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

Article

Modèle de régression semiparamétrique pour les données d'enquêtes complexes

par Zilin Wang et David R. Bellhouse

Décembre 2009

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2009
Vol. 35 N° 2 pp. 267-281
Statistique Canada, N° 12-001-X au catalogue

Modèle de régression semiparamétrique pour les données d'enquêtes complexes

Zilin Wang et David R. Bellhouse [1]

Résumé

Nous élaborons un modèle de régression semiparamétrique pour les enquêtes complexes. Dans ce modèle, les variables explicatives sont représentées séparément sous forme d'une partie non paramétrique et d'une partie linéaire paramétrique. Les méthodes d'estimation combinent l'estimation par la régression polynomiale locale non paramétrique et l'estimation par les moindres carrés. Nous élaborons également des résultats asymptotiques, tels que la convergence et la normalité des estimateurs des coefficients de régression et des fonctions de régression. Nous recourrons à la simulation et à des exemples empiriques tirés de l'Enquête sur la santé en Ontario de 1990 pour illustrer la performance de la méthode et les propriétés des estimations.

Mots clés : Enquête complexe ; estimation par domaine ; régression non paramétrique ; lissage.

1. Introduction

En pratique, nombre d'enquêtes sont utilisées pour étudier la relation entre une variable réponse et des variables explicatives, ainsi que pour construire des modèles prédictifs. Par conséquent, il est nécessaire de mettre au point des méthodes qui permettent d'appliquer les modèles de régression stochastiques à des données d'enquête. Alors que les méthodes de régression non paramétrique sont utilisées largement dans de nombreux domaines de la statistique, peu d'attention leur a été accordée dans celui des enquêtes complexes, à cause de la complexité de la structure des données. Étant donné la corrélation due au tirage de l'échantillon avec mise en grappes et les probabilités de sélection inégales, les données de ces enquêtes ne sont ni indépendantes ni identiquement distribuées. Les méthodes de régression non paramétrique standard sont donc souvent inappropriées pour l'analyse des données d'enquête par sondage.

Certains auteurs, par exemple Breidt et Opsomer (2000), Montanari et Ranalli (2005), et Zheng et Little (2004), se sont penchés sur l'élaboration de méthodes de régression non paramétrique applicables aux données d'enquête. Cependant, comme dans le cas de l'application classique des méthodes de régression, la plupart de ces travaux s'appuient sur des approches assistées par modèle pour estimer les grandeurs de population descriptives et les paramètres reliés à ces grandeurs. Dans le présent article, nous nous intéressons à l'application de méthodes de régression non paramétrique pour étudier la relation entre la variable réponse et les covariables, ainsi que la prédiction en utilisant l'information auxiliaire. Bellhouse et Stafford (2001) ont étendu une méthode de régression polynomiale locale pour procéder à une modélisation par régression souple des données d'enquête complexes. Toutefois, cet article traitait uniquement d'une fonction de régression non paramétrique simple. Ici, nous étendons l'étude au cas de plusieurs variables indépendantes, y compris les variables indicatrices qui figurent souvent dans l'analyse par régression des données d'enquête.

Nous considérons une fonction de régression semiparamétrique linéaire définie par $E(\mathbf{y} \mid \mathbf{X}, \mathbf{z}) = \mathbf{X}\boldsymbol{\beta} + G(\mathbf{z})$, où $G(\cdot)$ est une fonction arbitraire et $\boldsymbol{\beta}$ est un vecteur de paramètres de dimension p inconnu. Dans ce modèle de régression semiparamétrique, les variables explicatives sont représentées séparément dans deux parties, l'une non paramétrique et l'autre, linéaire paramétrique. Nous souhaitons estimer la forme fonctionnelle de la partie non paramétrique du modèle ainsi que les paramètres inclus dans la partie paramétrique. Nous plaçons les variables explicatives catégoriques et les variables continues pour lesquelles nous supposons qu'il existe une dépendance linéaire dans la partie paramétrique du modèle, $\mathbf{X}\boldsymbol{\beta}$, et une variable fournissant peu d'information sur la forme fonctionnelle dans la partie non paramétrique, $G(\mathbf{z})$. Non seulement ce modèle est dicté par la motivation a priori d'en faire un outil d'analyse des données et retient une fonction d'interprétation importante, mais il facilite aussi la résolution du problème de dimensionnalité élevée créé par les facteurs et certaines covariables, grâce à leur inclusion dans la partie paramétrique du modèle.

Un modèle semblable a été élaboré pour des données indépendantes et identiquement distribuées par Robinson (1988) et par Speckman (1988). Dans ces articles, l'estimation est effectuée en trois étapes. À la première étape, les moyennes de la variable réponse et des variables

1. Zilin Wang, Département de mathématiques, Université Wilfrid Laurier, Waterloo (Ontario) Canada, N2L 3C5. Courriel : zwang@wlu.ca ; David R. Bellhouse, Département de sciences statistiques et actuarielles, Université Western Ontario, Londres (Ontario) Canada, N6A 5B7. Courriel : bellhouse@stats.uwo.ca.

indépendantes paramétriques, sachant la variable non paramétrique, sont traitées comme une fonction de cette variable et lissées ; à la deuxième étape, les coefficients linéaires sont estimés par régression des résidus provenant de la variable réponse lissée sur les résidus provenant des covariables paramétriques lissées. Enfin, la différence entre la variable réponse et sa prédiction d'après le modèle de régression est lissée de façon semblable pour produire une estimation de la partie non paramétrique de la fonction de régression. Robinson (1988) et Speckman (1988) ont montré que les estimateurs résultants sont convergents à la racine carrée de n quand le modèle est correct et que les points de données sont indépendants et identiquement distribués. L'objectif de notre article est d'appliquer cette méthode de lissage à des données d'enquête tout en tenant compte d'un plan d'échantillonnage complexe.

Nous utilisons la méthode d'estimation par la régression polynomiale locale établie dans Bellhouse et Stafford (2001) pour effectuer le lissage durant le processus d'estimation. Un élément clé de l'exécution de cette méthode est le groupement par classe ou fenêtre (*binning*), qui découle des travaux de Bellhouse et Stafford (1999) sur l'estimation de la densité. Dans de nombreux ensembles de données d'enquête, une variable continue peut être naturellement groupée par classe ; par exemple, l'âge peut être enregistré comme l'âge au dernier anniversaire. En général, les classes ou fenêtres correspondent aux ensembles disjoints de valeurs d'une covariable continue et, par conséquent, peuvent être considérées comme des domaines. Au niveau de l'échantillon, nous estimons la moyenne de domaine de la variable d'intérêt en divisant la somme pondérée de la variable dans le domaine par la somme des poids dans le domaine. Dans Bellhouse et Stafford (2001), la variable réponse est groupée par classe en fonction des valeurs de la covariable et discrétisée, et les moyennes de domaine de la variable réponse sont lissées pour obtenir la fonction de régression. Quand la taille d'échantillon est grande et que le nombre de classes est relativement faible, les estimateurs basés sur le groupement par classe sont des fonctions des estimateurs de domaine dont les propriétés inférencielles peuvent être facilement établies d'après les résultats présentés dans Shao (1996) et dans Serfling (1980). L'un des avantages pratiques du groupement par classe est qu'il peut révéler l'information sur une tendance qui est obscurcie dans une enquête complexe et qui peut être importante si l'ensemble de données d'enquêtes complexes est très grand. Habituellement, il existe de multiples observations pour chaque ensemble de valeurs des covariables dans ces ensembles de données.

Un exemple qui illustre ces caractéristiques des données groupées par classe est tiré de l'Enquête sur la santé en Ontario. Cette enquête a été réalisée par Statistique Canada en 1990 auprès de 61 239 personnes vivant en Ontario, au Canada. Les données ont été obtenues au moyen d'un plan d'échantillonnage en grappes stratifié à deux degrés. Les strates correspondaient aux régions urbaines et rurales relevant de chacun des bureaux de santé de l'Ontario. Des secteurs de dénombrement ont été sélectionnés aléatoirement dans chaque strate et, de même, des ménages l'ont été dans chaque secteur de dénombrement. L'objectif de cette enquête est de mesurer l'état de santé des habitants de l'Ontario et de recueillir des données sur les facteurs de risque associés aux principales causes de mortalité et de morbidité dans la province. Dans notre exemple, nous examinons le poids de la personne en fonction de l'âge. Dans l'Enquête sur la santé en Ontario, l'âge déclaré est celui au dernier anniversaire. La mesure que nous utilisons ici comme substitut du poids est l'indice de masse corporelle (IMC), qui est calculé en divisant le poids exprimé en kilogrammes par le carré de la taille exprimée en mètres. L'IMC est l'un des indicateurs du degré d'obésité d'une personne. Normalement, un IMC inférieur à 18 est considéré comme une insuffisance pondérale et un IMC supérieur à 30, comme une indication d'obésité. L'IMC n'est une mesure appropriée que pour les personnes de 18 à 64 ans, à l'exception des femmes enceintes ou qui allaitent. Par conséquent, la taille de l'échantillon est réduite à 44 457 répondants admissibles qui peuvent être répartis entre 47 âges ou classes.

Le graphique de gauche de la figure 1 représente la tendance de l'indice de masse corporelle en fonction de l'âge. Il est facile de voir que le diagramme de dispersion semblable à un « nuage noir » masque la relation entre l'âge et l'indice de masse corporelle. Par contre, si nous calculons la moyenne de l'indice de masse corporelle à chaque point d'âge distinct et que nous représentons graphiquement les estimations moyennes par classe de l'indice de masse corporelle en fonction de l'âge, nous obtenons le graphique de droite de la figure 1. Il est évident qu'une moyenne groupée par classe fournit plus de renseignements visuels que les données brutes. Les grands ensembles de données peuvent non seulement donner lieu à des graphiques non informatifs, mais aussi rendre le calcul des estimations très fastidieux. Donc, il est naturel, dans l'analyse des données d'enquêtes complexes, de regrouper les données dans des domaines en fonction des valeurs distinctes d'une covariable discrétisée. En outre, les estimateurs résultant du groupement par classe sont des fonctions des estimateurs de domaine.

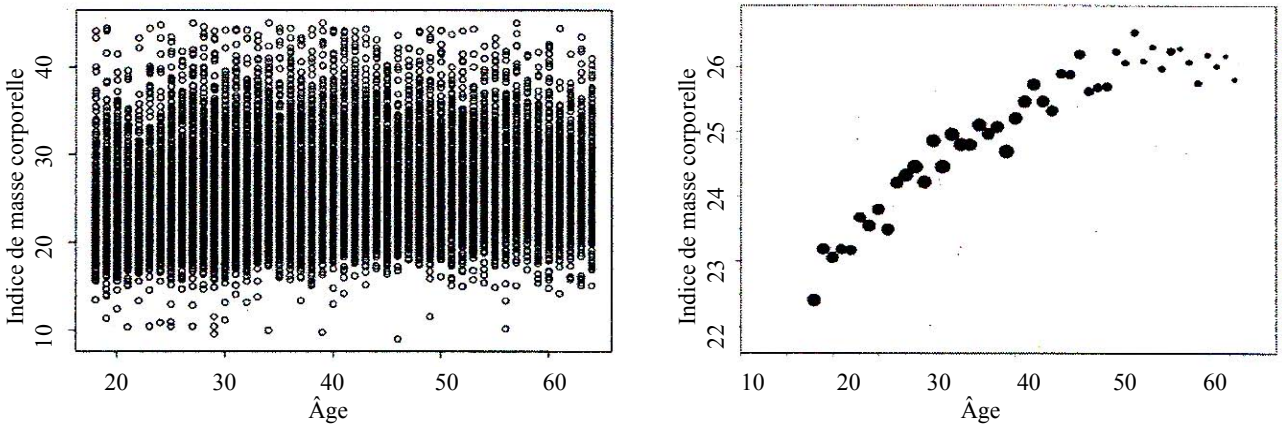

Figure 1 Comparaison des diagrammes de dispersion des données groupées et non groupées par classe provenant de l'Enquête sur la santé en Ontario

Un inconvénient du groupement par classe est que le nombre de classes ne peut pas croître asymptotiquement avec la population si les données sont groupées naturellement, comme cela est le cas de la variable d'âge dans l'exemple susmentionné. Dans de telles conditions, les estimateurs non paramétriques au niveau de la population demeurent entachés d'un biais en tant qu'estimateurs de fonctions de superpopulation, en raison de la taille de classe fixe. Dans notre cadre, nous supposons que les classes induites par les valeurs distinctes de la covariable sont les mêmes dans la population que dans l'échantillon ; de même, dans le lissage, nous supposerons que la largeur de fenêtre est la même au niveau de la population qu'au niveau de l'échantillon. Nous montrerons que les estimateurs sur échantillon sont des estimateurs convergents par rapport au plan des paramètres et fonctions de population finie correspondants, mais non de leurs analogues en superpopulation. Dans l'exemple des données de l'Enquête sur la santé en Ontario, le même ensemble d'âges distincts s'observe dans la population et dans l'échantillon.

La présentation de l'article est la suivante. À la section 2, nous introduisons les modèles de travail en superpopulation qui mènent aux méthodes d'estimation dans le cas de données d'enquête. À la section 3, nous calculons tous les moments des estimations obtenues et établissons certains résultats asymptotiques. Aux sections 4 et 5, nous présentons une étude par simulation et un exemple empirique de la méthode d'estimation appliquée en utilisant les données de l'Enquête sur la santé en Ontario de 1990 (1992). À la section 6, nous concluons par une discussion des hypothèses formulées et de certains futurs travaux. Les preuves de tous les lemmes et théorèmes présentés à la section 3 sont données en annexe.

2. Modèle de régression semiparamétrique et son estimation

Nous adoptons une approche typique de l'analyse des données d'enquêtes complexes. Pour commencer, nous considérons un modèle de travail appliqué à la population finie sous l'hypothèse que les observations sont indépendantes. Les estimations des paramètres du modèle deviennent alors les paramètres de population finie, ou paramètres sous recensement, qui doivent être estimés d'après l'échantillon. Une fois que nous avons défini les paramètres cibles de population finie, nous considérons un modèle hypothétique plus réaliste de la population finie afin d'obtenir des inférences au sujet de ces paramètres, ce que nous faisons à la section suivante. Considérons une population finie de taille N avec un vecteur de mesures $(y_k, \mathbf{x}_k, z_k)$ attaché à l'unité k, $k = 1, ..., N$, où y_k représente une observation de la variable réponse et $(\mathbf{x}_k, z_k)$ représente un vecteur d'observations des variables explicatives de dimension $p+1$. À titre de modèle de travail, nous imaginons que la variable réponse est produite par le modèle de régression linéaire partiel suivant

$$\mathbf{Y} = G(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

où $\mathbf{Y}$ est le vecteur de réponses et $\boldsymbol{\varepsilon}$ contient des entrées indépendantes et identiquement distribuées de moyenne nulle et de variance constante. La fonction $G(\cdot)$ est une fonction arbitraire de $\mathbf{z}$ et $\boldsymbol{\beta}$ est un vecteur de paramètres de dimension p inconnu. La matrice $\mathbf{X}$ de dimensions $N \times p$ correspond à la partie linéaire du modèle et contient des variables explicatives continues ou discrètes qui sont aléatoires. Le terme $G(\mathbf{z})$ est la partie non paramétrique du modèle. Nous supposons que z est non stochastique et mesurée sur une échelle continue, discrétisée en D valeurs

distinctes. En outre, nous imaginons que $E(\boldsymbol{\varepsilon} \mid \mathbf{z}, \mathbf{X}) = \mathbf{0}$. Il n'existe aucune interaction entre $\mathbf{X}$ et $\mathbf{z}$ dans le modèle.

Nous voulons estimer les versions au niveau de la population de $G(\cdot)$ et des paramètres $\boldsymbol{\beta}$. Nous commençons par élaborer des expressions pour ces entités, en nous inspirant des méthodes d'estimation décrites dans Robinson (1988) et dans Speckman (1988). En particulier, nous commençons par prendre l'espérance des deux membres de (1) sachant $\mathbf{z}$:

$$E(\mathbf{Y} \mid \mathbf{z}) = E(\mathbf{X} \mid \mathbf{z})\boldsymbol{\beta} + G(\mathbf{z}). \tag{2}$$

Puis, nous soustrayons (2) de (1) pour obtenir

$$\mathbf{Y} - E(\mathbf{Y} \mid \mathbf{z}) = \big(\mathbf{X} - E(\mathbf{X} \mid \mathbf{z})\big)\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

Pour définir la version de population de $\boldsymbol{\beta}$ dans (3), nous remplaçons $E(\mathbf{Y} \mid \mathbf{z})$ et $E(\mathbf{X} \mid \mathbf{z})$ dans (3) par leurs estimations au niveau de la population et nous estimons $\boldsymbol{\beta}$ par la méthode des moindres carrés.

Pour les estimations au niveau de la population de $E(\mathbf{Y} \mid \mathbf{z})$ et $E(\mathbf{X} \mid \mathbf{z})$, nous adoptons le lisseur à polynômes locaux de Jones (1989), dans lequel le groupement par classe est un élément essentiel de l'opération. Soit la variable discrétisée Z qui prend les valeurs $z_1, ..., z_D$; soit $\overline{\mathbf{Y}} = (\overline{Y}_1, ..., \overline{Y}_D)$ et $\overline{\mathbf{X}}_j = (\overline{X}_{j1}, ..., \overline{X}_{jD})$ pour $j = 1, ..., p$, respectivement, les vecteurs des moyennes dans les classes de $z_1, ..., z_D$. Soit aussi P_d la proportion dans la population des observations comprises dans la d^{e} classe pour $d = 1, ..., D$. Alors, désignons les espérances conditionnelles lissées de population de $\mathbf{Y}$ et $\mathbf{X}_j$ au point z_d par $m_{\mathbf{y}}(z_d)$ et $m_j(z_d)$, respectivement. Sachant que $K(\cdot)$ est une fonction noyau qui satisfait $\int K(t)\,dt = 1$ et $\int K(t)^2\,dt < \infty$ et que h est la largeur de fenêtre, et en utilisant le principe de la méthode de régression polynomiale locale, nous minimisons

$$\sum_{d'=1}^{D} \frac{P_d}{h}\{\overline{Y}_d - \alpha_0 - \alpha_1(z'_d - z_d), ..., -\alpha_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \tag{4}$$

et

$$\sum_{d'=1}^{D} \frac{P_d}{h}\{\overline{X}_{jd} - \gamma_0 - \gamma_1(z'_d - z_d), ..., -\gamma_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \tag{5}$$

par rapport aux α et aux γ de sorte que les espérances conditionnelles estimées (lissées) de population de y et X_j sur z_d, $m_{\mathbf{y}}(z_d)$ et $m_j(z_d)$ soient les solutions de α_0 et γ_0 pour les équations (4) et (5). En particulier,

$$m_j(z_d) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\overline{\mathbf{X}}_j$$

et

$$m_{\mathbf{y}}(z_d) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\overline{\mathbf{Y}}$$

où q est le degré du lisseur polynomial, $\mathbf{e}$ est un vecteur de dimension $(q+1)\times 1$ de la forme $(1, 0, 0, ..., 0)^T$, et $\mathbf{Z}$ et $\mathbf{K}_{\mathbf{W}}$ sont définis respectivement par

$$\mathbf{Z} = \begin{pmatrix} 1 & z_1 - z_d & \dots & (z_1 - z_d)^q \\ 1 & z_2 - z_d & \dots & (z_2 - z_d)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_D - z_d & \dots & (z_D - z_d)^q \end{pmatrix} \tag{6}$$

et $\mathbf{K}_{\mathbf{W}} = \text{diag}(\hat{P}_1 K((z_1 - z_d)/h), ..., \hat{P}_D K((z_D - z_d)/h))/h$.

Au moyen des estimateurs sous recensement des espérances conditionnelles $m_j(z_d)$ et $m_{\mathbf{y}}(z_d)$, nous définissons une matrice $\mathbf{M}_{\mathbf{X}}$ de dimensions $N \times p$ et un vecteur $\mathbf{M}_{\mathbf{y}}$ de dimension $N \times 1$ sous la forme

$$\mathbf{M}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} m_1(z_1) & m_2(z_1) & \cdots & m_p(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_1) & m_2(z_1) & \cdots & m_p(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} m_1(z_D) & m_2(z_D) & \cdots & m_p(z_D) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_D) & m_2(z_D) & \cdots & m_p(z_D) \end{pmatrix} \end{pmatrix} \tag{7}$$

et

$$\mathbf{M}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} m_{\mathbf{y}}(z_1) \\ \vdots \\ m_{\mathbf{y}}(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} m_{\mathbf{y}}(z_D) \\ \vdots \\ m_{\mathbf{y}}(z_D) \end{pmatrix} \end{pmatrix}.$$

Notons que les d^{e} blocs de $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$ sont de dimensions $N_d \times p$ et $N_d \times 1$, respectivement, où N_d est le nombre d'observations qui rentrent dans d^{e} classe et $\sum N_d = N$. En remplaçant la matrice, $E(\mathbf{X} \mid \mathbf{z})$, et le vecteur, $E(\mathbf{Y} \mid \mathbf{z})$, des espérances conditionnelles dans (3) par leurs estimations, $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$, et en utilisant le cadre des équations d'estimations générales proposé par Godambe et Thompson (1986) pour l'estimation des moindres carrés, nous pouvons

obtenir les versions en population finie des paramètres (estimateurs sous recensement) de $\boldsymbol{\beta}$, nommément $\mathbf{B}$, en résolvant

$$\begin{aligned}\mathbf{u}(\boldsymbol{\theta}) &= \sum_{k=1}^{N}(\mathbf{x}_k - \mathbf{M}_{xk})^T(y_k - M_{yk}) \\ &\quad - \sum_{k=1}^{N}(\mathbf{x}_k - \mathbf{M}_{xk})^T(\mathbf{x}_k - \mathbf{M}_{xk})\mathbf{B} \\ &= \mathbf{0}_{p\times 1}, \end{aligned} \tag{8}$$

où $\mathbf{M}_{xk}$ est la k^{e} ligne de la matrice $\mathbf{M}_{\mathbf{X}}$ de dimensions $N \times p$ et M_{yk} est le k^{e} élément du vecteur $\mathbf{M}_{\mathbf{y}}$ de dimension $N \times 1$. Le vecteur des paramètres de population finie $\boldsymbol{\theta}^T$ est composé de $(\mathbf{B}^T, \mathbf{m}_{\mathbf{x}}(\mathbf{z}), \mathbf{m}_{\mathbf{y}}(\mathbf{z})^T)$, où $\mathbf{m}_{\mathbf{x}}(\mathbf{z})$ est un vecteur de la forme $(\mathbf{m}_1(\mathbf{z})^T, ..., \mathbf{m}_p(\mathbf{z})^T)$ avec $\mathbf{m}_j(\mathbf{z}) = (m_j(z_1), ..., m_j(z_D))$ pour $j = 1, ..., p$ et $\mathbf{m}_{\mathbf{y}}(\mathbf{z}) = (m_{\mathbf{y}}(z_1), ..., m_{\mathbf{y}}(z_D))$. D'où l'expression explicite pour l'estimateur (paramètre sous recensement) $\mathbf{B}$ est

$$\mathbf{B} = ((\mathbf{X} - \mathbf{M}_{\mathbf{X}})^T(\mathbf{X} - \mathbf{M}_{\mathbf{X}}))^{-1}(\mathbf{X} - \mathbf{M}_{\mathbf{X}})^T(\mathbf{Y} - \mathbf{M}_{\mathbf{y}}).$$

Une fois que $\mathbf{B}$ est obtenu, la différence entre la variable réponse $\mathbf{Y}$ et le produit $\mathbf{XB}$ est traitée comme la variable aléatoire dépendante et la fonction $G(\cdot)$ est estimée conformément au modèle suivant

$$\mathbf{Y} - \mathbf{XB} = G(\mathbf{z}) + \boldsymbol{\varepsilon}.$$

La version en population finie de $G(\mathbf{z})$ à z_d, nommément $g(z_d)$, est

$$g(z_d) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}(\overline{\mathbf{Y}} - \overline{\mathbf{X}}\mathbf{B}),$$

où $\overline{\mathbf{X}}$ est une matrice de dimensions $D \times p$ de la forme $(\overline{\mathbf{X}}_1, ..., \overline{\mathbf{X}}_p)$.

En réalité, nous n'avons pas accès à l'ensemble de la population. Au contraire, nous ne pouvons observer qu'un échantillon tiré de la population en utilisant un certain plan d'échantillonnage probabiliste. Soit $\mathbf{s}$ l'ensemble de n unités échantillonnées avec l'échantillon $(y_k, \mathbf{x}_k, z_k, w_k)$ pour $k \in \mathbf{s}$, où w_k est le poids d'échantillonnage de l'unité k. En outre, nous supposons que la réponse est complète, si bien que la probabilité d'inclusion est égale à l'inverse du poids d'échantillonnage. Nous supposons aussi que les classes induites par les valeurs distinctes de $\mathbf{z}$ sont préservées lorsque l'on passe de la population à l'échantillon. Cette hypothèse est appropriée dans le cas d'une variable telle que l'âge enregistré comme étant l'âge au dernier anniversaire.

En appliquant la méthode de régression polynomiale locale pour données d'enquêtes complexes de Bellhouse et Stafford (2001), nous utilisons les versions d'échantillonnage des fonctions objectif dans (4) et (5) comme il suit,

$$\sum_{d'=1}^{D}\frac{\hat{p}_d}{h}\{\overline{y}_d - \alpha_0 - \alpha_1(z'_d - z_d), ..., -\alpha_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \tag{9}$$

et

$$\sum_{d'=1}^{D}\frac{\hat{p}_d}{h}\{\overline{x}_{jd} - \gamma_0 - \gamma_1(z'_d - z_d), ..., ,-\gamma_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right), \tag{10}$$

où $\overline{\mathbf{y}}$ et $\overline{\mathbf{x}}_j$ sont des estimateurs sur échantillon de $\overline{\mathbf{Y}}$ et $\overline{\mathbf{X}}_j$ et sont de la forme $(\overline{y}_1, ..., \overline{y}_D)^T$ et $(\overline{x}_{j1}, ..., \overline{x}_{jD})^T$, respectivement, et $\hat{p}_d$ est la proportion pondérée, dans l'échantillon, des observations comprises dans la classe d. Par conséquent, les estimateurs par sondage de $m_{\mathbf{y}}(z)$ et $m_j(z)$ à z_d sont donnés par

$$\begin{aligned}\hat{m}_j(z_d) &= \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\overline{\mathbf{x}}_j \\ \text{et} \qquad & \qquad (11) \\ \hat{m}_{\mathbf{y}}(z_d) &= \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\overline{\mathbf{y}}, \end{aligned}$$

où $\mathbf{Z}$ est de la même forme que dans (6) et $\hat{\mathbf{K}}_{\mathbf{W}}$ est défini comme étant

$$\hat{\mathbf{K}}_{\mathbf{W}} = \frac{1}{h}\text{diag}(\hat{p}_1 K((z_1 - z_d)/h), ..., \hat{p}_D K((z_D - z_d)/h)).$$

Nous pouvons aussi construire la matrice $\hat{\mathbf{M}}_{\mathbf{X}}$ de dimensions $n \times p$ et le vecteur $\hat{\mathbf{M}}_{\mathbf{X}}$ de dimension $n \times 1$ par la même méthode que celle utilisée pour construire $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$ dans les équations (7). Autrement dit, nous utilisons les estimateurs par échantillonnage $\hat{m}_j(z_d)$ et $\hat{m}_{\mathbf{y}}(z_d)$ donnés en (11) pour obtenir

$$\hat{\mathbf{M}}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} \hat{m}_{\mathbf{x}_1}(z_1) & \hat{m}_{\mathbf{x}_2}(z_1) & \cdots & \hat{m}_{\mathbf{x}_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{m}_{\mathbf{x}_1}(z_1) & \hat{m}_{\mathbf{x}_2}(z_1) & \cdots & \hat{m}_{\mathbf{x}_p}(z_1) \end{pmatrix} \\ \vdots \qquad\qquad \vdots \\ \begin{pmatrix} \hat{m}_{\mathbf{x}_1}(z_D) & \hat{m}_{\mathbf{x}_2}(z_D) & \cdots & \hat{m}_{\mathbf{x}_p}(z_D) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{m}_{\mathbf{x}_1}(z_D) & \hat{m}_{\mathbf{x}_2}(z_D) & \cdots & \hat{m}_{\mathbf{x}_p}(z_D) \end{pmatrix} \end{pmatrix}$$

et

$$\hat{\mathbf{M}}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \hat{m}_{\mathbf{y}}(z_1) \\ \vdots \\ \hat{m}_{\mathbf{y}}(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \hat{m}_{\mathbf{y}}(z_D) \\ \vdots \\ \hat{m}_{\mathbf{y}}(z_D) \end{pmatrix} \end{pmatrix}.$$

Soit n_d le nombre d'observations dans la d^{e} classe, tel que $\sum n_d = n$. Similairement à $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$ dans (7), les d^{e} blocs de $\hat{\mathbf{M}}_{\mathbf{X}}$ et $\hat{\mathbf{M}}_{\mathbf{y}}$ sont de dimensions $n_d \times p$ et $n_d \times 1$, respectivement.

Par analogie avec l'équation d'estimation en population (8), l'équation d'estimation sur échantillon pour $\mathbf{B}$ est

$$\begin{aligned} \hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) &= \sum_{k \in \mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{xk})^T (y_k - \hat{M}_{xk}) w_k \\ &\quad - \sum_{k \in \mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{xk})^T (\mathbf{x}_k - \hat{\mathbf{M}}_{xk}) \hat{\mathbf{B}} w_k \\ &= \mathbf{0}, \end{aligned} \tag{12}$$

où $\hat{\boldsymbol{\theta}}^T = (\hat{\mathbf{B}}^T, \hat{\mathbf{m}}_{\mathbf{x}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z})^T)$ est l'estimateur par échantillonnage de $\boldsymbol{\theta}^T = (\mathbf{B}^T, \mathbf{m}_{\mathbf{x}}(\mathbf{z}), \mathbf{m}_{\mathbf{y}}(\mathbf{z})^T)$. Notons qu'une approche semblable a été envisagée par Fuller (1975) et par Binder (1983). Néanmoins, la solution de (12) donne comme forme explicite de $\hat{\mathbf{B}}$ l'expression

$$\hat{\mathbf{B}} = ((\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{X}})^T \mathbf{W}_n (\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{X}}))^{-1} (\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{X}})^T \mathbf{W}_n (\mathbf{y} - \hat{\mathbf{M}}_{\mathbf{y}}),$$

où $\mathbf{W}_n$ est une matrice de poids de dimensions $n \times n$ avec les poids de sondage w_k sur la diagonale pour $k \in \mathbf{s}$, $\mathbf{y}$ est un vecteur de dimension $n \times 1$ contenant les observations sur échantillon de la variable réponse et $\mathbf{x}$ est une matrice de dimensions $n \times p$ contenant les observations sur échantillon des covariables.

En nous servant des estimations sur échantillon de $\mathbf{B}$ et en désignant par $\bar{\mathbf{x}}$ une matrice de dimensions $D \times p$ de la forme $(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_p)$, nous pouvons obtenir l'estimation sur échantillon de $g(z_d)$ sous la forme

$$\hat{g}(z_d) = \mathbf{e}^T (\mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} \mathbf{Z})^{-1} \mathbf{Z}^T \hat{\mathbf{K}}_{\mathbf{W}} (\bar{\mathbf{y}} - \bar{\mathbf{x}} \hat{\mathbf{B}}).$$

De nouveau, si q et h sont les mêmes que pour $\hat{m}_j(z_d)$, l'expression donnant $\hat{g}(z_d)$ se simplifie.

Quand nous appliquons les méthodes de régression polynomiale locale pour obtenir les estimateurs des espérances conditionnelles ainsi que la fonction arbitraire $G(\cdot)$, nous devons choisir une largeur de fenêtre h appropriée. Comme le groupement par classe intervient dans tous les aspects du processus d'estimation et que nous supposons que les classes induites par les valeurs distinctes de $\mathbf{z}$ sont préservées lorsqu'on passe de la population à l'échantillon, nous soutenons que la même largeur de fenêtre devrait être utilisée pour obtenir les estimateurs sous recensement et les estimateurs sous échantillonnage. Puisque nous ne disposons pas de toutes les observations de la population finie, nous utilisons l'échantillon pour choisir la largeur de fenêtre appropriée. Dans le présent article, nous adoptons la méthode de Fan et Gijbels (1995), qui ont élaboré un sélecteur de largeur de fenêtre dicté par les données qui combine les idées des méthodes d'insertion de données et de validation croisée pour les données indépendantes et identiquement distribuées. Afin d'appliquer cette méthode dictée par les données à notre cas, nous avons besoin de critères, tels que la somme des carrés des résidus et l'erreur quadratique moyenne, concernant les estimations résultantes des espérances conditionnelles. En notant que ces critères dépendent des espérances conditionnelles estimées ou des fonctions de régression et des dérivées des fonctions de régression, nous pouvons utiliser les fonctions objectif définies en (9) et (10) pour obtenir non seulement les estimations par sondage des fonctions de régression, mais aussi les dérivées de ces fonctions. Pour plus de détails, voir Wang (2004).

3. Propriétés sous le plan des estimateurs par échantillonnage

3.1 Notation et hypothèses

Afin d'illustrer les propriétés des estimateurs sous le plan d'échantillonnage, à l'instar de Särndal, Swensson et Wretman (1992) et d'Isaki et Fuller (1982), nous considérons une série de populations emboîtées U_ν, pour $\nu = 1, 2, \ldots$, telles que $U_1 \subset U_2 \subset U_3 \subset \ldots$. Toutes les grandeurs de population, les tailles et valeurs d'échantillon et les estimateurs par sondage ont l'indice ν. Cependant, pour simplifier la notation, nous laissons tomber l'indice inférieur ν pour ces quantités. Nous désignons l'espérance et la variance par rapport au plan d'échantillonnage par E_p et Var_p, respectivement, et conformément aux populations emboîtées susmentionnées, nous définissons la convergence par rapport au plan et l'absence asymptotique de biais comme dans Thompson (1997, page 167).

Dans la suite de l'exposé, le développement des résultats asymptotiques pour les estimateurs dépendra de la normalité et de la convergence asymptotiques des estimations des moyennes et des totaux. Nous ne nous limiterons pas à des plans d'échantillonnage particuliers ; au contraire, nous supposons que tous les totaux d'enquête qui figurent dans les estimateurs sont de type Horvitz-Thompson. Donc, la convergence et la normalité asymptotique des estimateurs sont soumises aux conditions de régularité standard appliquéees aux plans d'échantillonnage pour la convergence et

la normalité des estimateurs de type Horvitz-Thompson, qui ont été étudiées par Madow (1948), Hàjek (1960), Bickel et Freedman (1983), Krewski et Rao (1981) et Shao (1996). Les publications susmentionnées ont en commun certaines contraintes appliquées au plan d'échantillonnage. Une implication de ces contraintes est qu'aucun poids de sondage n'est anormalement grand, que le nombre total de grappes échantillonnées au premier degré ou d'unités primaires d'échantillonnage augmente, mais avec un écart croissant entre l'échantillon et la population. En outre, une condition de type Liapunov assure que les variables z, $\boldsymbol{x}$ et y se développent d'une manière régulière quand ν tend vers l'infini.

Nous utilisons le résultat selon lequel tout vecteur des estimateurs des totaux basés sur les données groupées en classe est asymptotiquement normal multivarié, si les conditions décrites au paragraphe précédent sont satisfaites et que le nombre de domaines est fixé. Nous obtenons cela en appliquant les résultats décrits dans Shao (1996, page 211) et dans Serfling (1980, page 18). Shao (1996) montre que, dans ce cadre, toute fonction lisse des estimations des totaux est asymptotiquement normale. Une estimation d'une moyenne de domaine est une fonction lisse de ce genre. De même, toute combinaison linéaire de diverses estimations de moyenne de domaine est une fonction lisse des estimations par sondage des totaux. Pour les besoins de notre étude, les classes forment les domaines et, donc, tout vecteur de moyennes de classe est asymptotiquement normal multivarié. Le résultat asymptotique utilisé ici dépend de l'existence d'un nombre fixe de classes. Cependant, en principe, il peut être intégré dans une théorie des paramètres de superpopulation, comme par exemple dans l'approche de Buskirk et Lohr (2005).

Définissons $\hat{\mathbf{m}}_\xi(\mathbf{z}) = (\hat{\mathbf{m}}_\mathbf{x}(\mathbf{z}), \hat{\mathbf{m}}_\mathbf{y}(\mathbf{z})^T)^T$ comme étant l'estimateur par sondage de $\mathbf{m}_\xi(\mathbf{z}) = (\mathbf{m}_\mathbf{x}(\mathbf{z}), \mathbf{m}_\mathbf{y}(\mathbf{z})^T)^T$. Appliquons une méthode de linéarisation de Taylor à (12) et désignons par ε une quantité qui s'approche de 0 et, quand $\hat{\boldsymbol{\theta}}$ s'approche de $\boldsymbol{\theta}$, nous avons

$$-\hat{\mathbf{u}}_B(\boldsymbol{\theta})(\hat{\mathbf{B}} - \mathbf{B}) \doteq \hat{\mathbf{u}}(\boldsymbol{\theta}) + \hat{\mathbf{U}}_\xi(\boldsymbol{\theta})(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|\,\varepsilon, \qquad (13)$$

où $\hat{\mathbf{u}}(\boldsymbol{\theta})$ est un estimateur par échantillonnage linéaire de $\mathbf{u}(\boldsymbol{\theta})$ dans (8) et est de la forme

$$\hat{\mathbf{u}}(\boldsymbol{\theta}) = \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \mathbf{M}_{xk})^T (y_k - M_{yk}) w_k - \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \mathbf{M}_{xk})^T (\mathbf{x}_k - \mathbf{M}_{xk}) \mathbf{B} w_k \ ; \qquad (14)$$

$\hat{\mathbf{u}}_\mathbf{B}(\boldsymbol{\theta})$ est le gradient de $\hat{\mathbf{B}}$ obtenu d'après $\hat{\mathbf{u}}(\boldsymbol{\theta})$; et $\hat{\mathbf{U}}_\xi(\boldsymbol{\theta})$ est une matrice de dimensions $p \times (p+1)D$ dont les composantes sont les dérivées premières de $\hat{\mathbf{u}}(\boldsymbol{\theta})$ par rapport à $\mathbf{m}_\xi(\mathbf{z})$. Désignons par $\mathbf{u}_B(\boldsymbol{\theta})$ et $\mathbf{U}_\xi(\boldsymbol{\theta})$ les paramètres de population correspondant à $\hat{\mathbf{u}}_\mathbf{B}(\boldsymbol{\theta})$ et $\hat{\mathbf{U}}_\xi(\boldsymbol{\theta})$, respectivement.

En plus des conditions de régularité susmentionnées, nous imposons les conditions suivantes, en désignant par $\mathcal{N}$ un voisinage de la valeur réelle des paramètres d'intérêt.

C1. $\lim_{\nu\to\infty} \mathbf{u}(\boldsymbol{\theta})/N$ existe et est finie pour tous $\boldsymbol{\theta}$ et $\mathcal{N}$.

C2. $\lim_{\nu\to\infty} \mathbf{u}_\mathbf{B}(\boldsymbol{\theta})/N = \mathbf{H}_\mathbf{B}$ et $\mathbf{H}_\mathbf{B}$ est de plein rang et inversible pour tous $\boldsymbol{\theta}$ et $\mathcal{N}$.

C3. $\lim_{\nu\to\infty} \mathbf{U}_\xi(\boldsymbol{\theta})/N = \mathbf{H}_\xi(\boldsymbol{\theta})$ et $\mathbf{H}_\xi(\boldsymbol{\theta})$ possède un déterminant fini pour tous $\boldsymbol{\theta}$ et $\mathcal{N}$.

C4. $\lim_{\nu\to\infty} n\,\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta})/N) = \mathrm{V}(\hat{\mathbf{u}}(\boldsymbol{\theta}))$, où Var_p est la variance par rapport au plan et $V(\hat{\mathbf{u}}(\boldsymbol{\theta}))$ est une matrice de variance définie positive pour tous $\boldsymbol{\theta}$ et $\mathcal{N}$.

C5. $\lim_{\nu\to\infty} N_d/N = \omega_d$ et $\lim_{\nu\to\infty} n/N = f$, où ω_d ainsi que f sont des constantes comprises entre 0 et 1.

C6. Soit $\mathbf{A}_d = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_\mathbf{W}$ la matrice de lissage de population ; alors $\lim_{\nu\to\infty} \mathbf{A}_d$ existe et est finie pour $d = 1, \ldots, D$.

C7. $\lim_{\nu\to\infty} n\,\mathrm{Var}_p(\hat{\mathbf{m}}_\xi(\mathbf{z})) = \mathrm{V}(\hat{\mathbf{m}}_\xi(\mathbf{z}))$.

C8. Les matrices des valeurs de population $\mathbf{Z}^T\mathbf{K}_\mathbf{W}\mathbf{Z}$ et $\mathbf{u}_\mathbf{B}(\boldsymbol{\theta})$ sont inversibles, ainsi que leurs estimateurs par échantillonnage $\mathbf{Z}^T\hat{\mathbf{K}}_\mathbf{W}\mathbf{Z}$ et $\hat{\mathbf{u}}_\mathbf{B}(\hat{\boldsymbol{\theta}})$.

3.2 Propriétés asymptotiques de $\hat{\mathbf{B}}$

Les preuves de tous les lemmes et théorèmes exposés dans la présente section et dans la suivante figurent en annexe. D'après les résultats de la linéarisation de Taylor dans (13), nous savons que les propriétés de $\hat{\mathbf{B}}$ dépendent de celles de $\hat{\mathbf{u}}(\boldsymbol{\theta})$, $\hat{\mathbf{u}}_\mathbf{B}(\boldsymbol{\theta})$, $\hat{\mathbf{U}}_\xi(\boldsymbol{\theta})$ et $\hat{\mathbf{m}}_\xi(\mathbf{z})$; leurs propriétés sont énoncées dans les deux lemmes qui suivent.

Lemme 1. *Si les conditions C*1 *à C*4 *sont satisfaites, nous avons, quand* $\nu \to \infty$:

1) $\sqrt{n}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta}))/N \longrightarrow N(\mathbf{0}, \mathrm{V}(\hat{\mathbf{u}}(\boldsymbol{\theta})))$;
2) $|\hat{\mathbf{u}}_B(\boldsymbol{\theta}) - \mathbf{u}_B(\boldsymbol{\theta})|/N$ *et* $|\hat{\mathbf{U}}_\xi(\boldsymbol{\theta}) - \mathbf{U}_\xi(\boldsymbol{\theta})|$ *convergent vers* $\mathbf{0}$ *en probabilité pour* $\boldsymbol{\theta}$ *et* $\mathcal{N}$;
3) $|\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta})|/N$ *converge vers zéro en probabilité.*

Lemme 2. *Sous les conditions C*5 *à C*7, $\sqrt{n}(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) = O_p(1)$.

En nous appuyant sur les lemmes 1 et 2, nous obtenons la normalité asymptotique de $\hat{\mathbf{B}}$ dans le théorème 1.

Théroème 1. *Sous les conditions C*1 *à C*7, *en supposant que l'espace des paramètres contient un voisinage du paramètre d'intérêt, nous avons, quand* ν *tend vers l'infini* :

1) $\sqrt{n}(\hat{\mathbf{B}}-\mathbf{B}) \xrightarrow{d} N(\mathbf{0}, \mathrm{V}(\hat{\mathbf{B}}))$ *où* $\mathrm{V}(\hat{\mathbf{B}}) = \lim_{v\to\infty} n\mathrm{Var}_p(\hat{\mathbf{B}})$;

2) $|\hat{\mathbf{B}} - \mathbf{B}|$ *converge vers zéro en probabilité.*

Pour obtenir les moments approximatifs de $\hat{\mathbf{B}}$, nous prenons les espérances des deux membres de l'équation (13), ce qui nous donne

$$E_p(-\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})(\hat{\mathbf{B}} - \mathbf{B})) \doteq E_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) + E_p\{\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})[\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})]\} + E_p(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||)\varepsilon. \quad (15)$$

L'hypothèse selon laquelle les deuxièmes moments des estimations sont bornés fait disparaître le dernier terme de l'équation (15) à la limite. En nous inspirant de Binder (1983), nous avons

$$E_p(-\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta}))E_p((\hat{\mathbf{B}}-\mathbf{B})) \doteq E_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) + E_p(\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta}))E_p\{[\hat{\mathbf{m}}_{\xi}(\mathbf{z})-\mathbf{m}_{\xi}(\mathbf{z})]\}.$$

Les totaux d'enquête qui définissent le vecteur $\hat{\mathbf{u}}(\boldsymbol{\theta})$ et la matrice $\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})$ sont des estimateurs de type Horvitz-Thompson et sont sans biais (Thompson 1997). D'où $E_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) = \mathbf{u}(\boldsymbol{\theta})$ et $E_p(\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})) = \mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})$. Puisque $\mathbf{u}(\boldsymbol{\theta})$ est l'équation d'estimation des coefficients linéaires partiels définis en (8), elle équivaut à un vecteur nul de dimension $1\times p$. En outre, Bellhouse et Stafford (2001) ont montré que $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ est un estimateur asymptotiquement sans biais de $\mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})$. Donc, $-\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})E_p((\hat{\mathbf{B}} - \mathbf{B})) \doteq \mathbf{0}$, ou, en nous basant sur les conditions que $\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})$ est inversible et que $\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})^{-1}$ est finie, nous avons $E_p(\hat{\mathbf{B}}) \doteq \mathbf{B}$.

En prenant la variance des deux membres de l'équation (13) et en utilisant les matrices de variance-covariance approximatives de $\hat{\mathbf{u}}(\boldsymbol{\theta})$ et $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$, nous obtenons la variance asymptotique de $\hat{\mathbf{B}}$ sous la forme

$$\mathrm{Var}_p(\hat{\mathbf{B}}) \doteq \mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})^{-1} (\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) + \mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})(\mathbf{A}(\mathbf{J}\otimes\mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}))\mathbf{A}^T)\mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})^T + 2(I_p\otimes\boldsymbol{\ell})(\boldsymbol{\ell}\otimes\mathbf{C})\mathbf{A}^T\mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})^T)(\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})^T)^{-1}, \quad (16)$$

où $\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta}))$ est une matrice de dimension $p\times p$ composée des variances des totaux compris dans le vecteur $\hat{\mathbf{u}}(\boldsymbol{\theta})$ et $\mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ est la matrice de variance-covariance des moyennes groupées par classe des covariables paramétriques et de la variable réponse. Les matrices $\mathbf{J}$ et $\boldsymbol{\ell}$ sont la matrice unitaire de dimension $D\times D$ et le vecteur unitaire de dimension $1\times D$ respectivement. Enfin, nous avons

$$\mathbf{A} = \begin{pmatrix} I_{p+1}\otimes\mathbf{A}_1 & 0 & 0 & 0 \\ 0 & I_{p+1}\otimes\mathbf{A}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & I_{p+1}\otimes\mathbf{A}_D \end{pmatrix}$$

et

$$\mathbf{C} = \begin{pmatrix} \mathrm{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{x}}_1) & \cdots & \mathrm{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{x}}_p) & \mathrm{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{y}}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{x}}_1) & \cdots & \mathrm{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{x}}_p) & \mathrm{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{y}}) \end{pmatrix},$$

où, pour $j = 1, ..., p$, $\hat{\mathbf{t}}_j$ est un vecteur de dimension $D\times 1$ dont la d^{e} entrée est $\sum_{k\in\mathbf{s}_d} w_{jk}\, u_{jk}(\boldsymbol{\theta})$ et $\mathbf{A}_d = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}$ pour $d = 1, ..., D$.

En remplaçant $\boldsymbol{\theta}$, $\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta}))$, $\mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, $\mathbf{A}$ et $\mathbf{C}$ par leurs estimateurs sur échantillon, nous obtenons l'estimateur par sondage de la variance de $\hat{\mathbf{B}}$:

$$\widehat{\mathrm{Var}}_p(\hat{\mathbf{B}}) = \hat{\mathbf{u}}_{\mathbf{B}}^{-1}(\hat{\boldsymbol{\theta}})(\widehat{\mathrm{Var}}_p(\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}})(\hat{\mathbf{A}}(\mathbf{J}\otimes\widehat{\mathrm{Cov}}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}))\hat{\mathbf{A}}^T)\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}})^T + 2(I_p\otimes\boldsymbol{\ell})(\boldsymbol{\ell}\otimes\hat{\mathbf{C}})\hat{\mathbf{A}}^T\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}})^T)(\hat{\mathbf{u}}_{\mathbf{B}}^T(\hat{\boldsymbol{\theta}})^T)^{-1},$$

où $\hat{\mathbf{A}}$ est l'estimateur par sondage de $\mathbf{A}$ et est composé de $\hat{\mathbf{A}}_d = \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}$.

3.3 Propriétés asymptotiques de $\hat{g}(\cdot)$

Soit $\bar{\mathbf{r}} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}$ l'estimateur sur échantillon de $\bar{\mathbf{R}} = \bar{\mathbf{Y}} - \bar{\mathbf{X}}\mathbf{B}$. Une linéarisation autour des paramètres de population, ainsi que l'absence de biais par rapport au plan pour les moyennes de domaine et $\hat{\mathbf{B}}$, aboutissent à l'absence asymptotique de biais par rapport au plan pour $\bar{\mathbf{r}}$. En réexprimant $\hat{g}(z_d)$, nous obtenons $\hat{g}(z_d) = \hat{\mathbf{A}}_d\bar{\mathbf{r}}$. Dans $\hat{\mathbf{A}}_d$, nous pouvons développer $(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}$ en utilisant le développement en série de Taylor $(\mathbf{I}+\mathbf{G})^{-1} = \mathbf{I} - \mathbf{G} + \mathbf{G}^2 - \cdots$ sachant que $\mathbf{G}$ est une matrice symétrique et inversible. En utilisant les deux premiers termes du développement, nous pouvons montrer que $E_p(\hat{\mathbf{A}}_d)$ est approximativement $\mathbf{A}_d$. Donc, nous avons l'absence asymptotique de biais par rapport au plan pour $\hat{g}(z_d)$. Selon la même méthode, nous obtenons la variance asymptotique approximative par rapport au plan de $\hat{g}(z_d)$ sous la forme

$$\mathrm{Var}_p(\hat{g}(z_d)) = \mathbf{A}_d\mathrm{Var}_p(\bar{\mathbf{r}})\mathbf{A}_d^T,$$

où, sachant que $\mathbf{Q} = (1, -B_1, ..., -B_p)$,

$$\begin{aligned}\mathrm{Var}_p(\bar{\mathbf{r}}) \doteq\ & (\mathbf{Q}\otimes\mathbf{I}_D)\mathrm{Cov}_p(\bar{\mathbf{x}},\,\bar{\mathbf{y}})(\mathbf{Q}\otimes\mathbf{I}_D)^T \\ & + \bar{\mathbf{X}}\,\mathrm{Var}_p(\hat{\mathbf{B}})\bar{\mathbf{X}}^T \\ & - 2(\mathbf{Q}\otimes\mathbf{I}_D)\mathrm{Cov}_p(\bar{\mathbf{y}},\,\hat{\mathbf{B}})\bar{\mathbf{X}}^T \\ & - \sum_{j=1}^{p} 2(\mathbf{Q}\otimes\mathbf{I}_D)\mathrm{Cov}_p(\bar{\mathbf{x}}_j,\,\hat{\mathbf{B}})\bar{\mathbf{X}}^T.\end{aligned}$$

Sachant la variance estimée de $\bar{\mathbf{r}}$, à savoir

$$\begin{aligned}\widehat{\mathrm{Var}}_p(\bar{\mathbf{r}}) =\ & (\hat{\mathbf{Q}}\otimes\mathbf{I}_D)\widehat{\mathrm{Cov}}_p(\bar{\mathbf{x}},\,\bar{\mathbf{y}})(\hat{\mathbf{Q}}\otimes\mathbf{I}_D)^T \\ & + \bar{\mathbf{x}}\,\widehat{\mathrm{Var}}_p(\hat{\mathbf{B}})\bar{\mathbf{x}}^T \\ & - 2(\mathbf{Q}\otimes\mathbf{I}_D)\widehat{\mathrm{Cov}}_p(\bar{\mathbf{y}},\,\hat{\mathbf{B}})\bar{\mathbf{x}}^T \\ & - \sum_{j=1}^{p} 2(\hat{\mathbf{Q}}\otimes\mathbf{I}_D)\widehat{\mathrm{Cov}}_p(\bar{\mathbf{x}}_j,\,\hat{\mathbf{B}})\bar{\mathbf{x}}^T,\end{aligned}$$

la variance estimée de $\hat{g}(z_d)$ est $\widehat{\mathrm{Var}}_p(\hat{g}(z_d)) = \hat{\mathbf{A}}_d\,\widehat{\mathrm{Var}}_p(\bar{\mathbf{r}})\hat{\mathbf{A}}_d^T$.

La normalité asymptotique de $\hat{g}(\cdot)$ dépend aussi de la normalité de $\bar{\mathbf{r}}$, qui est montrée dans le lemme suivant.

Lemme 3. *Sous les conditions* **C**1 *à* **C**7 *et en supposant que la dimension de* $\bar{\mathbf{r}}$ *est finie*, *nous avons*, *quand* ν *tend vers l'infini*

$$\sqrt{n}\,(\bar{\mathbf{r}}-\bar{\mathbf{R}}) \xrightarrow{d} N(\mathbf{0},\,\mathrm{V}(\bar{\mathbf{r}})),$$

où

$$\mathrm{V}(\bar{\mathbf{r}}) = \lim_{\nu\to\infty} n\,\mathrm{Var}_p(\bar{\mathbf{r}}).$$

En nous basant sur la normalité asymptotique établie dans le lemme 3 et sur l'estimateur de la variance de $\hat{g}(z_d)$, nous établissons les propriétés asymptotiques de $\hat{g}(z_d)$ dans le théorème suivant.

Théorème 2. *Sous les conditions* **C**1 *à* **C**7, *nous avons*, *quand* ν *tend vers l'infini* :

1) $|\hat{g}(z_d) - g(z_d)| \xrightarrow{P} 0$;

2) $(\hat{g}(z_d) - g(z_d))/\sqrt{\widehat{\mathrm{Var}}_p(\hat{g}(z_d))} \xrightarrow{d} N(0,\,1)$.

4. Études par simulation

4.1 Plan de l'expérience

L'étude par simulation dont l'exécution est décrite ici a été conçue en vue d'illustrer les résultats théoriques des théorèmes 1 et 2. Nous avons produit les données suivant un processus en deux étapes qui imitaient l'approche d'échantillonnage en superpopulation. Pour commencer, nous avons produit la population finie, puis nous avons sélectionné l'échantillon à partir de cette population. En particulier, nous avons considéré une population finie de $L = 500$ grappes avec $M(= M_i) = 20\,000$ dans chacune d'elles. Nous avons obtenu les observations de population pour la mesure d'intérêt y_{ij} à partir du modèle

$$\begin{aligned}y_{ij} =\ & \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \\ & + 0,5\exp\left(\frac{z_{ij} - 40}{10}\right) + \mu_i + \varepsilon_{ij} \qquad (17)\end{aligned}$$

pour $i = 1, \ldots, L$ et $j = 1, \ldots, M$ où les termes d'erreur μ_i et ε_{ij} sont mutuellement indépendants avec $\mu_i \sim N(0, \sigma_\mu^2)$ et $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Nous posons que $\sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$, de sorte que le coefficient de corrélation intragrappe est $\rho = \sigma_\mu^2/\sigma^2$. Parmi les covariables comprises dans le modèle, x_{1ij} ainsi que x_{2ij} ont été traitées comme étant la partie linéaire paramétrique du modèle et z_{ij}, comme étant la partie non paramétrique. Nous avons produit les x_{1ij} à partir de la loi de Bernoulli(1/2) et les x_{2ij} à partir de la loi uniforme(0, 1). Les z_{ij} ont été produites d'après la distribution de l'âge de la population canadienne (selon le Recensement de 1996) pour la tranche d'âge de 18 à 64 ans et elles étaient indépendantes des termes d'erreur. Nous présentons dans cette étude les résultats pour les valeurs $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$, $\sigma^2 = 3$ et $\rho = 0,\ 0{,}2,\ 0{,}5$. Nous avons utilisé pour l'étude un plan d'échantillonnage à deux degrés, avec $l(= 10,\ 25,\ 50,\ 100)$ grappes sélectionnées au hasard dans L et $m(= 1\,000)$ unités secondaires d'échantillonnage sélectionnées au hasard dans chaque grappe de taille M. Pour chaque taille d'échantillon et valeur de ρ, nous avons répété la simulation 300 fois. Au niveau de la population, nous avons appliqué la méthode de sélection de la largeur de fenêtre de Fan et Gijbels (1995) et déterminé que les largeurs de fenêtre pour l'estimation des espérances conditionnelles de X_1 et X_2 sur z étaient de 1,2 et 1,5, respectivement. Quand nous avons lissé les résidus pour estimer $g(z)$, la largeur de fenêtre était de 0,6.

4.2 Résultats

En utilisant la population finie produite, nous avons trouvé que les estimations sous recensement étaient $B_1 = 2{,}01$ et $B_2 = 3{,}00$. Pour confirmer l'absence de biais par rapport au plan et l'efficacité de $\hat{\mathbf{B}}$, nous avons calculé le carré du biais simulé (Biais^2), qui est le carré de la différence entre la moyenne des estimations simulées et des estimations sous recensement. En outre, nous présentons le ratio des estimations moyennes de la variance à la variance simulée de chaque estimateur d'un coefficient linéaire (RVar) pour montrer la validité de l'estimateur de variance $\mathrm{Var}_p(\hat{\mathbf{B}})$. Pour évaluer la normalité de $\hat{\mathbf{B}}$, nous avons standardisé les estimations des coefficients linéaires en utilisant l'écart-type empirique et la valeur de population de $\mathbf{B}$, et nous avons tracé les graphiques quantile - quantile des valeurs standardisées.

En appliquant la méthode semiparamétrique de Speckman (1988) au modèle (17), nous avons obtenu les estimations sous recensement $g(z)$ pour $z = 18, ..., 64$. Pour évaluer l'exactitude de $\hat{g}(z)$ par rapport au plan, nous avons calculé la différence entre $\hat{g}(z)$ et $g(z)$ à chaque point distinct. La moyenne des carrés des différences sur 47 valeurs distinctes de z est alors présentée comme le carré du biais moyen MBiais^2. Nous avons calculé deux erreurs quadratiques moyennes pour vérifier l'efficacité par rapport au plan de $\hat{g}(z)$ et la convergence de $\widehat{\mathrm{Var}}_p(\hat{g}(z))$. L'une d'elles est la moyenne des estimations de l'erreur quadratique moyenne intégrée (MEQMI), que nous obtenons en calculant d'abord la somme des $\widehat{\mathrm{Var}}_p(\hat{g}(z))$ sur $z = 18, ..., 64$ pour chaque simulation, puis en prenant la moyenne des sommes sur le nombre total des simulations. L'erreur quadratique moyenne intégrée (EQMI) simulée est une autre erreur quadratique moyenne que nous avons calculée par sommation des erreurs quadratiques moyennes simulées à chaque point distinct de z. La moyenne des ratios de la moyenne simulée de $\widehat{\mathrm{Var}}_p(\hat{g}(z))$ à la variance simulée de $\hat{g}(z)$ (Reff) montre la convergence de $\widehat{\mathrm{Var}}_p(\hat{g}(z))$. En outre, nous avons calculé la couverture de l'intervalle de confiance à 95 % ponctuel à chaque point distinct de z.

Les résultats pour les propriétés de $\hat{\mathbf{B}}$, $\widehat{\mathrm{Var}}_p(\hat{\mathbf{B}})$, $\hat{g}(z)$ et $\widehat{\mathrm{Var}}_p(\hat{g}(z))$ sont présentés aux tableaux 1 et 2 et aux figures 2 et 3. Les tableaux 1 et 2 donnent l'information sur l'exactitude et la précision des estimations simulées de $\hat{\mathbf{B}}$ et $\hat{g}(\cdot)$. La figure 2 donne les graphiques quantile-quantile de la valeur normalisée en échantillon de $\hat{B}_2$. Il convient de souligner que les graphiques quantile-quantile de $\hat{B}_1$ se comportent de manière semblable à ceux de $\hat{B}_2$. La figure 3 représente graphiquement la couverture des intervalles de confiance à 95 % pour $g(\cdot)$. Dans les figures 2 et 3, nous ne présentons que les cas où $l = 10, 25, 100$ et $\rho = 0, 0{,}5$. La performance globale des estimateurs concorde avec la théorie des théorèmes 1 et 2.

Le tableau 1 confirme l'absence de biais par rapport au plan dans $\hat{\mathbf{B}}$. Il montre aussi que, à mesure que la taille d'échantillon augmente, les propriétés des estimations des coefficients linéaires s'améliorent pour toutes les structures d'erreur. En particulier, le carré du biais et la variance de $\hat{\mathbf{B}}$ diminuent quand le nombre d'échantillons primaires augmente. La variance estimée de $\hat{\mathbf{B}}$ se rapproche de la variance simulée de $\hat{\mathbf{B}}$ à mesure que la taille d'échantillon augmente, ce qui confirme la convergence des estimations de variance de $\hat{\mathbf{B}}$. Si nous comparons les variances et les biais de $\hat{\mathbf{B}}$ dans les cas où $\rho = 0{,}2$ et $\rho = 0{,}5$ au cas où $\rho = 0$, nous constatons que la corrélation intragrappe (effet de grappe) n'a pas eu d'incidence sur les propriétés de $\hat{\mathbf{B}}$. Cela pourrait être dû au fait que la taille d'échantillon intragrappe était grande.

L'examen de la figure 2 révèle que le nombre d'unités primaires échantillonnées et l'effet de grappe jouent un certain rôle dans la normalité de $\hat{\mathbf{B}}$. En particulier, quand la taille de l'échantillon est faible, par exemple $l = 10$, la normalité de l'estimateur $\hat{\mathbf{B}}$ standardisé présente un certain écart par rapport à la théorie pour $\rho = 0$ et $\rho = 0{,}5$. Quand l passe à 25, nous constatons que la performance de $\hat{\mathbf{B}}$ pour $\rho = 0$ commence à s'améliorer, tandis que pour $\rho = 0{,}5$, nous n'observons aucune amélioration jusqu'à $l = 100$. Empiriquement, ce résultat donne à penser que si le nombre de grappes est faible, nous ne devrions pas nous fier à la normalité théorique des estimations du coefficient ; nous pourrions plutôt utiliser la loi t pour effectuer l'inférence.

En ce qui concerne les résultats de la partie non paramétrique de l'estimation, le tableau 2 montre que la moyenne des estimations de l'erreur quadratique moyenne intégrée est très proche des erreurs quadratiques moyennes intégrées simulées pour toutes les tailles d'échantillon et structures d'erreur. L'absence de biais par rapport au plan est de nouveau confirmée par la moyenne du carré du biais (MBiais^2). Les valeurs du ratio moyen de la variance estimée à la variance simulée (RVar), qui sont proches de 1 dans tous les cas, sont en harmonie avec la convergence par rapport au plan de l'estimateur de la variance de $\hat{g}(z)$. Les erreurs quadratiques moyennes intégrées de $\hat{g}(\cdot)$ sont influencées par les corrélations intragrappe, ce que l'on peut déduire du fait que l'erreur quadratique moyenne intégrée ainsi que l'erreur quadratique moyenne intégrée estimée moyenne convergent plus lentement vers zéro quand $\rho = 0{,}2$ et $\rho = 0{,}5$ que quand $\rho = 0$.

Tableau 1
Résultats des simulations pour les estimateurs ponctuels de $\hat{B}$

		ρ = 0			ρ = 0,2			ρ = 0,5		
	l	Biais^2 ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Biais^2 ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Biais^2 ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar
$\hat{B}_1$	10	5,77	1,07	1,13	3,12	1,1	1,01	0,23	1,19	1,33
	25	9,97	0,46	1,07	0,38	0,44	1,08	0,30	0,53	0,98
	50	0,54	0,21	1,08	0,13	0,27	0,93	0,026	0,21	1,18
	100	0,22	0,13	0,96	0,019	0,11	1,06	0,039	0,13	0,98
$\hat{B}_2$	10	0,36	3,32	1,13	1,54	3,74	0,92	1,26	3,5	1,78
	25	0,64	1,31	1,10	2,40	1,34	1,06	0,14	1,42	1,03
	50	0,31	0,75	0,94	1,27	0,85	0,94	0,16	0,76	0,97
	100	0,15	0,38	0,94	1,11	0,38	0,98	0,072	0,33	1,03

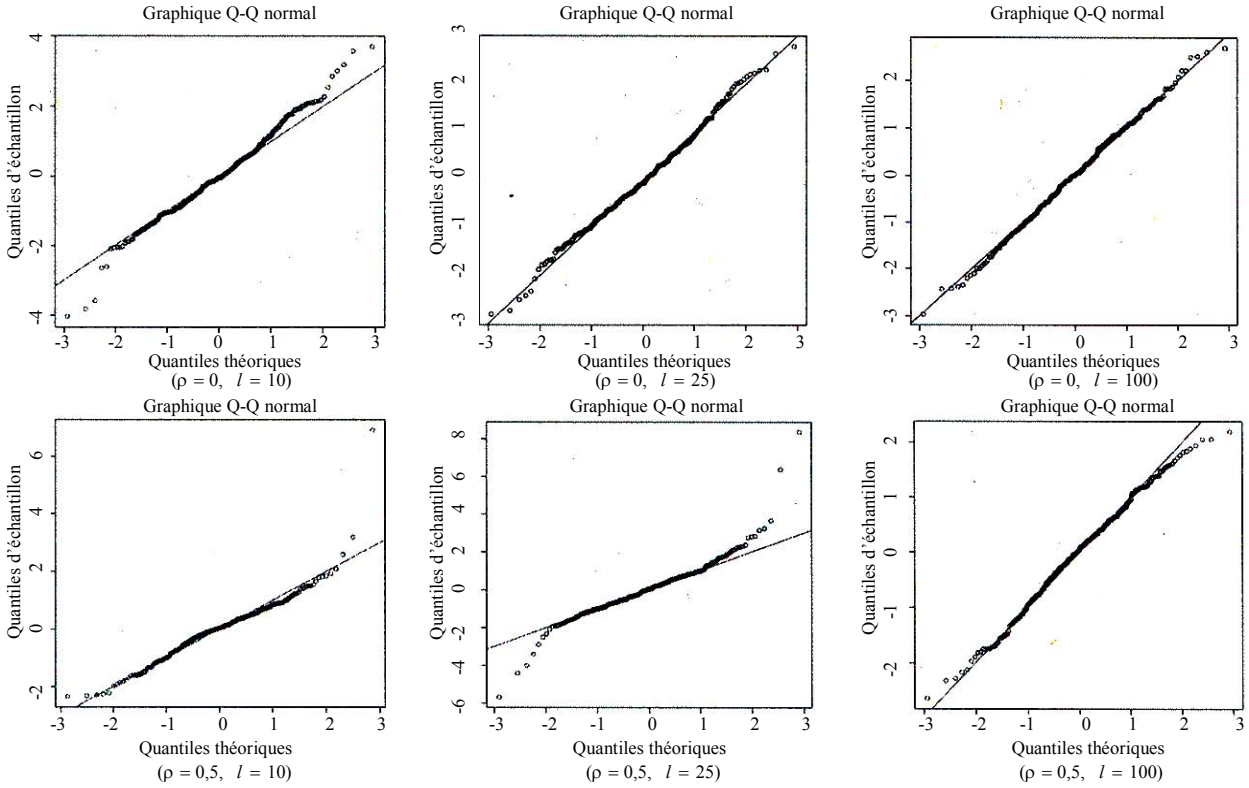

Figure 2 Graphiques quantile-quantile pour $\hat{B}_2$ standardisé

Tableau 2
Biais et efficacité de $\hat{g}(z)$

ρ	l	MEQMI	EQMI	MBiais^2 (×10-5)	RVar
0	10	0,37	0,42	5,29	1,27
	25	0,15	0,17	3,20	1,10
	50	0,074	0,086	3,29	1,09
	100	0,037	0,044	2,34	1,08
0,2	10	2,95	3,25	6,13	0,91
	25	1,22	1,17	3,71	1,04
	50	0,74	0,54	2,34	1,0
	100	0,26	0,27	7,08	0,98
0,5	10	8,143	8,877	3,73	0,92
	25	3,155	3,073	6,56	1,03
	50	1,461	1,599	2,86	1,15
	100	0,659	0,607	3,59	1,09

À la figure 3, la couverture des intervalles de confiance à 95 % ponctuel pour $g(\cdot)$ varie entre 85 % et 96 %. Elle s'améliore à mesure qu'augmente la taille de l'échantillon. Toutefois, la performance de $\hat{g}(\cdot)$ est plus sensible à la réduction de la taille effective d'échantillon causée par la corrélation intragrappe. En particulier, les couvertures des intervalles de confiance à 95 % dans les cas où $\rho = 0{,}2$ et $\rho = 0{,}5$ sont inférieures au niveau de confiance nominal de 95 % quand $l = 10$. La couverture s'améliore à mesure que le nombre d'unités primaires d'échantillonnage augmente pour les cas où $\rho = 0$ et $\rho = 0{,}2$. Pour $\rho = 0{,}5$, le défaut de couverture persiste quand la taille d'échantillon croît jusqu'à 100. Nous voyons aussi qu'à $z = 18$ ou 64, la couverture est

plus élevée que le niveau nominal, parce que l'effet de borne de l'estimation par la régression polynomiale locale cause un plus grand biais aux deux bornes des données. Pour $\rho = 0{,}5$, la taille effective d'échantillon est faible, de sorte que l'effet de borne devient sévère, ce qui crée les pics dirigés vers le bas à 18 et à 63.

Soulignons que, si la taille des unités primaires d'échantillonnage est grande (1 000), la fraction d'échantillonnage est très faible (0,05). Par conséquent, ces propriétés des estimations ne changeraient pas même si la taille des unités primaires d'échantillonnage était faible.

5. Exemples empiriques

Revenons maintenant à l'exemple de la section 1. Afin d'illustrer le modèle linéaire partiel, nous examinons les effets de l'âge, du sexe, de la situation d'usage du tabac et de l'activité physique sur l'indice de masse corporelle (IMC) et sur l'indice de masse corporelle désiré (IMCD). Construite comme l'IMC, l'IMCD est une variable dérivée produite d'après les réponses à la question sur le poids souhaité par la personne. Puisque la croissance des personnes faisant partie du groupe d'âge qui nous intéresse est terminée, nous nous servons de la taille réelle pour calculer l'IMCD. Nous utilisons l'âge comme covariable non paramétrique et traitons les autres facteurs comme des variables discrètes. Comme il n'existe que 47 points distincts dans la variable d'âge, nous groupons l'ensemble de données par classe en fonction de l'âge. La taille de classe est fixée à l'unité, de sorte qu'il existe 47 classes, dont les points médians sont $18, 19, \ldots, 64$. Parmi les variables explicatives catégoriques, le sexe possède deux niveaux : masculin = 1 et féminin = 0 ; la situation d'usage du tabac comprend les niveaux tels que ancien fumeur = 0, n'a jamais fumé = 1, fume à l'occasion = 2, fume quotidiennement = 3 ; et l'activité physique est répartie en trois niveaux : personne active = 0, personne moyennement active = 1 et personne inactive = 2. Les modèles de régression sont a) IMC = $g_1(\text{âge}) + \mathbf{XB}_1 + \varepsilon_1$ et b) IMCD = $g_2(\text{âge}) + \mathbf{XB}_2 + \varepsilon_2$, où $\mathbf{X}$ est la matrice de plan contenant toutes les variables indicatrices.

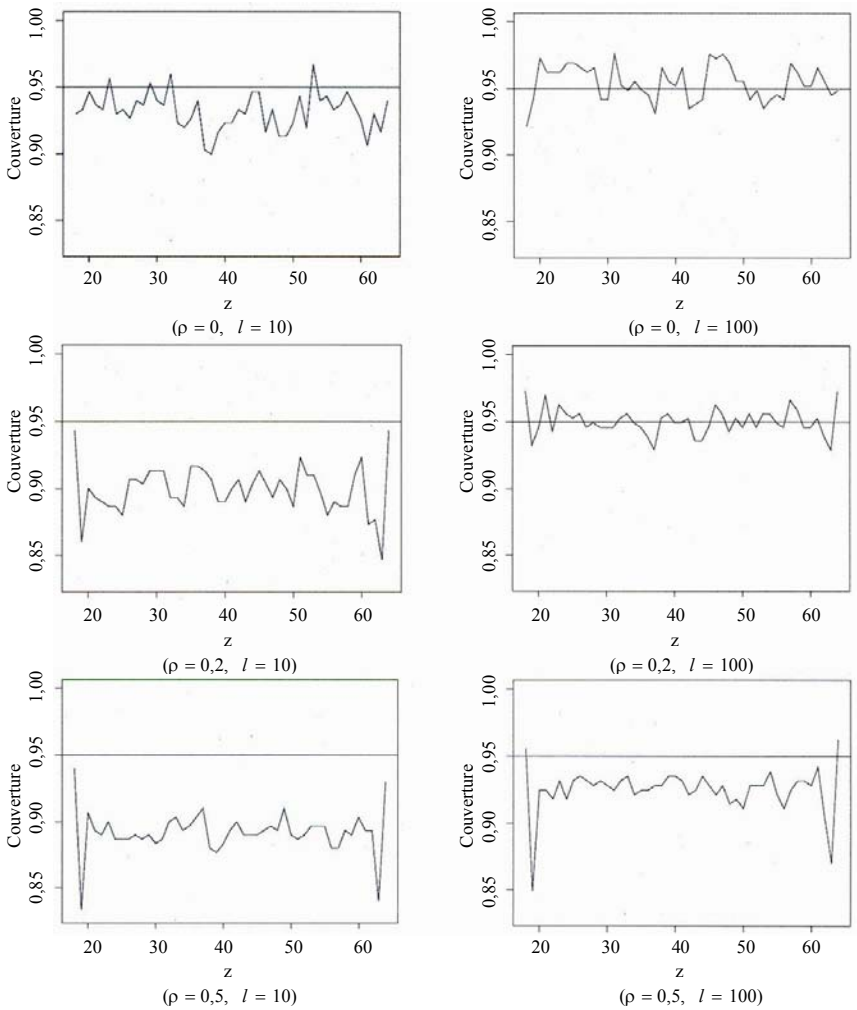

Figure 3 Couverture des intervalles de confiance à 95 % ponctuels pour $g(z)$

Le tableau 3 donne les estimations par sondage des coefficients linéaires dans les modèles (a) et (b). Si nous comparons l'IMC selon le sexe, nous constatons qu'il est plus élevé chez les hommes. Si la catégorie de base est ancien fumeur, les coefficients de la situation d'usage du tabac sont tous négatifs et significatifs, ce qui fait penser que les anciens fumeurs ont tendance à être plus lourds que les personnes qui ont une autre situation d'usage du tabac. Les estimations indiquent aussi que les personnes inactives ont un IMC plus élevé. En ce qui concerne l'IMCD, les valeurs p suggèrent que la plupart des facteurs liés au mode de vie n'ont pas d'effet significatif.

Tableau 3
Résultats pour les modèles de régression semiparamétrique (a) et (b) (les erreurs-types sont indiquées entre parenthèses)

Variable	$\hat{\mathbf{B}}_1$	valeur p	$\hat{\mathbf{B}}_2$	valeur p
Sexe	1,45	0,00	2,80	0,00
	(0,05)		(0,05)	
N'a jamais fumé	-0,45	0,00	-0,06	0,34
	(0,10)		(0,06)	
Fume à l'occasion	-0,31	0,04	-0,00	0,96
	(0,17)		(0,10)	
Fume quotidiennement	-0,61	0,00	-0,12	0,03
	(0,09)		(0,06)	
Moyennement actif	-0,33	0,00	-0,07	0,24
	(0,09)		(0,06)	
Actif	-0,50	0,00	-0,14	0,07

À la figure 4, les fonctions estimées de l'âge, $\hat{g}_1(\text{Âge})$ et $\hat{g}_2(\text{Âge})$, et leurs bandes de confiance sont représentées graphiquement en fonction de l'âge. Nous constatons que, dans les deux cas, l'IMC et l'IMCD sont des fonctions croissantes de l'âge.

La figure 5 donne les fonctions estimées de l'âge, $\hat{g}_1(\text{Âge})$ et $\hat{g}_2(\text{Âge})$, pour les personnes actives et moyennement actives. Si nous examinons l'effet de l'âge séparément chez les femmes et les hommes, nous voyons que chez les femmes actives ou moyennement actives, l'IMCD est, en moyenne, plus faible que l'IMC, tandis que les hommes dont l'intensité d'activité physique est la même manifestent le désir d'augmenter leur poids avant l'âge de 21 ans. En outre, nous comparons les tendances de l'IMC et de l'IMCD selon l'âge chez les femmes ainsi que les hommes. Étant donné la non-convergence de ces tendances, nous pouvons conclure qu'il existe des interactions entre le sexe et l'âge.

6. Conclusion

Au moyen d'un modèle linéaire partiel, nous étendons les méthodes de régression semiparamétrique aux données d'enquêtes complexes. Nous élaborons les propriétés asymptotiques des estimateurs par sondage. Le calcul des estimations de la variance des coefficients linéaires et de la fonction de régression s'appuie sur les estimations de variance des totaux et des moyennes d'enquête. À condition d'obtenir les estimations de variance requises de ces totaux et moyennes, nous pouvons appliquer la méthode en nous servant de progiciels statistiques standard.

Dans le modèle linéaire partiel de travail, nous supposons qu'il n'existe aucune interaction entre la composante paramétrique et la composante non paramétrique. Cependant, l'exemple empirique des tendances de l'indice de masse corporelle selon l'âge montre qu'il faut vérifier cette hypothèse. Dans de futurs travaux, nous relâcherons l'hypothèse d'absence d'interaction. Une approche directe de la modélisation des termes d'interaction consiste à laisser la composante non paramétrique paraître linéaire dans le terme d'interaction. Autrement dit, nous définissons le modèle linéaire partiel sous la forme

$$\mathbf{y} = G(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \mathbf{X}H(\mathbf{z}) + \boldsymbol{\varepsilon}.$$

En testant l'écart de $H(\mathbf{z})$ par rapport à zéro, nous pouvons déceler l'existence d'une interaction.

Pour estimer l'espérance conditionnelle sur les composantes non paramétriques pour des variables indicatrices aléatoires discrètes, nous proposons d'utiliser des modèles linéaires ou additifs généralisés.

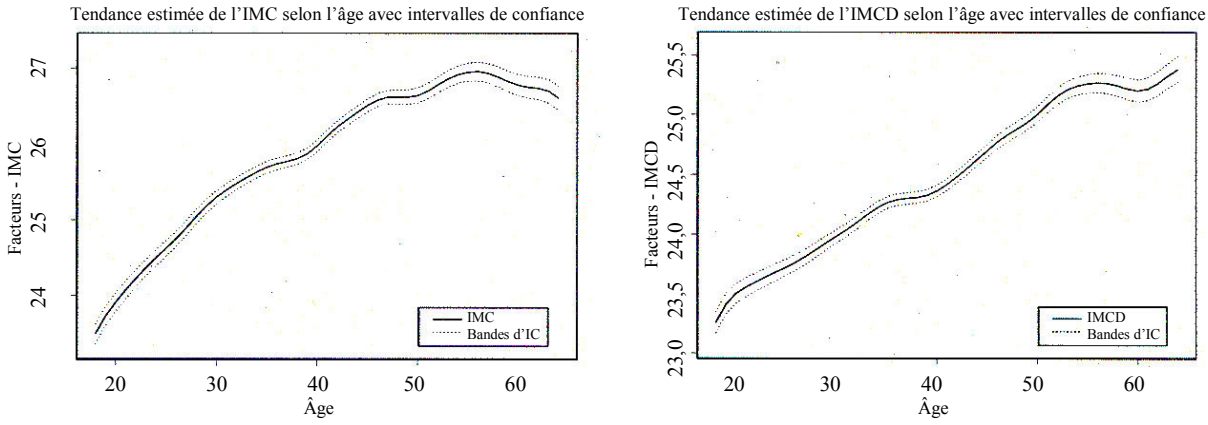

Figure 4 Tendances estimées de l'IMC et de l'IMCD selon l'âge avec intervalles de confiance à 95 % ponctuels

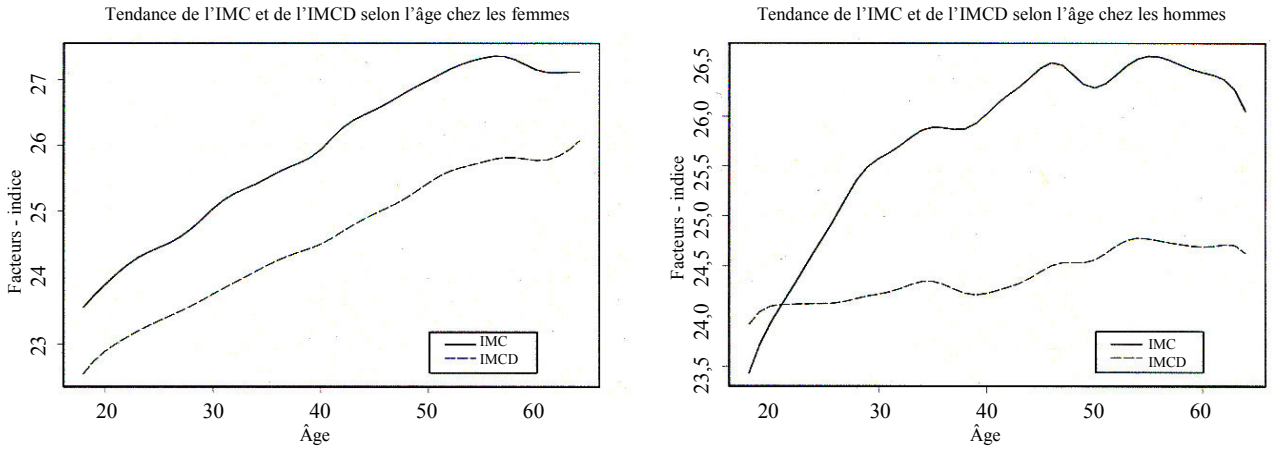

Figure 5 Comparaison des tendances estimées de l'IMC et de l'IMCD selon l'âge chez les femmes et chez les hommes qui sont actifs ou moyennement actifs

Annexe

A.1 Preuve du lemme 1

En observant que les entrées de $\hat{\mathbf{u}}(\boldsymbol{\theta})$, $\hat{\mathbf{u}}_B(\boldsymbol{\theta})$ et $\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ sont soit des totaux d'échantillon ou des ratios de totaux d'échantillon, nous pouvons appliquer les lemmes 1.2.5 et 1.2.6 présentés dans Wang (2004) pour établir ce lemme.

A.2 Preuve du lemme 2

Chaque entrée de $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ est simplement une fonction de régression estimée par la méthode des polynômes locaux établie par Bellhouse et Stafford (2001). Le théorème 2.2.1 dans Wang (2004) montre que $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ est convergent à la vitesse racine carrée de n. Donc, puisque la dimension de $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ est finie, nous pouvons montrer que $\sqrt{n}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z}))$ est borné en probabilité.

A.3 Preuve du théorème 1

Puisque, pour le vrai $\boldsymbol{\theta}$, nous avons $\mathbf{u}(\boldsymbol{\theta}) = \mathbf{0}$, nous pouvons réécrire l'équation (13) comme il suit :

$$-\frac{\sqrt{n}\hat{\mathbf{u}}_B(\boldsymbol{\theta})}{N}(\hat{\mathbf{B}} - \mathbf{B}) \doteq$$

$$\left(\frac{\sqrt{n}}{N}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})\frac{\sqrt{n}}{N}\left(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})\right)\right)$$

$$+ \frac{\sqrt{n}}{N}||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||\varepsilon.$$

L'argument standard présenté dans Rao (1973, page 387) donne

$$\sqrt{n}/N\,||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||\varepsilon \xrightarrow{P} 0.$$

En utilisant la condition voulant que la fraction d'échantillonnage $f = n/N$ soit constante quand ν tend vers l'infini, nous avons

$$\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) = -\left(\frac{\hat{\mathbf{u}}_B(\boldsymbol{\theta})}{N}\right)^{-1}$$

$$\left(\frac{\sqrt{n}}{N}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})\frac{f\sqrt{n}}{n}\left(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})\right)\right).$$

Il découle des résultats du lemme 1 que $(\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})/N)^{-1}$ et $\hat{\mathbf{U}}_{\mathbf{m}_{\boldsymbol{\xi}}}(\boldsymbol{\theta})$ convergent tous deux en probabilité vers leurs valeurs de population. Le lemme 2 indique que le vecteur $\sqrt{n}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(z) - \mathbf{m}_{\boldsymbol{\xi}}(z)) = O_p(1)$. Donc $(f\sqrt{n}/n)(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z}))$ converge en probabilité vers un vecteur nul quand n tend vers l'infini. Enfin, partant de la normalité de $\sqrt{n}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta}))/N$ énoncée dans le lemme 1, nous utilisons le théorème de Slutsky pour démontrer la normalité asymptotique de $\hat{\mathbf{B}}$.

A.4 Preuve du lemme 3

Sachant que $\bar{\mathbf{r}} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}$, nous avons $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) = \sqrt{n}[(\bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}) - \bar{\mathbf{R}}]$. D'après le théorème 1, nous savons que, à la limite quand ν tend vers l'infini, $\hat{\mathbf{B}}$ converge vers $\mathbf{B}$ en probabilité. Donc, nous avons $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) \doteq \sqrt{n}((\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B}) - \bar{\mathbf{R}})$. La d^{e} entrée de $(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B})$ est

$$\bar{y}_d - \bar{x}_{1d}B_1 - \cdots - \bar{x}_{pd}B_p =$$

$$\frac{1}{\hat{N}_d}\sum\nolimits_{k\in s_d} w_k(y_k - x_{1k}B_1 - \cdots - x_{pk}B_p).$$

Autrement dit, $(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B})$ est simplement un vecteur des moyennes par classe estimées. En utilisant le résultat de Shao (1996) sur les fonctions des moyennes d'échantillon et

les résultats de l'« astuce de Cramér-Wold » (*Cramér-Wold device*) donnés dans Serfling (1980, page 18), nous voyons que $\sqrt{n}(\overline{\mathbf{y}} - \overline{\mathbf{x}}\mathbf{B} - \overline{\mathbf{R}})$ converge vers un vecteur aléatoire normalement distribué. Donc, en utilisant cette idée indirecte de Slutsky, nous avons prouvé la normalité de $\sqrt{n}(\overline{\mathbf{r}} - \overline{\mathbf{R}}) = \sqrt{n}[(\overline{\mathbf{y}} - \overline{\mathbf{x}}\hat{\mathbf{B}}) - \overline{\mathbf{R}}]$.

A.5 Preuve du théorème 2

La preuve découle du même argument que dans la preuve du théorème 2.2.1 de Wang (2004), selon lequel $\hat{g}(z_d)$ est une fonction de la moyenne et des proportions de domaine.

Remerciements

Les présents travaux ont été financés par une subvention du Conseil de recherches en sciences naturelles et en génie (CRSNG) du Canada. L'auteur remercie Mary Thompson de ses suggestions et commentaires constructifs formulés lors d'une version antérieure du présent article. Les auteurs remercient aussi le rédacteur associé et les deux examinateurs de leurs commentaires très utiles.

Bibliographie

Bellhouse, D.R., et Stafford, J.E. (1999). Density estimation from complex survey. *Statistca Sinica*, 9, 407-424.

Bellhouse, D.R., et Stafford, J.E. (2001). Régression polynomiale locale dans le cas des enquêtes complexes. *Techniques d'enquête*, 27, 219-226.

Bickel, P.J., et Freedman, D.A. (1983). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

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

Buskirk, D.T., et Lohr, L.S. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Fan, J., et Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society*, Séries B, 57, 371-394.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, *C*, 37, 117-132.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Hàjek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koelemenyei*, 5, 361-374.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Jones, M.C. (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84, 733-741.

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

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535–545.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.

Ontario Health Survey (1992). *Ontario Health Survey*: *User's Guide.* Ministry of Health, Toronto, Ontario, Canada.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2ième Éd.). New York : John Wiley & Sons, Inc.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56, 931-954.

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

Serfling, R.J. (1980). *Approximation Theorem of Mathematical Statistics*. New York : John Wiley & Sons, Inc.

Shao, J. (1996). Resampling methods in sample survey. *Statistics*, 27, 203-254.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society*, Séries B, 50, 413-436.

Thompson, M.E. (1997). *Theory of Sample Survey* (1ière Éd.). New York : Chapman and Hall.

Wang, Z. (2004). *Some Nonparametric Regression Techniques for Complex Survey Data*. Thèse de doctorat non-publiée, The University of Western Ontario, Londres, Ontario, Canada.

Zheng, H., et Little, R.J.A. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 233-243.