-1}\mathbf{X}_{si}\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right);\ \bar{K} = m^{-1}\sum_{i\in s}K_i \\ G_i &= \mathbf{1}_{n_i}^{\top}\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\left[\mathbf{H}_{ii}\mathbf{y}_{si} - \hat{\mathbf{y}}_{si}\right];\ \bar{G} = m^{-1}\sum_{i\in s}G_i \\ \mathbf{R}_i &= \mathbf{A}^{-1}\mathbf{X}_{si}^{\top}\mathbf{Q}_i\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i. \end{aligned}$$

Cette forme de υ_{Jack} réduit considérablement les calculs, puisqu'une seule estimation GREG est nécessaire, au lieu de m estimations. (Il va de soi qu'il peut être avantageux de recalculer l'estimation par l'estimation par la régression généralisée GREG pour chaque réplique jackknife si un ajustement de non-réponse élaboré influe sur la taille de la vraie variance.)

Dans les grands échantillons, on peut établir approximativement υ_{Jack} par :

$$\upsilon_{J1} = \frac{m-1}{m}\sum_{i\in s}\left(D_i - \bar{D}\right)^2 \tag{2.13}$$

ou par

$$\begin{aligned}\upsilon_{J2} &= \frac{m-1}{m}\sum_{i\in s} D_i^2 \\ &= \frac{m-1}{m}\sum_{i\in s}\mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i .\end{aligned} \tag{2.14}$$

Les estimateurs υ_{J1} et υ_{J2} sont des versions en grappes des approximations à un degré du jackknife dans Valliant (2002, équations (3.5), (3.6)).

Comme l'esquisse l'annexe A.6, $\upsilon_{\text{Jack}}, \upsilon_{JL}, \upsilon_{J1}, \upsilon_{J2}, \upsilon_D$ et υ_R équivalent tous asymptotiquement à $m \to \infty$. Comme υ_{Jack} et υ_{JL} sont convergents par rapport au plan de sondage, on peut s'attendre à ce que les autres estimateurs ci-dessus donnent de bons résultats sur des échantillons répétés quand la taille de l'échantillon au premier degré est grande et que le modèle (2.1) est approximativement correct. Il faut cependant garder en tête que la fraction d'échantillonnage des grappes doit être petite pour que les estimateurs construits à partir d'un échantillon au premier degré sans remise aient les mêmes performances que si l'échantillon avait été sélectionné avec remise.

Aucun de ces estimateurs de type sandwich ne comprend de facteurs de correction de la population finie. Ils peuvent par conséquent avoir tendance à surestimer la variance d'échantillonnage quand une grande proportion des grappes d'échantillon est sélectionnée. Pour tenir compte de cela, nous pouvons rajuster davantage tous les estimateurs de la variance de façon ponctuelle en multipliant les estimateurs de la variance par un facteur de correction de population finie, noté f_{pc}, tel qu'il a été élaboré par Kott (1988). Il en résulte les estimateurs ajustés suivants :

$$\begin{aligned}\upsilon_R^* &= f_{pc}\sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\mathbf{e}_i\mathbf{e}_i^\top\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i \\ \upsilon_D^* &= f_{pc}\sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i \\ \upsilon_{\text{Jack}}^* &= f_{pc}\frac{m}{m-1}\left[\sum_{i\in s}\left(D_i - \bar{D}\right)^2 - 2\sum_{i\in s}\left(D_i - \bar{D}\right)F_i + \sum_{i\in s}F_i^2\right] \\ \upsilon_{J1}^* &= f_{pc}\frac{m}{m-1}\sum_{i\in s}\left(D_i - \bar{D}\right)^2 \\ \upsilon_{J2}^* &= f_{pc}\sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i .\end{aligned}$$

Quand un échantillon aléatoire simple est sélectionné au premier degré, $f_{pc} = 1 - m/M$. D'après Kott (1988), une correction appropriée quand le premier degré est sélectionné avec des probabilités variables est $f_{pc} = 1 - m\sum_{i=1}^{M} p_i^2$ où p_i est la probabilité de tirage unique pour la grappe i, c'est-à-dire la probabilité que la grappe i soit sélectionnée dans un échantillon de taille 1.

3 Simulation

Nous avons réalisé une série d'études par simulations pour mettre à l'épreuve les performances des nouveaux estimateurs de la variance dans différentes populations. Dans chaque échantillon simulé, nous avons calculé les quantités énumérées dans le tableau 3.1. Pour évaluer les estimateurs de la variance, nous avons calculé la moyenne des estimations de la variance, comparé ces moyennes à l'erreur quadratique moyenne empirique, et calculé les probabilités de couverture de l'intervalle de confiance en fonction des différentes estimations de la variance. Le tableau 3.2 résume les plans d'échantillonnage des 18 études par simulations. La colonne intitulée Étiquette donne les titres qui seront utilisés dans les tableaux suivants. Les plans d'échantillonnage sont utilisés dans les trois populations décrites ci-dessous.

Tableau 3.1
Statistiques d'intérêt pour la simulation de variance de l'estimation GREG en grappes

Statistiques	Description
$\hat{t}_y^{\pi}$	Estimation du total à partir de l'estimateur de Horvitz-Thompson
$\hat{t}_y^{gr}$	Total estimé à partir de l'estimateur GREG
υ_E	Variance empirique
υ_g	Estimateur de la variance fondé sur le plan en supposant un échantillonnage de Poisson aux deux degrés de Särndal et coll. (1992) dans (2.3)
υ_{wr}	Estimateur de la variance avec remise dans (2.4)
υ_{JL}	Estimateur de la variance par linéarisation par la méthode du jackknife de Yung et Rao (1996) dans (2.5)
υ_R	Estimateur sandwich dans (2.8)
υ_D	Premier estimateur sandwich à la matrice chapeau ajustée dans (2.11)
υ_{Jack}	Estimateur de la variance par la méthode du jackknife dans (2.6)
υ_{J1}	Première approximation de l'estimateur de la variance par la méthode du jackknife dans (2.13)
υ_{J2}	Deuxième approximation de l'estimateur de la variance par la méthode du jackknife dans (2.14)
υ_R^*	Estimateur sandwich avec ajustement de la population finie
υ_D^*	Premier estimateur sandwich ajusté à la matrice chapeau avec correction de la population finie
υ_{Jack}^*	Estimateur de la variance jackknife par la méthode du jackknife avec correction de population finie
υ_{J1}^*	Première approximation par la méthode du jackknife avec correction de population finie
υ_{J2}^*	Deuxième approximation par la méthode du jackknife avec ajustement de population finie

3.1 Données

Nous avons effectué des simulations sur trois populations pour évaluer les performances fondées sur le plan des estimateurs de la variance dans différentes situations. Dans la première population, nous avons étudié les performances des estimateurs de la variance en cas de grande fraction de sondage au premier degré et d'échantillon de taille moyenne. La deuxième étude par simulations portait sur les performances

des estimateurs de la variance dans un jeu de données relativement compliqué et une petite taille d'échantillon au premier degré. La dernière étude par simulations montre les performances des estimateurs de la variance dans de grands échantillons.

Tableau 3.2
Plans des simulations pour trois populations

	Étiquette	Population	Échantillon au premier degré	m	Échantillon au deuxième degré	Nbre d'échantillons
1	EAS fixe	Troisième année	EASSR	25	$n_i = 5$	1 000
2	EAS fixe	Troisième année	EASSR	50	$n_i = 5$	1 000
3	EAS epsem	Troisième année	EASSR	25	$f_i = \frac{675}{2\,427}$	1 000
4	EAS epsem	Troisième année	EASSR	50	$f_i = \frac{675}{2\,427}$	1 000
5	PPT epsem	Troisième année	PPTSR	25	$n_i = 5$	1 000
6	PPT epsem	Troisième année	PPTSR	50	$n_i = 5$	1 000
7	EAS fixe	ACS	EASSR	3	$n_i = 9$	5 000
8	EAS fixe	ACS	EASSR	15	$n_i = 9$	5 000
9	EAS epsem	ACS	EASSR	3	$f_i = \frac{30\,430}{194\,329}$	5 000
10	EAS epsem	ACS	EASSR	15	$f_i = \frac{30\,430}{194\,329}$	5 000
11	PPT epsem	ACS	PPTSR	3	$n_i = 9$	5 000
12	PPT epsem	ACS	PPTSR	15	$n_i = 9$	5 000
13	EAS fixe	Simulée	EASSR	300	$n_i = 2$	1 000
14	EAS fixe	Simulée	EASSR	1 500	$n_i = 2$	100
15	EAS epsem	Simulée	EASSR	300	$f_i = \frac{60\,000}{195\,164}$	1 000
16	EAS epsem	Simulée	EASSR	1 500	$f_i = \frac{60\,000}{195\,164}$	100
17	PPT epsem	Simulée	PPTSR	300	$n_i = 3$	1 000
18	PPT epsem	Simulée	PPTSR	1 500	$n_i = 3$	100

3.1.1 Population d'élèves de troisième année

La première étude par simulations a utilisé la population d'élèves de troisième année de l'annexe B.6 de Valliant et coll. (2000). Ce jeu de données contenait les résultats en mathématiques de 2 427 élèves de troisième année dans 135 écoles. Le nombre relativement faible d'écoles de la population et le nombre assez constant d'élèves de chaque école faisaient de cette population un objet idéal pour l'étude d'échantillons avec de grandes fractions d'échantillonnage.

Au moyen de l'estimation par la régression généralisée (GREG), nous avons estimé la note moyenne en mathématiques des élèves de troisième année. Au total, nous avons sélectionné 1 000 échantillons dans chacun des six plans d'échantillonnage du tableau 3.2. Dans le premier plan d'échantillonnage, nous avons sélectionné 1 000 échantillons aléatoires simples sans remise (EASSR) dans 25 écoles. Dans chaque école

échantillonnée, nous avons sélectionné exactement cinq élèves par EASSR. Étant donné que le nombre d'élèves variait d'une école à l'autre, le plan d'échantillonnage a donné lieu à différentes probabilités inconditionnelles de sélection, mais à un échantillon fixe de 125 élèves. Le deuxième plan d'échantillonnage était semblable au premier, mis à part le fait que nous avons sélectionné 50 écoles. Parce que le choix de 50 des 135 écoles a donné lieu à une grande fraction de sondage au premier degré de 0,37, un facteur de correction de population finie était nécessaire. Les échantillons $m = 25$ et de 50 écoles peuvent tous deux être considérés comme étant de taille « moyenne ».

Dans le troisième plan d'échantillonnage, nous avons sélectionné 1 000 échantillons aléatoires simples dans 25 écoles sans remise. Au sein de chaque école échantillonnée, nous avons sélectionné des élèves à un taux constant de $\frac{675}{2\,427}$, ce qui a produit 1 000 échantillons avec des tailles aléatoires centrées autour de 125 élèves. Dans ce plan, chaque élève avait une probabilité de sélection inconditionnelle égale. Le quatrième plan d'échantillonnage était semblable au troisième, mis à part le fait que nous avons sélectionné 50 écoles. Les tailles d'échantillon étaient également aléatoires dans ce plan, avec une moyenne de 250 élèves. Comme les troisième et quatrième plans d'échantillonnage ont donné à chaque unité la même probabilité de sélection, ils sont intitulés EAS epsem (pour l'anglais *equal probability selection*, soit mécanisme d'échantillonnage avec probabilités égales) dans les tableaux suivants.

Dans le cinquième plan, nous avons sélectionné 1 000 échantillons dans 25 écoles avec des probabilités proportionnelles au nombre d'élèves de chaque école. Dans chaque école échantillonnée, nous avons sélectionné exactement cinq élèves, ce qui a donné 1 000 échantillons comprenant exactement 125 élèves chacun. Le sixième plan d'échantillonnage était semblable au cinquième, mis à part le fait que nous avons sélectionné 50 écoles. Nous avons sélectionné 1 000 échantillons de 250 élèves au moyen de ce plan. Les cinquième et sixième plans sont des plans d'échantillonnage avec probabilités égales (ou epsem). Comme les deuxième et quatrième plans d'échantillonnage, ce plan d'échantillonnage comportait également une grande fraction d'échantillonnage et justifiait la nécessité d'un facteur de correction de la population finie aux fins d'ajustement des estimateurs de la variance.

À partir de chaque échantillon, nous avons estimé les notes moyennes en mathématiques pour la population finie au moyen d'un estimateur GREG et en supposant que le nombre d'élèves de la population était connu. Le modèle auxiliaire visait à reproduire le modèle de régression linéaire en grappes de la section 9.6 de Valliant et coll. (2000). Les onze variables explicatives utilisées dans la modélisation des résultats en mathématiques de chaque élève étaient : une ordonnée à l'origine, le sexe (masculin ou féminin), l'origine ethnique (blanc/asiatique, noir, autochtone des États-Unis/autre ou hispanique), si la langue parlée à la maison est celle de l'examen (toujours, parfois/jamais), le type de collectivité (banlieue de petite ou grande ville), et inscription dans un établissement d'enseignement. On a divisé le total des résultats en mathématiques estimés au moyen de l'estimateur GREG par le nombre d'élèves de la population, soit 2 427, pour obtenir le résultat moyen. Le résultat moyen de la population est de 477,7. Pour l'ensemble de la population, la valeur de R au carré pour le modèle linéaire au niveau des élèves était de 0,9735, ce qui indique une relation linéaire très forte.

3.1.2 Population de l'Enquête sur les collectivités américaines (American Community Survey ou ACS)

La deuxième étude par simulations a utilisé les données du fichier sommaire 3 du recensement de 2000 et celles du fichier sommaire 2005 – 2009 de l'Enquête sur les collectivités américaines (ACS). Elle visait à estimer le nombre total de logements dans l'État américain de l'Alabama, selon le fichier sommaire de l'ACS. Les nombres des groupes d'îlots du recensement de 2000 ont été utilisés comme covariables dans le modèle auxiliaire.

Pour créer la population, on a d'abord extrait toutes les données sur les groupes d'îlots du fichier sommaire de l'ACS et du fichier sommaire 3 du recensement de 2000. On a ensuite fusionné les deux fichiers au niveau du groupe d'îlots. Les groupes d'îlots comptant 1 000 logements ou plus dans le recensement de 2000 ont été supprimés, car leurs caractéristiques différaient de celles de la majorité des îlots. Dans de nombreux plans d'échantillonnage, les unités de grande taille comme celles-ci seraient placées dans une strate à tirage complet distincte et ne contribueraient pas à la variance des estimations. On a également retiré les groupes d'îlots ayant connu une croissance extrême du nombre total de logements. Plus précisément, les groupes d'îlots comptant plus de 10 unités en plus du double du nombre du recensement de 2000 ont été supprimés.

Les grappes étaient définies comme des comtés et les groupes d'îlots étaient traités comme des unités. Le fait de traiter le groupe d'îlots comme une unité est motivé par la tâche commune consistant à sélectionner l'échantillon d'îlots, à en établir la liste, puis à utiliser les listes pour estimer le nombre total de logements dans la population finie.

Les grappes comptant moins de 10 groupes d'îlots ou plus de 120 groupes d'îlots ont été retirées de la base de sondage des grappes. En tout, il y avait 61 grappes (comtés) contenant un total de 2 051 groupes d'îlots et 1 109 499 logements dans le jeu de données vérifié. Au total, six comtés et 1 278 groupes d'îlots comprenant 1 030 471 logements ont été retirés du fichier de l'Alabama.

La figure 3.1 montre deux diagrammes de dispersion. Le premier graphique montre le nombre total de logements dans le groupe d'îlots déclaré dans le fichier sommaire de l'ACS comme une fonction du nombre de logements du recensement de 2000. Chaque point représente un des 2 051 groupes d'îlots de la population finie. La ligne diagonale est un lisseur non paramétrique, qui indique une relation forte entre les deux variables. Le graphique indique également des signes d'hétéroscédasticité parce que les points semblent s'éloigner à mesure que le nombre du recensement de 2000 augmente. Le deuxième diagramme montre les résidus obtenus par la régression du nombre de logements du recensement de 2000 sur le nombre de logements de l'ACS au moyen des moindres carrés ordinaires (MCO) représentés par rapport au nombre de logements de l'ACS. À mesure que le nombre de logements déclaré dans le fichier de l'ACS augmente, les prédictions du modèle semblent sous-estimer considérablement le nombre réel de logements. Cela semble indiquer un certain degré de non-linéarité dans la fonction moyenne. De plus, la variance est remarquablement hétéroscédastique.

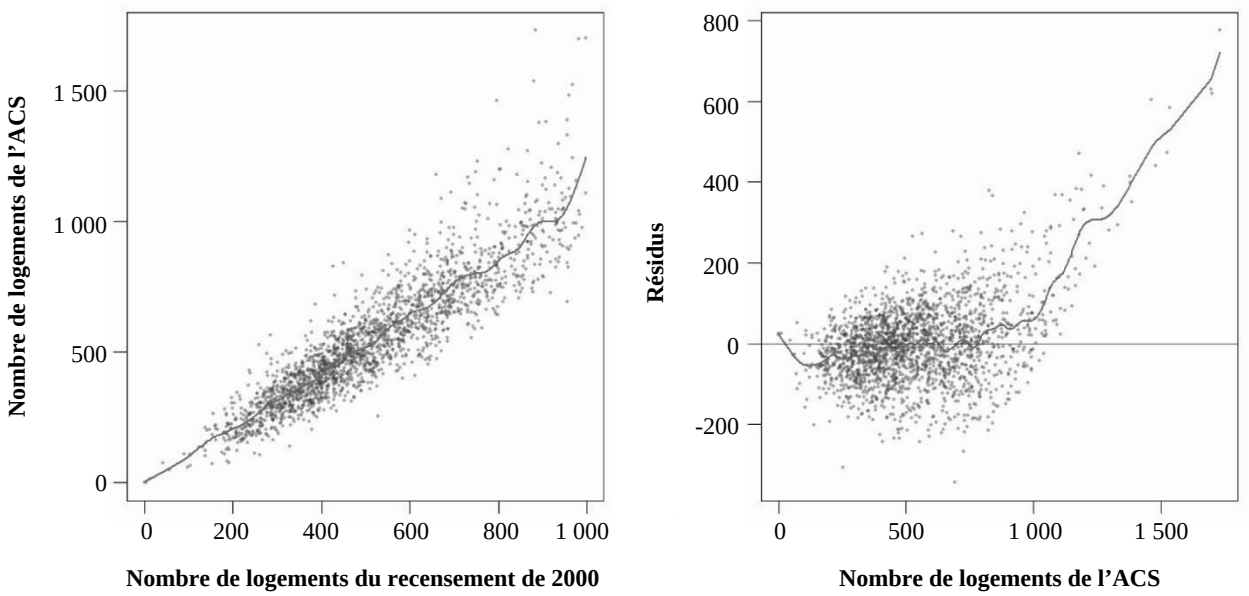


Figure 3.1 Diagramme de dispersion et graphique des résidus pour la population de l'ACS. Les lignes grises représentent des lisseurs non paramétriques.

Comme dans la première étude par simulations, nous avons essayé six plans d'échantillonnage différents. Nous avons sélectionné 5 000 échantillons dans chacun des six mécanismes de sélection indiqués au tableau 3.2. Dans le premier plan d'échantillonnage, nous avons sélectionné 5 000 échantillons aléatoires simples dans 3 grappes sans remise. Dans les grandes enquêtes nationales, il n'est pas rare de sélectionner un petit nombre d'unités primaires d'échantillonnage dans chaque strate. Dans ce cas, nous traitons l'Alabama comme une seule strate de plan d'échantillonnage et ses 61 comtés comme des grappes. Trois comtés de la strate ont été échantillonnés. Dans chaque grappe, nous avons sélectionné neuf groupes d'îlots au moyen d'un EASSR. Le deuxième plan était similaire, mais avec 15 grappes et 9 groupes d'îlots par grappe. Les deux premiers plans d'échantillonnage ont produit des pondérations très variables. Les autres plans (lignes 9 à 12) étaient parallèles à ceux des lignes 3 à 6 pour la population d'élèves de troisième année. Les tailles d'échantillon de $m = 3$ et 15 sont petites, si bien que les propriétés de grands échantillons théoriques sont moins susceptibles de se vérifier.

À partir de chaque échantillon, nous avons estimé le nombre total de logements dans la population finie à l'aide d'un estimateur GREG. Le modèle auxiliaire comprenait une ordonnée à l'origine et le nombre de logements du recensement de 2000; l'hétéroscédasticité mentionnée ci-dessus n'a pas été prise en compte dans l'estimation par la régression généralisée. Pour l'ensemble de la population, la valeur de R au carré était de 0,819, ce qui indique encore une fois une relation linéaire forte.

3.1.3 Population simulée

On a créé une population avec un grand nombre de grappes pour évaluer les caractéristiques asymptotiques des estimateurs de la variance. Produites à l'aide d'un modèle linéaire classique, 30 000 grappes ont été créées au total, chacune ayant un nombre aléatoire d'unités. On a déterminé le nombre d'unités de chaque grappe en ajoutant trois à un nombre entier aléatoire uniforme entre 0 et 7. La taille des grappes créées varie de 3 à 10 unités. Au total, la population contenait 195 164 unités dans 30 000 grappes. Pour chaque unité, on a créé une covariable positive en tant que $x_k \sim 1\,000 \exp N(0, 1)$ où $N(0, 1)$ est une variable aléatoire normale avec une moyenne de 0 et un écart-type de 1. On a créé une réponse aléatoire de sorte que $y_k \sim N\left(1\,000 + 2x_k, \frac{x_k}{2}\right)$. La figure 3.2 montre des diagrammes de dispersion de la relation entre x_k et y_k pour la population finie.

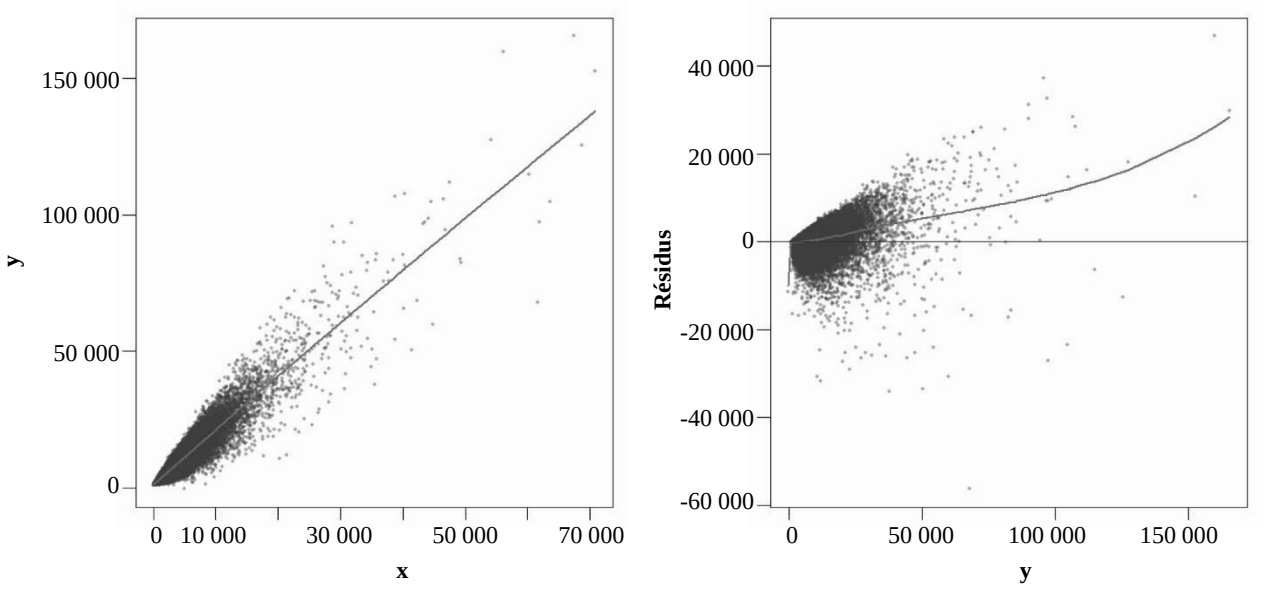


Figure 3.2 Diagramme de dispersion et résidus pour la population simulée. Les lignes grises représentent des lisseurs non paramétriques.

Nous avons sélectionné des échantillons au moyen des six différents mécanismes d'échantillonnage avec probabilités aux lignes 13 à 18 du tableau 3.2. Les types de plans d'échantillonnage sont parallèles à ceux utilisés pour les populations d'élèves de troisième année et de l'ACS. Dans les plans d'échantillonnage 14, 16 et 18, nous avons sélectionné 100 échantillons aléatoires simples de 1 500 grappes sans remise. Nous n'avons sélectionné que 100 échantillons, car le traitement et la sélection informatiques de chaque échantillon prenaient trop de temps. Étant donné que les tailles d'échantillon de $m = 300$ et 1 500 sont grandes, les propriétés de grands échantillons théoriques devraient se vérifier.

À partir de chaque échantillon, nous avons estimé le total de la réponse au moyen d'un estimateur GREG. La population réelle finie était de 839 149 969 personnes. Le modèle auxiliaire comprenait une ordonnée à

l'origine et x avec $\mathbf{Q} = \mathbf{I}$. Pour l'ensemble de la population, la valeur de R au carré était de 0,953, ce qui indique une relation linéaire très forte. La figure 3.2 présente un diagramme de dispersion de la population ainsi qu'un graphique des résidus basé sur une régression des moindres carrés ordinaires de x_k sur y_k pour l'ensemble de la population. Des éléments indiquent de manière probante l'hétéroscédasticité des erreurs.

3.2 Résultats

Nous avons examiné le biais, la variabilité et la couverture de l'intervalle de confiance des estimateurs de la variance nouveaux et anciens. Les tableaux présentent seulement certaines des simulations pour des questions d'espace. Le tableau 3.3 montre les moyennes de l'estimateur π et de l'estimateur GREG ainsi que les ratios des valeurs moyennes des estimateurs de la variance par rapport aux *erreurs quadratiques moyennes* empiriques pour toutes les populations et les combinaisons de taille d'échantillon dans toutes les simulations. L'estimateur π et l'estimateur GREG sont approximativement sans biais, mais l'estimateur GREG est beaucoup plus efficace.

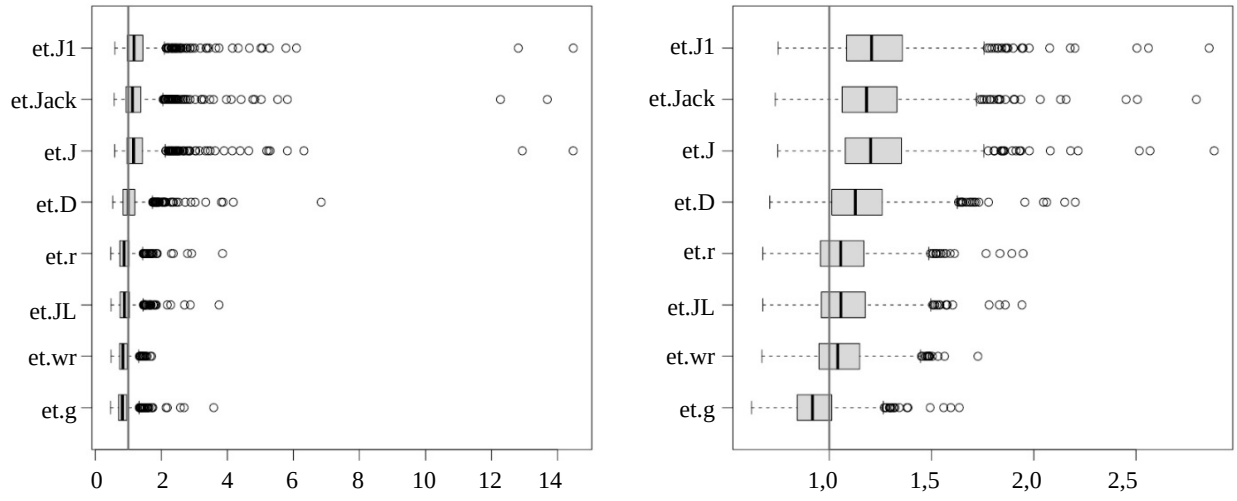
Tableau 3.3
Résultats de la simulation pour les estimations des moyennes et des estimateurs de la variance de trois populations et six plans d'échantillonnage dans chaque population. Les valeurs des lignes des estimateurs de la variance sont des ratios de la variance moyenne estimée par rapport à la *l'erreur quadratique moyenne* empirique de l'estimateur GREG. Voir la description des estimateurs de la variance dans le tableau 3.1

Estimateur	EAS fixe		EAS epsem		PPT epsem	
	Population des élèves de troisième année					
	$m = 25$	$m = 50$	$m = 25$	$m = 50$	$m = 25$	$m = 50$
moyenne $\hat{t}_y^{\pi} / N$	477,23	477,11	476,29	476,85	477,31	477,75
eqm $\hat{t}_y^{\pi} / N$	663,12	264,75	2 013,90	981,54	142,93	53,17
moyenne $\hat{t}_y^{g} / N$	474,27	476,37	476,95	477,24	477,50	477,85
eqm $\hat{t}_y^{g} / N$	218,96	66,66	114,08	50,10	121,57	41,32
$\upsilon_g / \text{eqm}(\hat{t}_y^{g})$	0,76	0,87	0,73	0,82	0,66	0,91
$\upsilon_{wr} / \text{eqm}(\hat{t}_y^{g})$	0,75	1,11	0,79	1,06	0,73	1,19
$\upsilon_{JL} / \text{eqm}(\hat{t}_y^{g})$	0,88	1,16	0,85	1,10	0,78	1,24
$\upsilon_R / \text{eqm}(\hat{t}_y^{g})$	0,87	1,15	0,82	1,08	0,74	1,22
$\upsilon_D / \text{eqm}(\hat{t}_y^{g})$	1,26	1,32	1,09	1,25	0,95	1,36
$\upsilon_{J2} / \text{eqm}(\hat{t}_y^{g})$	2,22	1,54	1,50	1,46	1,23	1,54
$\upsilon_{\text{Jack}} / \text{eqm}(\hat{t}_y^{g})$	2,03	1,49	1,44	1,43	1,19	1,51
$\upsilon_{J1} / \text{eqm}(\hat{t}_y^{g})$	2,22	1,55	1,56	1,49	1,28	1,57
$\upsilon_R^* / \text{eqm}(\hat{t}_y^{g})$	0,71	0,73	0,67	0,68	0,60	0,74
$\upsilon_D^* / \text{eqm}(\hat{t}_y^{g})$	1,02	0,83	0,88	0,79	0,76	0,83
$\upsilon_{J2}^* / \text{eqm}(\hat{t}_y^{g})$	1,81	0,97	1,22	0,92	0,99	0,93
$\upsilon_{\text{Jack}}^* / \text{eqm}(\hat{t}_y^{g})$	1,66	0,94	1,17	0,90	0,95	0,92
$\upsilon_{J1}^* / \text{eqm}(\hat{t}_y^{g})$	1,81	0,98	1,27	0,94	1,03	0,95

Tableau 3.3 (suite)
Résultats de la simulation pour les estimations des moyennes et des estimateurs de la variance de trois populations et six plans d'échantillonnage dans chaque population. Les valeurs des lignes des estimateurs de la variance sont des ratios de la variance moyenne estimée par rapport à *l'erreur quadratique moyenne* empirique de l'estimateur GREG. Voir la description des estimateurs de la variance dans le tableau 3.1

Estimateur	EAS fixe		EAS epsem		PPT epsem	
	Population de l'ACS (nombres en milliers)					
	$m = 3$	$m = 15$	$m = 3$	$m = 15$	$m = 3$	$m = 15$
moyenne $\hat{t}_y^{\pi} / N$	1 119,13	1 108,23	1 112,89	1 113,89	1 111,48	1 109,02
eqm $\hat{t}_y^{\pi} / N$	181 329,24	27 650,01	201 618,77	32 926,98	15 991,69	2 619,32
moyenne $\hat{t}_y^{g} / N$	1 081,68	1 103,34	1 104,45	1 108,45	1 106,36	1 108,46
eqm $\hat{t}_y^{g} / N$	11 220,86	921,82	2 111,84	408,19	1 874,39	352,65
$\upsilon_g / \text{eqm}(\hat{t}_y^{g})$	2,70	0,90	0,44	0,83	0,53	0,92
$\upsilon_{wr} / \text{eqm}(\hat{t}_y^{g})$	1,17	0,98	0,68	1,03	0,87	1,14
$\upsilon_{JL} / \text{eqm}(\hat{t}_y^{g})$	2,18	0,91	0,65	0,99	0,79	1,11
$\upsilon_R / \text{eqm}(\hat{t}_y^{g})$	2,80	1,00	0,43	0,92	0,53	1,03
$\upsilon_D / \text{eqm}(\hat{t}_y^{g})$	6,09	1,32	0,84	1,08	0,89	1,15
$\upsilon_{J2} / \text{eqm}(\hat{t}_y^{g})$	17 191,52	1,85	2,36	1,27	1,64	1,29
$\upsilon_{\text{Jack}} / \text{eqm}(\hat{t}_y^{g})$	4 678,25	1,47	1,37	1,19	1,05	1,21
$\upsilon_{J1} / \text{eqm}(\hat{t}_y^{g})$	17 190,86	1,72	3,07	1,36	2,35	1,38
$\upsilon_R^* / \text{eqm}(\hat{t}_y^{g})$	2,66	0,76	0,41	0,70	0,49	0,68
$\upsilon_D^* / \text{eqm}(\hat{t}_y^{g})$	5,79	0,99	0,80	0,82	0,83	0,76
$\upsilon_{J2}^* / \text{eqm}(\hat{t}_y^{g})$	16 346,03	1,40	2,25	0,96	1,52	0,85
$\upsilon_{\text{Jack}}^* / \text{eqm}(\hat{t}_y^{g})$	4 448,17	1,11	1,30	0,90	0,97	0,80
$\upsilon_{J1}^* / \text{eqm}(\hat{t}_y^{g})$	16 345,41	1,30	2,92	1,03	2,19	0,91
	Population simulée (nombres en millions)					
	$m = 300$	$m = 1\,500$	$m = 300$	$m = 1\,500$	$m = 300$	$m = 1\,500$
moyenne $\hat{t}_y^{\pi} / N$	838,91	838,71	838,13	843,13	838,74	839,06
eqm $\hat{t}_y^{\pi} / N$	1 588,43	250,20	2 303,19	563,77	1 218,73	253,13
moyenne $\hat{t}_y^{g} / N$	838,57	839,10	838,81	840,01	839,39	839,08
eqm $\hat{t}_y^{g} / N$	156,29	23,07	117,18	19,63	105,64	25,24
$\upsilon_g / \text{eqm}(\hat{t}_y^{g})$	0,91	1,11	0,91	1,13	1,01	0,89
$\upsilon_{wr} / \text{eqm}(\hat{t}_y^{g})$	0,94	1,13	0,91	1,17	1,01	0,90
$\upsilon_{JL} / \text{eqm}(\hat{t}_y^{g})$	0,91	1,13	0,92	1,15	1,02	0,90
$\upsilon_R / \text{eqm}(\hat{t}_y^{g})$	0,91	1,13	0,92	1,14	1,02	0,90
$\upsilon_D / \text{eqm}(\hat{t}_y^{g})$	1,03	1,15	0,96	1,16	1,07	0,91
$\upsilon_{J2} / \text{eqm}(\hat{t}_y^{g})$	1,50	1,17	1,03	1,18	1,13	0,93
$\upsilon_{\text{Jack}} / \text{eqm}(\hat{t}_y^{g})$	1,48	1,17	1,03	1,18	1,12	0,93
$\upsilon_{J1} / \text{eqm}(\hat{t}_y^{g})$	1,50	1,17	1,03	1,18	1,13	0,93
$\upsilon_R^* / \text{eqm}(\hat{t}_y^{g})$	0,90	1,07	0,91	1,09	1,01	0,85
$\upsilon_D^* / \text{eqm}(\hat{t}_y^{g})$	1,02	1,09	0,96	1,11	1,05	0,86
$\upsilon_{J2}^* / \text{eqm}(\hat{t}_y^{g})$	1,48	1,11	1,02	1,12	1,12	0,88
$\upsilon_{\text{Jack}}^* / \text{eqm}(\hat{t}_y^{g})$	1,47	1,11	1,01	1,12	1,11	0,88
$\upsilon_{J1}^* / \text{eqm}(\hat{t}_y^{g})$	1,48	1,11	1,02	1,13	1,12	0,88

Les performances des estimateurs de la variance dépendent du plan d'échantillonnage et de la population. Certaines des estimations du tableau 3.3 de la population de l'ACS avec un échantillon aléatoire simple de 3 grappes et 9 unités dans chaque grappe se démarquent comme étant très peu fiables. Les inverses des probabilités de sélection varient considérablement pour ce plan d'échantillonnage. La variabilité de ces pondérations, conjuguée à certaines observations extrêmes dans la population, cause l'instabilité de certains estimateurs de la variance. Pour être plus précis, $\upsilon_{J2}, \upsilon_{\text{Jack}}, \upsilon_{J1}, \upsilon^*_{J2}, \upsilon^*_{\text{Jack}}, \upsilon^*_{J1}$ sont des surestimations extrêmes en moyenne. Ces six estimateurs contiennent des ajustements explicites ou implicites de la matrice chapeau qui peuvent être assez grands et accroissent considérablement les estimateurs de la variance lorsqu'ils sont conjugués à de grands poids d'échantillonnage. En revanche, υ_D, qui a également une matrice chapeau ajustée, a des performances satisfaisantes pour toutes les populations et toutes les tailles d'échantillon. Il faut souligner le résultat selon lequel υ_D est une bien moindre surestimation de *l'erreur quadratique moyenne* dans la combinaison (ACS, EAS fixe, $m = 3$, $n_i = 9$) tandis que les autres estimateurs à la matrice chapeau ajustée sont des surestimations extrêmes. Les estimateurs υ_g, υ_{wr} et, dans une moindre mesure, υ_R et υ_{JL} tendent à des sous-estimations aux plus petites tailles d'échantillon dans les populations d'élèves de troisième et de l'ACS et pour tous les plans d'échantillonnage dans ces populations, mais ce problème diminue en cas d'échantillons de grande taille.



(A) Échantillons de taille $m = 25$ (b) Échantillons de taille $m = 50$

Figure 3.3 Diagrammes de quartiles des ratios d'estimations d'erreurs-types par rapport aux erreurs-types empiriques pour 1 000 échantillons aléatoires simples de la population d'élèves de troisième année. Lignes de référence verticales à 1.

Les diagrammes de quartiles de la figure 3.3 montrent mieux la variabilité des estimateurs pour les *échantillons aléatoires simples* de taille $m = 25$ et 50 de la population d'élèves de troisième année. Les diagrammes de quartiles représentent les erreurs-types (ET) estimées en tant que fraction de l'ET empirique pour les échantillons de chaque simulation. Un ratio de 1 signifie que la variance estimée est égale à la variance empirique. Certains échantillons donnent de grandes estimations de l'ET, mais la majorité des échantillons sont beaucoup plus près de la variance empirique. Le degré de surestimation et l'incidence des

valeurs extrêmes diminuent considérablement pour la plus grande taille d'échantillon, comme la comparaison des nombres le montre visiblement. Les estimateurs à la matrice chapeau ajustée ont également tendance à légèrement surestimer la variance véritable, comme en témoignent les rectangles déplacés au-dessus des lignes de référence tracées à 1. Cela peut constituer un avantage pour la couverture de l'intervalle de confiance.

Le tableau 3.4 présente les sommaires à six nombres des ratios des estimations de l'ET, $\sqrt{v}$, à la racine carrée de la variance empirique, $\sqrt{v_E}$, pour la population d'élèves de troisième année dans quatre des plans d'échantillonnage. Comme l'indique la valeur médiane des ratios de υ_{J2}, υ_{Jack}, υ_{J1}, υ^*_{J2}, υ^*_{Jack} et υ^*_{J1}, ils sont généralement centrés près des ET empiriques, mais ils peuvent avoir des valeurs extrêmement grandes dans certains échantillons qui influent sur leurs moyennes. (Le problème des valeurs aberrantes est encore plus prononcé dans la population de l'ACS, mais les détails n'en sont pas présentés ici.) Les estimateurs les moins touchés par les extrêmes sont υ_g, υ_{wr}, υ_{JL}, υ_R, υ_D, υ^*_R et υ^*_D. Cependant, les estimateurs qui incorporent les *corrections pour population finie* (CPF) sont souvent des sous-estimations, sauf en cas d'EAS et $m = 25$.

Tableau 3.4
Résumés à six nombres pour d'autres estimateurs d'erreurs-types pour la population d'élèves de troisième année dans quatre plans d'échantillonnage. v_E est la variance empirique dans les échantillons simulés. Voir la description des estimateurs de la variance dans le tableau 3.1

	$\sqrt{v}$	Distribution de $\sqrt{v}/\sqrt{\upsilon_E}$					
		Min.	**1er qu.**	**Médiane**	**Moyenne**	**3e qu.**	**Max.**
EAS $m = 25$	$\sqrt{\upsilon_g}$	0,46	0,71	0,82	0,86	0,96	3,59
	$\sqrt{\upsilon_{wr}}$	0,48	0,73	0,84	0,87	0,97	1,71
	$\sqrt{\upsilon_{JL}}$	0,48	0,75	0,88	0,92	1,03	3,75
	$\sqrt{\upsilon_R}$	0,47	0,74	0,87	0,92	1,02	3,85
	$\sqrt{\upsilon_D}$	0,53	0,84	1,00	1,08	1,20	6,84
	$\sqrt{\upsilon_{J2}}$	0,59	0,96	1,16	1,31	1,43	14,47
	$\sqrt{\upsilon_{\text{Jack}}}$	0,57	0,93	1,13	1,26	1,38	13,69
	$\sqrt{\upsilon_{J1}}$	0,59	0,97	1,17	1,32	1,44	14,48
	$\sqrt{\upsilon^*_R}$	0,42	0,67	0,79	0,83	0,92	3,48
	$\sqrt{\upsilon^*_D}$	0,48	0,76	0,90	0,97	1,08	6,17
	$\sqrt{\upsilon^*_{J2}}$	0,53	0,87	1,05	1,18	1,29	13,06
	$\sqrt{\upsilon^*_{\text{Jack}}}$	0,52	0,84	1,02	1,14	1,25	12,35
	$\sqrt{\upsilon^*_{J1}}$	0,54	0,88	1,06	1,19	1,30	13,07
EAS $m = 50$	$\sqrt{\upsilon_g}$	0,62	0,84	0,92	0,94	1,01	1,64
	$\sqrt{\upsilon_{wr}}$	0,67	0,95	1,04	1,06	1,15	1,73
	$\sqrt{\upsilon_{JL}}$	0,68	0,96	1,06	1,08	1,18	1,94
	$\sqrt{\upsilon_R}$	0,68	0,96	1,06	1,07	1,17	1,95
	$\sqrt{\upsilon_D}$	0,71	1,01	1,13	1,15	1,26	2,20
	$\sqrt{\upsilon_{J2}}$	0,75	1,08	1,20	1,24	1,35	2,88
	$\sqrt{\upsilon_{\text{Jack}}}$	0,74	1,06	1,18	1,22	1,33	2,79
	$\sqrt{\upsilon_{J1}}$	0,75	1,09	1,21	1,24	1,36	2,86
	$\sqrt{\upsilon^*_R}$	0,54	0,76	0,84	0,85	0,93	1,55
	$\sqrt{\upsilon^*_D}$	0,56	0,80	0,89	0,91	1,00	1,75
	$\sqrt{\upsilon^*_{J2}}$	0,59	0,86	0,95	0,98	1,07	2,29
	$\sqrt{\upsilon^*_{\text{Jack}}}$	0,58	0,84	0,94	0,97	1,06	2,22
	$\sqrt{\upsilon^*_{J1}}$	0,60	0,86	0,96	0,99	1,08	2,27

Tableau 3.4 (suite)
Résumés à six nombres pour d'autres estimateurs d'erreurs-types pour la population d'élèves de troisième année dans quatre plans d'échantillonnage. v_E est la variance empirique dans les échantillons simulés. Voir la description des estimateurs de la variance dans le tableau 3.1

	$\sqrt{v}$	Distribution de $\sqrt{v}/\sqrt{v_E}$					
		Min.	**1er qu.**	**Médiane**	**Moyenne**	**3e qu.**	**Max.**
PPT $m = 25$	$\sqrt{v_g}$	0,48	0,71	0,79	0,80	0,88	1,33
	$\sqrt{v_{wr}}$	0,51	0,76	0,84	0,84	0,92	1,30
	$\sqrt{v_{JL}}$	0,50	0,76	0,86	0,87	0,96	1,46
	$\sqrt{v_R}$	0,49	0,75	0,84	0,85	0,94	1,43
	$\sqrt{v_D}$	0,53	0,83	0,94	0,96	1,06	1,66
	$\sqrt{v_{J2}}$	0,59	0,94	1,06	1,09	1,21	2,15
	$\sqrt{v_{\text{Jack}}}$	0,57	0,92	1,04	1,07	1,18	2,10
	$\sqrt{v_{J1}}$	0,60	0,96	1,08	1,11	1,23	2,19
	$\sqrt{v_R^*}$	0,43	0,67	0,76	0,76	0,84	1,30
	$\sqrt{v_D^*}$	0,47	0,75	0,84	0,86	0,95	1,51
	$\sqrt{v_{J2}^*}$	0,52	0,84	0,95	0,98	1,08	1,90
	$\sqrt{v_{\text{Jack}}^*}$	0,51	0,82	0,93	0,96	1,06	1,86
	$\sqrt{v_{J1}^*}$	0,53	0,86	0,97	1,00	1,10	1,93
PPT $m = 50$	$\sqrt{v_g}$	0,72	0,88	0,95	0,95	1,01	1,28
	$\sqrt{v_{wr}}$	0,78	1,00	1,09	1,09	1,16	1,47
	$\sqrt{v_{JL}}$	0,81	1,01	1,11	1,11	1,19	1,52
	$\sqrt{v_R}$	0,80	1,00	1,09	1,09	1,18	1,50
	$\sqrt{v_D}$	0,84	1,06	1,15	1,16	1,25	1,64
	$\sqrt{v_{J2}}$	0,88	1,11	1,22	1,23	1,33	1,83
	$\sqrt{v_{\text{Jack}}}$	0,88	1,10	1,21	1,22	1,31	1,81
	$\sqrt{v_{J1}}$	0,89	1,13	1,23	1,24	1,34	1,85
	$\sqrt{v_R^*}$	0,62	0,78	0,85	0,85	0,92	1,16
	$\sqrt{v_D^*}$	0,65	0,82	0,90	0,90	0,97	1,28
	$\sqrt{v_{J2}^*}$	0,68	0,87	0,95	0,96	1,03	1,43
	$\sqrt{v_{\text{Jack}}^*}$	0,67	0,86	0,94	0,95	1,02	1,42
	$\sqrt{v_{J1}^*}$	0,69	0,88	0,96	0,97	1,04	1,44

Enfin, le tableau 3.5 montre la couverture de l'intervalle de confiance de 95 % pour tous les estimateurs fondés sur les distributions t. Cela signifie que nous avons calculé $\left[\hat{t}_y^{gr} - t_{0{,}975,\, m-1}\sqrt{v},\ \hat{t}_y^{gr} + t_{0{,}975,\, m-1}\sqrt{v}\ \right]$ où $t_{0{,}975,\, m-1}$ est le 97,5e percentile d'une distribution t avec $m-1$ degrés de liberté. Nous avons ensuite constaté la fréquence à laquelle la valeur vraie tombait en dessous, au-dessus et à l'intérieur de cette fourchette. En plus des nouveaux et des anciens estimateurs, le tableau 3.5 montre également la couverture de l'intervalle de confiance atteinte quand la variance empirique, v_E, a été utilisée pour former les intervalles de confiance. Idéalement, le total de la population doit se situer dans l'intervalle de confiance estimé à 95 % pour 95 % des échantillons. Le total réel doit être inférieur aux limites de confiance de 95 % pour 2,5 % des échantillons et supérieur aux limites de confiance pour le même pourcentage d'échantillons.

Les estimateurs par la méthode du jackknife v_D^*, v_{Jack}^* et v_{J2} donnent des taux de couverture supérieurs à ceux des autres estimateurs de la variance, car ils sont plus grands. Dans les petits échantillons, les estimateurs par la méthode du jackknife couvrent au-dessus du niveau nominal. Les estimateurs de la

variance classiques, υ_g, υ_{wr} et υ_{JL} donnent une couverture insuffisante dans un certain nombre de cas, bien que leur couverture ait presque toujours été supérieure à 90 %. Il faut noter que υ_g est généralement meilleur que υ_R en raison de l'ajustement de la matrice chapeau qui rend υ_D plus grand.

Les estimateurs de la variance qui intègrent des ajustements de matrice chapeau (υ_D, υ_{J2}, υ_{Jack} et υ_R^*) augmentent généralement les taux de couverture de l'intervalle de confiance par rapport aux autres choix. Cet avantage était particulièrement remarquable pour la population de l'ACS population où, par exemple, υ_{wr} couvre dans moins de 90 % des échantillons dans les combinaisons (υ_{Jack}^*, $m = 3$), (EAS epsem, $m = 3$), et (EAS epsem, $m = 15$). Bien qu'en principe, une CPF semble utile dans certaines combinaisons de population et de tailles d'échantillon, les intervalles de confiance fondés sur des estimateurs de la variance avec CPF ont des taux de couverture inférieurs à ceux sans CPF. Par exemple, dans l'ACS (EAS epsem, $m = 15$) les taux de couverture de υ_R^*, υ_D^*, υ_{J2}^*, υ_{Jack}^* et υ_{J1}^* vont de 86,1 à 90,6 % tandis que les versions sans CPF vont de 90,2 à 93,4 %.

Tableau 3.5
Couverture de l'intervalle de confiance de 95 % pour les totaux de population fondés sur des distributions *t* et d'autres estimateurs de la variance. Voir la description des estimateurs de la variance dans le tableau 3.1

Est. variance	Troisième année			ACS			Simulation			Troisième année			ACS			Simulation		
	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.
	EAS $m = 25$			EAS $m = 3$			EAS $m = 300$			EAS $m = 50$			EAS $m = 15$			EAS $m = 1\,500$		
υ_E	2,9	95,6	1,5	0,7	99,3	0	2,7	95,0	2,3	3,4	95,1	1,5	3,3	95,8	1,0	1,0	96,0	3,0
υ_g	7,4	90,7	1,9	2,4	97,3	0,4	4,3	93,5	2,2	5,9	92,8	1,3	6,6	92,3	1,0	1,0	95,0	4,0
υ_{wr}	7,0	90,5	2,5	9,2	88,8	2,0	3,9	92,8	3,3	4,1	95,0	0,9	7,5	91,0	1,5	1,0	96,0	3,0
υ_{JL}	5,5	93,2	1,3	6,5	92,1	1,4	4,4	93,4	2,2	3,3	96,1	0,6	7,2	91,4	1,4	1,0	95,0	4,0
υ_R	5,9	92,7	1,4	3,1	96,3	0,6	4,3	93,5	2,2	3,4	96,0	0,6	6,5	92,5	1,0	1,0	95,0	4,0
υ_D	3,8	95,4	0,8	1,6	98,0	0,4	3,7	94,2	2,1	2,4	97,1	0,5	5,1	94,3	0,6	1,0	95,0	4,0
υ_{J2}	1,7	98,0	0,3	0,6	99,3	0,1	3,6	94,4	2,0	2,0	97,7	0,3	3,9	95,7	0,4	1,0	95,0	4,0
υ_{Jack}	2,1	97,6	0,3	3,2	95,9	0,8	3,6	94,4	2,0	2,0	97,7	0,3	5,6	93,7	0,7	1,0	95,0	4,0
υ_{J1}	1,6	98,1	0,3	1,6	98,0	0,3	3,6	94,4	2,0	2,0	97,7	0,3	4,5	95,0	0,5	1,0	95,0	4,0
υ_R^*	8,6	89,4	2,0	3,4	96,0	0,7	4,4	93,4	2,2	7,8	89,8	2,4	9,5	88,5	2,0	1,0	95,0	4,0
υ_D^*	5,5	93,3	1,2	1,6	98,0	0,4	3,8	94,1	2,1	6,4	92,2	1,4	7,5	91,1	1,4	1,0	95,0	4,0
υ_{J2}^*	2,9	96,6	0,5	0,6	99,3	0,1	3,6	94,4	2,0	5,2	93,8	1	5,8	93,3	0,8	1,0	95,0	4,0
υ_{Jack}^*	3,7	95,7	0,6	3,4	95,7	0,9	3,6	94,4	2,0	5,5	93,4	1,1	7,9	90,6	1,6	1,0	95,0	4,0
υ_{J1}^*	2,7	96,9	0,4	1,7	97,9	0,4	3,6	94,4	2,0	5,0	93,9	1,1	6,6	92,3	1,1	1,0	95,0	4,0
	EAS epsem $m = 25$			EAS epsem $m = 3$			EAS epsem $m = 300$			EAS epsem $m = 50$			EAS epsem $m = 15$			EAS epsem $m = 1\,500$		
υ_E	1,7	96,2	2,1	0,0	99,9	0,1	2,4	94,7	2,9	2,3	95,5	2,2	1,1	97,1	1,8	3,0	94,0	3,0
υ_g	5,6	91,2	3,2	6,5	91,5	2,0	2,6	94,1	3,3	5,1	92,2	2,7	8,3	90,4	1,3	3,0	96,0	1,0
υ_{wr}	5,8	91,2	3,0	9,6	87,2	3,2	3,1	93,3	3,6	3,4	95,1	1,5	9,3	89,7	1,1	3,0	95,0	2,0
υ_{JL}	5,1	92,4	2,5	6,5	91,2	2,3	2,6	94,1	3,3	2,8	96,0	1,2	8,2	90,9	0,9	3,0	96,0	1,0
υ_R	5,2	92,3	2,5	8,4	88,3	3,3	2,6	94,1	3,3	2,9	95,7	1,4	8,8	90,2	1,0	3,0	96,0	1,0
υ_D	3,7	94,3	2,0	5,5	92,8	1,7	2,5	94,3	3,2	2,3	96,9	0,8	7,8	91,6	0,7	3,0	96,0	1,0
υ_{J2}	1,9	97,3	0,8	2,6	96,7	0,7	2,3	94,9	2,8	2,0	97,9	0,1	6,9	92,6	0,5	3,0	96,0	1,0
υ_{Jack}	2,2	96,8	1,0	4,7	94,0	1,3	2,3	94,9	2,8	2,1	97,8	0,1	7,3	92,1	0,6	3,0	96,0	1,0
υ_{J1}	1,8	97,5	0,7	2,5	96,9	0,6	2,3	94,9	2,8	2,0	97,9	0,1	6,2	93,4	0,4	3,0	96,0	1,0
υ_R^*	6,6	89,5	3,9	8,9	87,8	3,4	2,7	93,9	3,4	7,7	88,7	3,6	11,7	86,1	2,2	3,0	96,0	1,0
υ_D^*	5,1	92,5	2,4	5,7	92,4	1,9	2,5	94,3	3,2	6,0	91,6	2,4	10,6	88,0	1,5	3,0	96,0	1,0
υ_{J2}^*	3,4	94,9	1,7	2,8	96,5	0,7	2,3	94,9	2,8	4,6	93,7	1,7	9,2	89,7	1,1	3,0	96,0	1,0
υ_{Jack}^*	3,5	94,8	1,7	4,9	93,7	1,4	2,3	94,9	2,8	4,7	93,3	2	9,9	89,0	1,2	3,0	96,0	1,0
υ_{J1}^*	3,0	95,4	1,6	2,6	96,8	0,6	2,3	94,9	2,8	4,6	93,7	1,7	8,6	90,6	0,8	3,0	96,0	1,0

Tableau 3.5 (suite)
Couverture de l'intervalle de confiance de 95 % pour les totaux de population fondés sur des distributions *t* et d'autres estimateurs de la variance. Voir la description des estimateurs de la variance dans le tableau 3.1

Est. variance	Troisième année			ACS			Simulation			Troisième année			ACS			Simulation		
	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.	Inf.	Moy.	Sup.
	PPT m = 25			PPT m = 3			PPT m = 300			PPT m = 50			PPT m = 15			PPT m = 1 500		
v_E	1,7	95,9	2,4	0	100,0	0,0	2,9	94,2	2,9	2,3	95,3	2,4	0,7	98,0	1,3	2,0	95,0	3,0
v_g	6,2	90,0	3,8	4,7	94,3	1,0	2,9	93,9	3,2	3,1	94,1	2,8	5,1	94,4	0,5	2,0	92,0	6,0
v_{wr}	5,1	91,1	3,8	5,6	92,8	1,5	3,1	93,6	3,3	2,0	97,0	1,0	5,3	94,3	0,4	3,0	92,0	5,0
v_{JL}	4,9	92,0	3,1	4,9	93,5	1,5	2,9	94,0	3,1	1,9	96,9	1,2	4,9	94,7	0,3	2,0	92,0	6,0
v_R	5,3	91,5	3,2	7,2	90,5	2,3	2,9	93,9	3,2	2,0	96,8	1,2	5,6	94,1	0,4	2,0	92,0	6,0
v_D	3,8	94,1	2,1	4,4	94,4	1,1	2,7	94,7	2,6	1,7	97,4	0,9	4,8	94,9	0,3	2,0	92,0	6,0
v_{J2}	2,7	96,1	1,2	2,6	97,0	0,4	2,6	95,0	2,4	1,6	97,9	0,5	4,3	95,5	0,2	2,0	92,0	6,0
v_{Jack}	2,8	95,8	1,4	4,2	94,9	0,9	2,6	95,0	2,4	1,6	97,9	0,5	4,7	95,1	0,2	2,0	92,0	6,0
v_{J1}	2,2	96,7	1,1	2,1	97,5	0,4	2,6	95,0	2,4	1,5	98,0	0,5	3,9	96,0	0,1	2,0	92,0	6,0
v_R^*	7,4	87,8	4,8	7,6	90,0	2,4	2,9	93,9	3,2	5,0	90,6	4,4	8,9	89,8	1,3	2,0	92,0	6,0
v_D^*	5,3	91,6	3,1	4,7	94,0	1,3	2,7	94,5	2,8	4,1	92,2	3,7	8,1	90,9	1,0	2,0	92,0	6,0
v_{J2}^*	3,6	94,3	2,1	2,8	96,8	0,4	2,6	95,0	2,4	3,0	94,1	2,9	7,2	92,0	0,7	2,0	92,0	6,0
v_{Jack}^*	4,0	93,7	2,3	4,5	94,5	1,0	2,6	95,0	2,4	3,1	94,0	2,9	7,9	91,1	1,0	2,0	92,0	6,0
v_{J1}^*	3,5	94,6	1,9	2,2	97,4	0,4	2,6	95,0	2,4	2,9	94,4	2,7	6,8	92,6	0,6	2,0	92,0	6,0

Une des caractéristiques de v_D et v_D^* est que les contributions propres aux grappes, $v_{D,i}$ et $v_{D,i}^*$, ainsi que les estimations de la variance globales peuvent être négatives. Dans les simulations, on a utilisé l'ajustement décrit après (2.11) pour éviter les contributions négatives. Les estimations négatives étaient plus courantes quand les tailles d'échantillon au deuxième degré étaient petites et que les pondérations étaient très variables. Par exemple, pour la population de l'ACS, près de 28 % des échantillons aléatoires simples de 3 grappes et $m_i = 9$ ont donné lieu à au moins une contribution de variance négative pour une grappe. Plus souvent, environ 10 % des échantillons contenaient au moins une estimation de la variance négative pour une grappe. Dans la population des élèves de troisième année, de 16 % à 27 % des échantillons avaient au moins une valeur négative de v_{Di}. Dans la population simulée ayant de grandes tailles d'échantillon, la valeur de v_{Di} était négative dans moins de 5 % des échantillons. La correction ponctuelle consistant à modifier $I_i - H_{ii}$ en I_i, v_D est un des estimateurs de la variance les plus attrayants, car il a tendance à surestimer légèrement la variance empirique, a une des meilleures couvertures d'intervalle de confiance et a une variabilité raisonnable comparativement à d'autres estimateurs de la variance.

4 Conclusion

Il a été démontré que les ajustements d'effets de levier des estimateurs standards de la variance réduisent le biais et améliorent la couverture de l'intervalle de confiance fondée sur les estimateurs par régression généralisée dans les échantillons à un degré. Le présent article étend ces résultats à des échantillons à deux degrés en présentant de nouveaux ajustements fondés sur des matrices chapeaux. Notre théorie justifie les ajustements et illustre que certains estimateurs proposés sont liés au jackknife avec suppression de grappe, qui est une procédure commune dans l'estimation par sondage.

Pour mettre à l'épreuve la théorie, nous avons mené une série d'études par simulations sur trois populations conçues pour évaluer le rendement dans des situations diverses. Pour ce, nous avons utilisé une grande fraction de sondage d'unités au premier degré dans une population d'âge scolaire. Dans une deuxième population, constituée à partir des données de l'Enquête sur les collectivités américaines (ACS), nous avons mis à l'épreuve les effets des petites tailles d'échantillon. Dans une troisième population simulée, nous avons examiné les performances d'un grand échantillon. Nous avons employé à la fois un échantillonnage aléatoire simple et un échantillonnage avec probabilités proportionnelles à la taille des grappes.

Les relations des estimateurs de la variance étaient semblables dans tous les plans d'échantillonnage. L'estimateur de la variance avec remise, υ_{wr}, qui est le choix par défaut dans les progiciels pour données d'enquête, l'estimateur par linéarisation jackknife, υ_{JL}, et l'estimateur de la variance fondé sur le plan, υ_g, qui suppose un échantillonnage de Poisson à chaque degré pour faciliter les calculs, présentent souvent un biais négatif, ce qui entraîne des intervalles de confiance au taux de couverture inférieur au taux souhaité. Certains estimateurs liés au jackknife – υ_{Jack}, υ_{J1} et υ_{J2} – qui comprennent explicitement ou implicitement des ajustements de matrice chapeau, ont tendance à produire de grandes valeurs aberrantes quand l'échantillon au premier degré est petit. Cela est particulièrement vrai quand le premier degré est sélectionné par EAS, mais moins dans l'échantillonnage avec PPT quand une mesure de taille efficace est utilisée.

Les estimateurs de la variance proposés ici, en particulier υ_D, offrent des solutions de rechange à l'estimation de la variance des estimateurs GREG dans des échantillons complexes. Au détriment d'une légère inflation de la variabilité de l'estimateur de la variance, les estimateurs sandwich à la matrice chapeau ajustée, notés ici par v_D, v_{J1} et v_{J2}, donnent une couverture de l'intervalle de confiance plus proche de la valeur nominale dans les échantillons petits à moyens. Selon le plan d'échantillonnage et les caractéristiques de la population, les estimateurs à la matrice chapeau ajustée peuvent produire des estimations de la variance moins biaisées et de meilleures inférences comparativement aux méthodes standards.

Remerciements

Les auteurs remercient le rédacteur associé et deux examinateurs, dont les commentaires ont considérablement amélioré l'article.

Annexe

Résultats théoriques

A.1 Hypothèses

Voici les hypothèses utilisées pour l'obtention de résultats asymptotiques. Le nombre de populations et de grappes d'échantillons tend vers l'infini. Cependant, le nombre de grappes de population augmente plus rapidement que le nombre de grappes d'échantillon. Certaines quantités de population sont supposées bornées.

A.1.1 $m/M \to 0$ *quand* $m \to \infty$ *et* $M \to \infty$.

A.1.2 *Tous les* N_i *et* n_i *sont bornés.*

A.1.3 $\pi_{ik} = O(m/M)$ *pour tous les* ik.

A.1.4 *Tous les éléments de* $\mathbf{X}$, $\mathbf{\Psi}$ *et* $\mathbf{Q}$ *sont bornés.*

A.1.5 *Le plan d'échantillonnage est tel que* $\frac{\sqrt{m}}{M}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_{Ux}\right) \xrightarrow{d} N(0, \mathbf{V})$, *où* $\mathbf{V}$ *est une matrice définie positive* $p \times p$, *c'est-à-dire que* $\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_{Ux}\right) = O_p\left(M/\sqrt{m}\right)$.

Étant donné que $\mathbf{\Pi} = O\left(\frac{m}{M}\right)$ élément par élément et $\mathbf{A} = \mathbf{X}_s^\top \mathbf{Q}^{-1} \mathbf{\Pi}^{-1} \mathbf{X}_s$ peut être écrit comme la somme de termes n et que n_i est borné quand $m \to \infty$, $\mathbf{A} = O(M)$. Par définition, $\mathbf{g}_i^\top = \mathbf{1}_{n_i} + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1} \mathbf{X}_i^\top \mathbf{Q}_i$. Le second terme dans g_i est $O_p\left(m^{-1/2}\right)$. Par conséquent, $\mathbf{g}_i$ converge vers un vecteur de valeurs 1. Si on utilise $\mathbf{A} = O(M)$ ainsi que les hypothèses A.1.3 et A.1.4, $\mathbf{H}_{ij}$ est $O\left(m^{-1}\right)$ élément par élément.

A.2 Variation du modèle de l'estimateur GREG

Soit $\mathbf{y}_{si}$ le vecteur de tous les éléments d'échantillon dans la grappe i et soit $\mathbf{y}_i$ le vecteur de tous les éléments de la grappe i. La variance du GREG, en ce qui concerne le modèle de travail (2.1), est :

$$\begin{aligned}
\mathrm{var}_\xi\left(\hat{t}_y^{gr} - t_y\right) &= \mathrm{var}_\xi\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{y}_{si} - \sum_{i\in U} \mathbf{1}_{N_i}^\top \mathbf{y}_i\right) \\
&= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2\mathrm{cov}_\xi\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{y}_{si}, \sum_{i\in U} \mathbf{1}_{N_i}^\top \mathbf{y}_i\right) + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N.
\end{aligned}$$

Étant donné que $\sum_{i\in U} \mathbf{1}_i^\top \mathbf{y}_i = \sum_{i\in s} \mathbf{1}_i^\top \mathbf{y}_i + \sum_{i\in(U-s)} \mathbf{1}_i^\top \mathbf{y}_i$ et les éléments des différentes grappes ne sont pas corrélés, nous obtenons :

$$\begin{aligned}
\mathrm{var}_\xi\left(\hat{t}_y^{gr} - t_y\right) &= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2\sum_{i\in s}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathrm{cov}_\xi(\mathbf{y}_{si}, \mathbf{y}_i) \mathbf{1}_{N_i}\right] + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N \\
&= L_1 - 2L_2 + L_3.
\end{aligned}$$

Puisque $\mathbf{A}^{-1} = O\left(M^{-1}\right)$ et $\mathbf{g}_i$ et $\mathbf{\Psi}_{si}$ sont bornés, nous avons $L_1 = O\left(M^2/m\right)$. Étant donné que $\mathbf{\Psi}_{si}$ est borné, $\mathrm{cov}_\xi(\mathbf{y}_{si}, \mathbf{y}_i) = O(1)$ et $L_2 = O(M)$. L_3 est la somme des termes N. Puisque les valeurs N_i sont bornées, $L_3 = O(M)$. Ainsi, L_1 est le terme dominant de la variance de prédiction.

A.3 Démonstration de $\mathrm{var}_\xi\left(\mathbf{e}_i\right) \approx \mathbf{\Psi}_{si}$

Dans la présente section, pour simplifier la notation, nous omettons l'indice s dans $\mathbf{y}_{si}$, $\hat{\mathbf{y}}_{si}$ et $\mathbf{\Psi}_{si}$. Le résidu peut s'écrire en termes de matrice chapeau comme suit.

$$\begin{aligned}
\mathbf{e}_i &= \mathbf{y}_i - \hat{\mathbf{y}}_i \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j\neq i;\, j\in s} \mathbf{H}_{ij} \mathbf{y}_j
\end{aligned}$$

où $\mathbf{I}_{n_i}$ est la matrice d'identité $n_i \times n_i$. La variance du modèle de $\mathbf{e}_i$ est alors

$$\begin{aligned}\text{var}_\xi(\mathbf{e}_i) &= \text{var}_\xi\left[\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j\neq i}\mathbf{H}_{ij}\mathbf{y}_j\right] \\ &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\text{var}_\xi(\mathbf{y}_i)\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i}\mathbf{H}_{ij}\text{var}_\xi\left(\mathbf{y}_j\right)\mathbf{H}_{ij}^\top \\ &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top. \qquad \text{(A.1)}\end{aligned}$$

Comme on l'a indiqué plus haut, $\mathbf{H}_{ii} = O\left(m^{-1}\right)$. Alors, $\text{var}_\xi(\mathbf{e}_i) = \boldsymbol{\Psi}_i + O\left(m^{-1}\right)$.

Pour justifier υ_D, notons que le second terme de (A.1) peut s'écrire comme suit :

$$\sum_{j\neq i}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top - \mathbf{H}_{ii}\boldsymbol{\Psi}_i\mathbf{H}_{ii}^\top.$$

La somme sur l'échantillon en grappes complet est

$$\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \mathbf{X}_i\mathbf{A}^{-1}\left(\sum_{j\in s}\mathbf{X}_j^\top\mathbf{Q}_j\boldsymbol{\Pi}_j^{-1}\boldsymbol{\Psi}_j\boldsymbol{\Pi}_j^{-1}\mathbf{Q}_j\mathbf{X}_j\right)\mathbf{A}^{-1}\mathbf{X}_i^\top.$$

Dans le cas particulier de $\mathbf{Q}_j = \boldsymbol{\Psi}_j^{-1}$ et $\boldsymbol{\Pi}_i = c\mathbf{I}_{n_i}$ pour une constante $c \in (0, 1)$ (c'est-à-dire que l'échantillon est autopondéré), nous avons

$$\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = c^{-2}\mathbf{X}_i\mathbf{A}^{-1}\left(\sum_{j\in s}\mathbf{X}_j^\top\boldsymbol{\Psi}_j^{-1}\mathbf{X}_j\right)\mathbf{A}^{-1}\mathbf{X}_i^\top,$$

ainsi que $\mathbf{H}_{ii} = c\mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Psi}_i^{-1}$ et $\mathbf{A} = c^{-1}\mathbf{X}\boldsymbol{\Psi}^{-1}\mathbf{X}$. À partir de ces simplifications, nous obtenons $\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \mathbf{H}_{ii}\boldsymbol{\Psi}_i$. Si on substitue ce résultat dans (A.1) et qu'on simplifie, on a

$$\begin{aligned}\text{var}_\xi(\mathbf{e}_i) &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top \\ &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i. \qquad \text{(A.2)}\end{aligned}$$

Il s'agit de la base de l'ajustement de υ_R pour obtenir υ_D.

A.4 Démonstration de $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} - \mathbf{R}_i$ pour les échantillons en grappes

Dans la présente section, nous omettons l'indice s dans $\mathbf{X}_s$, $\mathbf{y}_s$, $\mathbf{X}_{si}$, $\mathbf{y}_{si}$, $\mathbf{X}_{s(i)}$ et $\mathbf{y}_{s(i)}$ pour simplifier la notation. L'indice (i) désigne la suppression de la i^{e} grappe du vecteur ou de la matrice de l'échantillon complet. Par exemple, $\hat{\mathbf{B}}_{(i)}$ est l'estimation de $\mathbf{B}$ fondée sur toutes les grappes d'échantillon sauf la grappe i soit

$$\hat{\mathbf{B}}_{(i)} = \left(\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}$$

où $\mathbf{W}_{(i)} = \mathbf{Q}_{(i)}\boldsymbol{\Pi}_{(i)}^{-1}$. Si nous utilisons le lemme 9.5.1 de Valliant et coll. (2000), nous obtenons

$$\hat{\mathbf{B}}_{(i)} = \left(\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}^{-1}\right)\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}.$$

Étant donné que $\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)} = \mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i^\top\mathbf{W}_i\mathbf{y}_i$ et $\hat{\mathbf{B}} = \mathbf{A}^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y}$, nous avons

$$\begin{aligned}
\hat{\mathbf{B}}_{(i)} &= \mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i^\top\mathbf{W}_i\mathbf{y}_i\right) \\
&\quad + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i\mathbf{W}_i\mathbf{y}_i\right) \\
&= \hat{\mathbf{B}} - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\hat{\mathbf{y}}_i \\
&\quad - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{H}_{ii}\mathbf{y}_i \\
&= \hat{\mathbf{B}} - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i.
\end{aligned}$$

Par conséquent, $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} - \mathbf{R}_i$.

A.5 Estimateur de la variance par la méthode du jackknife de GREG en grappes en termes de leviers

Nous simplifions maintenant l'estimateur de la variance par la méthode du jackknife avec suppression de grappe de GREG en grappes. Comme dans les sections A.3 et A.4, nous omettons l'indice s dans plusieurs termes pour simplifier la notation. Le total estimé après la suppression de la grappe i^{e} est défini comme étant

$$\begin{aligned}
\hat{t}_{y(i)}^{gr} &= \frac{m}{m-1}\hat{t}_{y(i)}^{\pi} + \left[\mathbf{t}_{Ux} - \frac{m}{m-1}\hat{\mathbf{t}}_{x(i)}^{\pi}\right]\hat{\mathbf{B}}_{(i)} \\
&= \frac{m\mathbf{1}_n^\top\boldsymbol{\Pi}^{-1}\mathbf{y}}{m-1} - \frac{m\mathbf{1}_{n_i}^\top\boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} + \left[\mathbf{1}_N^\top\mathbf{X}_U - \frac{m\mathbf{1}_n^\top\boldsymbol{\Pi}^{-1}\mathbf{X}}{m-1} + \frac{m\mathbf{1}_{n_i}^\top\boldsymbol{\Pi}_i^{-1}\mathbf{X}_i}{m-1}\right]\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) \\
&= \frac{m\mathbf{1}_n^\top\boldsymbol{\Pi}^{-1}\mathbf{y}}{m-1} - \frac{m\mathbf{1}_{n_i}^\top\boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} \\
&\quad + \frac{m}{m-1}\left(\mathbf{1}_N^\top\mathbf{X}_U - \mathbf{1}_n^\top\boldsymbol{\Pi}^{-1}\mathbf{X}\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) - \frac{1}{m-1}\left(\mathbf{1}_N^\top\mathbf{X}_U - m\mathbf{1}_{n_i}^\top\boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) \\
&= \frac{m}{m-1}\hat{t}_y^{gr} - \frac{m\mathbf{1}_{n_i}^\top\boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} - \frac{m}{m-1}\left(\mathbf{1}_N^\top\mathbf{X}_U - \mathbf{1}_n^\top\boldsymbol{\Pi}^{-1}\mathbf{X}\right)\mathbf{R}_i - \frac{1}{m-1}K_i.
\end{aligned}$$

L'ajout et la soustraction de $\frac{m}{m-1}\mathbf{1}_{n_i}^\top\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$ et une importante simplification donnent

$$\hat{t}_{y(i)}^{gr} = \frac{m}{m-1}\hat{t}_y^{gr} - \frac{m}{m-1}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{m}{m-1}G_i - \frac{1}{m-1}K_i.$$

La différence entre les estimations avec suppression d'une unité et la moyenne de ces estimations donne

$$\begin{aligned}\hat{t}_{y(i)}^{gr} - \hat{t}_{y(\cdot)}^{gr} &= -\frac{m}{m-1}\left(D_i - \bar{D}\right) + \frac{m}{m-1}\left(G_i - \bar{G}\right) - \frac{1}{m-1}\left(K_i - \bar{K}\right) \\ &= -\frac{m}{m-1}\left(D_i - \bar{D}\right) + \frac{m}{m-1}\left[\left(G_i - \bar{G}\right) - \frac{1}{m}\left(K_i - \bar{K}\right)\right].\end{aligned}$$

Soit $F_i = \left(G_i - \bar{G}\right) - m^{-1}\left(K_i - \bar{K}\right)$ qui donne la formule de υ_{Jack} dans l'équation (2.12). Puis, étant donné que $\mathbf{H}_{ii} = O\left(m^{-1}\right)$ et $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\mathbf{B}}$,

$$\begin{aligned}F_i &= \left(G_i - \bar{G}\right) - \frac{1}{m}\left(K_i - \bar{K}\right) \\ &\approx \left[-\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\hat{\mathbf{y}}_i + \frac{1}{m}\sum_{i\in s}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\hat{\mathbf{y}}_i\right] - \frac{1}{m}\left[-m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\hat{\mathbf{B}} + \sum_{i\in s}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\hat{\mathbf{B}}\right] \\ &= \mathbf{0}.\end{aligned}$$

Ainsi, $F_i = o(1)$, et υ_{Jack} dans (2.6) et (2.12) équivaut asymptotiquement à υ_{J1} dans (2.13).

Enfin, pour justifier υ_{J2} dans (2.14), nous écrivons υ_{J1} sous la forme du calcul

$$\upsilon_{J1} = \frac{m}{m-1}\left[\sum_{i\in s}\left(\mathbf{g}_i^\top U_i\mathbf{e}_i\right)^2 - \frac{1}{m}\left(\sum_{i\in s}\mathbf{g}_i^\top U_i\mathbf{e}_i\right)^2\right] \tag{A.3}$$

où $U_i = \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}$. Notons que la variance du modèle de D_i est

$$\begin{aligned}\text{var}_\xi\left(D_i\right) &= \text{var}_\xi\left(\mathbf{g}_i^\top U_i\mathbf{e}_i\right) \\ &= \mathbf{g}_i^\top U_i^\top \text{var}_\xi\left(\mathbf{e}_i\right) U_i\mathbf{g}_i.\end{aligned}$$

Puisque $U_i = O(M/m)$ et que la somme dans $\sum_{i\in s}\text{var}_\xi\left(D_i\right)$ contient des termes $n = m\bar{n}$, la variance de $\sum_{i\in s}\mathbf{g}_i^\top U_i\mathbf{e}_i$ est $O\left(M^2/m\right)$. Ensuite, on met à l'échelle υ_{J1} pour que la valeur soit appropriée pour une moyenne, le premier terme entre parenthèses dans (A.3) est $N^{-2}\sum_{i\in s}D_i^2 = O\left(m^{-1}\right)$. Puisque le second terme entre parenthèses a une espérance de modèle de 0 et une variance $O\left(m^{-1}\right)$, il converge en probabilité à 0, et υ_{J2} équivaut asymptotiquement à υ_{J1}.

A.6 Équivalence asymptotique des estimateurs de la variance

Dans la présente annexe, nous esquissons des arguments pour expliquer pourquoi plusieurs estimateurs de la variance sont asymptotiquement équivalents. En utilisant des arguments fondés sur le plan de sondage, Yung et Rao (1996, Annexe) ont montré que l'estimateur par linéarisation jackknife, υ_{JL}, pour l'estimation par la régression généralisée (GREG), équivaut asymptotiquement à l'estimateur convergent par rapport au plan, υ_{Jack}, dans des plans à plusieurs degrés stratifiés avec un grand nombre de strates et un nombre borné de grappes d'échantillon sélectionnées dans chaque strate. Si on utilise les conditions de régularité de Rao et Shao (1985), on peut étendre le résultat à des plans dans lesquels soit (i) le nombre de strates est grand et le nombre de grappes par strate est limité ou (ii) le nombre de strates est limité et le nombre de grappes d'échantillon par strate est grand, comme cela est le cas dans le présent article.

L'estimateur par linéarisation jackknife de la section 2 peut être étendu comme suit

$$N^{-2}\upsilon_{JL} = N^{-2}\sum_{i\in s}\mathbf{g}_i^{\top}\mathbf{\Pi}_i^{-1}\mathbf{e}_i\mathbf{e}_i^{\top}\mathbf{\Pi}_i^{-1}\mathbf{g}_i - N^{-2}m\left(m^{-1}\sum_{i\in s}\mathbf{g}_i^{\top}\mathbf{\Pi}_i^{-1}\mathbf{e}_i\right)^2. \quad \text{(A.4)}$$

Le premier terme dans (A.4) est égal à v_R. Parce que, dans certaines hypothèses raisonnables, $\mathbf{g}_i$ et $\mathbf{e}_i$ sont bornés, et $\mathbf{\Pi}_i^{-1} = O(M/m)$ selon les hypothèses A.1.2 et A.1.3, le premier terme dans (A.4) est $O(1/m)$. Le second terme est aussi $O(1/m)$, mais l'espérance du modèle de $\bar{\mathbf{e}}_2 = m^{-1}\sum_{i\in s}\mathbf{g}_i^{\top}\mathbf{\Pi}_i^{-1}\mathbf{e}_i$ est nulle tant que (2.1) se vérifie. Étant donné que $\bar{\mathbf{e}}_2$ est une moyenne, sa variance de modèle tend vers 0 quand $m \to \infty$. Ainsi, le second terme dans (A.4) converge en probabilité à 0 et $\upsilon_{JL} \approx \upsilon_R$.

À la section A.5, il a été démontré que υ_{Jack} et υ_{J1} sont asymptotiquement équivalents. Dans A.1.1-A.1.4, $\mathbf{H}_{ii} = O(m^{-1})$. Par conséquent, υ_{J2} et υ_D sont approximativement identiques à υ_R et $m \to \infty$. Ainsi, $\upsilon_{\text{Jack}} \approx \upsilon_{JL}$ par extension de Yung et Rao (1996), les deux étant convergents par rapport au plan de sondage. De plus, υ_{JL} équivaut asymptotiquement à $\upsilon_{J1}, \upsilon_{J2}, \upsilon_D$ et υ_R. Par conséquent, les autres estimateurs de la variance examinés ici ont tous des justifications fondées sur le modèle et sur le plan de sondage.

Bibliographie

Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75(4), 797-799.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.

Li, J., et Valliant, R. (2009). Matrice chapeau et effets de levier pondérés par les poids de sondage. *Techniques d'enquête*, 35, 1, 17-27. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10881-fra.pdf.

Long, J.S., et Ervin, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217-224.

MacKinnon, J.G., et White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305-325.

Rao, J.N.K., et Shao, J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.

Royall, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73(362), 351-358.

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.

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

Valliant, R. (2002). Estimation de la variance de l'estimateur de régression généralisée. *Techniques d'enquête*, 28, 1, 109-122. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2002001/article/6424-fra.pdf.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley Series in Probability and Statistics: Survey Methodology Section. New York: John Wiley & Sons, Inc.

Yung, W., et Rao, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 1, 23-31. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1996001/article/14388-fra.pdf.