N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

TECHNIQUE DE RÉGRESSION SEMIPARAMÉTRIQUE POUR LES DONNÉES D'ENQUÊTE COMPLEXE

Zilin Wang et David Bellhouse[1]

RÉSUMÉ

Étant donné la complexité habituelle des plans d'échantillonnage, les données d'enquête ne sont ni indépendantes ni identiquement distribuées. Par conséquent, les méthodes d'estimation élaborées pour des données indépendantes et identiquement distribuées ne conviennent pas pour des modèles fondés sur des données d'enquête complexe. Le but du présent article est de discuter de ce genre de méthodes d'estimation en mettant l'accent sur les modèles de régression semiparamétrique pour les données d'enquête complexes dans lesquels les variables explicatives comprennent des éléments paramétriques et non paramétriques. Les méthodes d'estimation applicables à ces modèles combinent l'estimation par régression polynomiale locale non paramétrique et l'estimation classique par les moindres carrés dans les enquêtes complexes. Nous discutons des moments et des propriétés asymptotiques des estimateurs. Enfin, nous illustrons la méthodologie et les résultats théoriques au moyen des données de l'Enquête sur la santé en Ontario de 1990.

MOTS CLÉS : Échantillonnage, enquête sur la santé en Ontario, fonction de décalage, groupement par classe (binning), lissage, régression.

1. INTRODUCTION

Si **y** est la variable réponse et **X**, une matrice des variables explicatives correspondantes, un modèle de régression classique est de la forme :

$$E(\mathbf{y}|\mathbf{X}) = G(\mathbf{X}),$$

où G(.) est appelé fonction de régression. Selon l'hypothèse émise au sujet de la fonction de régression, la modélisation par régression peut être subdivisée en deux grandes catégories, à savoir la régression paramétrique et la régression non paramétrique. Une régression paramétrique est paramétrisée au moyen d'un vecteur de paramètres inconnu de dimension *p*, dénoté $\boldsymbol{\beta}$ et G(**X**) est habituellement dénotée G(**X,** $\boldsymbol{\beta}$). Nous estimons $\boldsymbol{\beta}$ en faisant une hypothèse sur la forme fonctionnelle de G(. **,** .). Si l'hypothèse concernant la forme de G(. , .) est correcte, les propriétés du modèle de régression paramétrique sont fort utiles, mais, si elle est erronée, le modèle peut donner des résultats incorrects. Dans le cas d'un modèle de régression non paramétrique, nous atténuons l'hypothèse concernant la forme de G(.) et nous utilisons l'information locale pour obtenir les estimations ponctuelles de la fonction G(.). Un modèle de régression non paramétrique peut être estimé au moyen d'un lisseur. Bien que l'utilité des techniques de régression non paramétrique ait été démontrée, quand nous procédons à une modélisation par régression souple, nous payons un prix pour le relâchement de l'hypothèse d'une forme fonctionnelle particulière dans l'analyse de régression non paramétrique. Plus précisément, outre la difficulté qu'il y a à choisir la taille correcte de fenêtre (voisinage), nous devons résoudre le problème plus grave du « fléau de la dimensionnalité » que posent toutes les méthodes de lissage pour un modèle de régression multiple, problème qui se manifeste lorsque les voisinages comptant un nombre fixe de points deviennent moins locaux à mesure que le nombre de dimensions augmente. Le « fléau de la dimensionnalité » rend la vitesse de convergence de l'estimateur si faible que les propriétés de l'estimation non paramétrique ne sont pas prometteuses pour les régressions multiples. L'un des résultats de ce « fléau de la dimensionnalité » est qu'il est impossible d'inclure des variables explicatives discrètes dans l'analyse par régression non paramétrique.

[1] Zilin Wang et David Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada, N6A 5B7.

Pour profiter de la force de l'estimation paramétrique et réduire au minimum le « fléau de la dimensionnalité », nous élaborons ce qu'il est convenu d'appeler un modèle de régression semiparamétrique partiellement linéaire de la forme

$$E(\mathbf{y}|\mathbf{X},\mathbf{z}) = \mathbf{X}\boldsymbol{\beta} + G(\mathbf{z}),$$

où les variables explicatives sont représentées en deux parties, la partie non paramétrique ($G(\mathbf{z})$) et la partie linéaire paramétrique ($\mathbf{X}\boldsymbol{\beta}$). Dans ce modèle de régression semiparamétrique, nous estimerons la forme fonctionnelle de la partie non paramétrique du modèle ainsi que les paramètres. Ce modèle semiparamétrique partiellement linéaire a pour justification a priori le fait qu'il s'agit d'un outil d'analyse des données et qu'il conserve une caractéristique d'interprétation importante. Nous pouvons placer dans la partie paramétrique du modèle les variables pour lesquelles nous possédons le plus d'information connue sur la forme fonctionnelle et dans la partie non paramétrique du modèle, celles pour lesquelles nous en possédons peu. En outre, les variables explicatives discrètes ont toujours posé des problèmes lors de l'estimation par régression non paramétrique, à cause de tailles d'échantillon peu efficaces. Il est fort naturel d'inclure les variables explicatives discrètes dans la partie linéaire du modèle.

L'objectif du présent article est d'appliquer ce modèle de régression semiparamétrique partiellement linéaire à des enquêtes complexes. Nous nous intéressons aux méthodes d'estimation élaborées indépendamment par Robinson (1988) et par Speckman (1988) pour des données indépendantes et identiquement distribuées. Étant donné le plan d'échantillonnage, les données provenant d'une enquête complexe ne sont ni indépendantes ni identiquement distribuées. Donc, nous ne pouvons appliquer directement la méthode d'estimation de Robinson (1988) ni celle de Speckman (1988) à ce genre de données. Pour résoudre la difficulté technique que posent les données complexes, nous établissons deux superpopulations de façon à ce que nous puissions adapter la méthode d'estimation établie pour des données indépendantes et identiquement distribuées et faire des inférences pour les estimateurs fondés sur les échantillons d'enquête.

Telle qu'elle a été établie par Robinson (1988) et par Speckman (1988), la méthode d'estimation pour les modèles de régression partiellement linéaires consiste en une méthode d'estimation non paramétrique et une méthode d'estimation par les moindres carrés. Par conséquent, nous avons besoin d'un lisseur pour appliquer la méthode d'estimation dans le contexte de l'échantillonnage. Celui que nous utiliserons a été développé par Bellhouse et Stafford (2001) pour les enquêtes complexes. L'une des caractéristiques des données d'enquête complexes est que la taille de l'ensemble de données peut être très grande. Habituellement, dans un grand ensemble de données d'enquête, il existe des observations multiples pour certaines valeurs. Les grands ensembles de données peuvent non seulement produire des tendances non informatives de la relation entre la variable réponse et les covariables lorsqu'on représente graphiquement les données, mais peuvent aussi rendre le calcul des estimations très fastidieux. Donc, il est fort naturel, lors de l'analyse de données d'enquête complexes, de grouper les données par domaine selon les valeurs distinctes des variables caractéristiques. Bellhouse et Stafford (2001) proposent l'application de méthodes de régression polynomiale locale aux données d'enquête à grande échelle et recourent au groupement des données par classe selon les valeurs de la variable explicative.

En combinant la méthode bien établie d'estimation par les moindres carrés et les techniques de régression polynomiale locale élaborées par Bellhouse et Stafford (2001) pour les enquêtes complexes, nous développons les estimateurs fondés sur l'échantillon d'enquête pour le modèle de régression partiellement linéaire et établissons leurs propriétés asymptotiques. La présentation de l'article est la suivante. À la section 2, nous introduisons le modèle de régression partiellement linéaire dans le contexte de l'échantillonnage. À la section 3, nous discutons des propriétés asymptotiques des estimateurs fondés sur l'échantillon d'enquête. À la section 4, nous donnons une illustration empirique de la méthode d'estimation à l'aide des données de l'Enquête sur la santé en Ontario de 1990. Enfin, à la section 5, nous présentons nos conclusions.

2. UN MODÈLE PARTIELLEMENT LINÉAIRE DANS LE CONTEXTE DE L'ÉCHANTILLONNAGE

2.1 Préliminaire

Un modèle semiparamétrique est défini comme suit :

$$\mathbf{y} = \mathrm{G}(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

où $\mathbf{y}$ est le vecteur de variable réponse et les $\boldsymbol{\varepsilon}$ sont indépendantes et identiquement distribuées de moyenne zéro et de variance constante. G(.) est une fonction arbitraire de $\mathbf{z}$. D'après l'information du modèle sur les variables indépendantes, ces dernières sont réparties en deux catégories. Les variables indépendantes incluses dans la matrice $\mathbf{X}$ de dimensions $n \times p$ correspondent à la partie paramétrique ou linéaire du modèle et la variable indépendante, $\mathbf{z}$, est la partie non paramétrique du modèle. Chaque variable indépendante paramétrique, $\mathbf{x}_j$, est un vecteur de variables aléatoires dont la loi est F_j. $\mathbf{z}$ est mesurée sur une échelle continue et $\mathbf{X}$ contient des variables explicatives continues ou discrètes. La forme fonctionnelle de G(.) ainsi que les paramètres $\beta_1, \ldots, \beta_p$ sont inconnus. En outre, nous supposons que $\mathrm{E}(\boldsymbol{\varepsilon}|\, \mathbf{z}, \mathbf{X})=\mathbf{0}$ et qu'il n'existe aucune interaction entre $\mathbf{X}$ et $\mathbf{z}$.

Le problème que pose l'estimation de $\boldsymbol{\beta}$ dans le modèle partiellement linéaire énoncé en (1) est que la forme de la fonction $\mathrm{G}(\mathbf{z})$ est inconnue. S'il existait un moyen d'éliminer cette fonction, la méthode des moindres carrés pourrait être utilisée pour estimer le modèle de régression linéaire résultant.

En prenant l'espérance des deux membres de l'équation (1) conditionnellement à $\mathbf{z}$, nous obtenons

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

sachant que $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{z}) = \mathbf{0}$. Maintenant, soustrayons (2) de (1) pour obtenir

$$\mathbf{y} - \mathrm{E}(\mathbf{y} \mid \mathbf{z}) = (\mathbf{X} - \mathrm{E}(\mathbf{X} \mid \mathbf{z}))\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

En définissant $\boldsymbol{Y} \equiv \mathbf{y} - \mathrm{E}(\mathbf{y}|\mathbf{z})$ et $\boldsymbol{X} \equiv \mathbf{X} - \mathrm{E}(\mathbf{X}|\mathbf{z})$, nous obtenons le modèle de régression linéaire

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

Or, une démarche évidente consiste à estimer $\boldsymbol{\beta}$ par la méthode des moindres carrés. Malheureusement, puisque $\mathrm{E}(\mathbf{y}|\mathbf{z})$ et $\mathrm{E}(\mathbf{X}|\mathbf{z})$ sont inconnues, l'estimation de $\boldsymbol{\beta}$ par les moindres carrés n'est pas réalisable. Par conséquent, nous estimons $\boldsymbol{\beta}$ en deux étapes. À la première étape, nous estimons les espérances conditionnelles de l'équation (3) par la technique de lissage par la méthode du noyau de Nadaya-Watson. À la deuxième étape, nous remplaçons $\mathrm{E}(\mathbf{y}|\mathbf{z})$ et $\mathrm{E}(\mathbf{X}|\mathbf{z})$ dans l'équation (3) par leurs estimations obtenues à la première étape et nous estimons $\boldsymbol{\beta}$ par la méthode des moindres carrés.

Une fois que nous avons obtenu l'estimation de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, nous traitons la différence entre la variable réponse $\mathbf{y}$ et les $\mathbf{X}\hat{\boldsymbol{\beta}}$ comme étant la variable aléatoire dépendante et nous estimons la fonction $\mathrm{G}(\cdot)$ conformément au modèle suivant,

$$\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathrm{G}(\mathbf{z}) + \boldsymbol{\mu} \tag{5}$$

L'avantage de cette méthode semiparamétrique applicable à des données indépendantes et identiquement distribuées est qu'il n'est pas nécessaire de recourir à l'itération et que l'on peut obtenir des estimateurs des coefficients linéaires qui convergent à la vitesse racine carrée de n.

2.2 Plan d'échantillonnage et superpopulations

Supposons que nous ayons une population $\boldsymbol{U}$ comprenant N unités distinctes. La caractéristique d'intérêt est une unité évaluée vectoriellement $(y_k, \boldsymbol{x}_k, z_k)$ pour tous les $k=1,\ldots,N$. y_k représente la k^e valeur de population de la variable réponse, et $(\boldsymbol{x}_k, z_k)$ représente la k^e observation des variables explicatives et est un vecteur de longueur $p+1$. Soit $\mathbf{s}$ un ensemble d'unités dans l'échantillon avec $(y_k, \boldsymbol{x}_k, z_k, w_k)$ pour $k \in \boldsymbol{s}$ obtenu conformément au plan d'échantillonnage pour une taille d'échantillon n. Le poids de sondage w_k est associé à la k^e unité d'échantillonnage. De plus, nous supposons que la non-réponse est nulle pour nous assurer que les probabilités d'inclusion soient égales à l'inverse des poids d'échantillonnage.

Notons qu'il est nécessaire de recourir à plusieurs méthodes pour obtenir l'estimation du modèle partiellement linéaire. Afin de produire des estimations et de faire des inférences d'après celles-ci, nous devons définir un cadre de superpopulation hypothétique. Habituellement, lors de l'analyse de données d'enquête, on émet l'hypothèse d'un modèle de travail ou d'un modèle de superpopulation à partir de la population finie. Les estimations des paramètres de ce modèle donnent des paramètres de population finie ou des estimations dans des conditions de recensement fondées sur le modèle. On utilise l'échantillon d'enquête pour obtenir des estimations de ces « estimations dans des conditions de recensement ». Normalement, on produit les calculs asymptotiques pour justifier les inférences au sujet de la population à partir de l'échantillon au moyen d'un deuxième modèle de superpopulation.

Superpopulation 1

Les N unités de population finie sont un échantillon d'unités indépendantes et identiquement distribuées tirées de la superpopulation infinie. Les unités d'une population finie sont des réalisations du modèle défini par l'équation (1). Nous dénotons respectivement par $\mathbf{B} = (B_1, \ldots, B_p)$ et $g(z)$ les paramètres de population finie des coefficients linéaires et la fonction de régression au point fixe z dans le modèle de travail. En nous fondant sur cette superpopulation, nous pouvons adopter directement la méthode applicable aux données indépendantes et identiquement distribuées pour obtenir les paramètres de population finie d'intérêt que nous pouvons estimer. Notre seule préoccupation est que les paramètres de population finie soient des estimateurs convergents des paramètres de la superpopulation si l'on supprime l'hypothèse d'indépendance. La superpopulation 1 nous permet de produire des résultats asymptotiques valides dans le contexte d'indépendance, mais elle peut créer des problèmes de spécification si les unités de population finie ne concordent pas avec le modèle de superpopulation. Par conséquent, une fois que nous avons obtenu les estimateurs d'échantillon, nous utilisons une autre superpopulation, qui est composée d'une série emboîtée de populations finies, pour établir la loi asymptotique et les inférences à partir des estimateurs.

Superpopulation 2

La superpopulation 2 comprend une série emboîtée de populations finies indicées par υ de sorte que toutes les quantités de la population finie et les quantités d'échantillon dépendent de l'indice υ et que tous les cadres asymptotiques soient établis pour $\upsilon \to \infty$.

2.3. Estimation

En utilisant le cadre de la superpopulation 1, nous étendons la méthode d'estimation de la section 2.1 aux données d'enquête complexe avec certaines modifications. Plus précisément, au lieu d'utiliser la technique de lissage par la méthode du noyau de Nadaya-Watson décrite dans Robinson (1988), nous utiliserons la technique de régression polynomiale locale pour estimer les espérances conditionnelles. Comme le mentionne Wand et Jones (1995), la technique de lissage par la méthode du noyau de Nadaya-Watson peut être considérée comme un ajustement d'une constante locale et on a montré qu'elle produit un biais aux bornes plus important que certains ajustements de modèles de régression polynomiale locale d'un autre degré. Il convient de souligner que l'estimation des coefficients linéaires et de la fonction de régression non paramétrique se fait en deux temps. À la première étape, nous estimons les coefficients linéaires et à la seconde, la fonction non paramétrique.

Durant la première étape de la méthode d'estimation, il est nécessaire d'utiliser un lisseur pour estimer les espérances conditionnelles de la variable réponse et des variables explicatives paramétriques sur la variable

explicative non paramétrique, **z**. Nous dénotons respectivement $m_y(z)$ et $m_{x_j}(z)$ les espérances conditionnelles de population de **y** et $\mathbf{x}_j$ à un point fixe *z*. Pour estimer $m_y(z)$ et $m_{x_j}(z)$, nous groupons par classe les données observées en fonction de **z**. Supposons que **z** possède *m* valeurs distinctes dans la population finie. Représentons par z_i la i^e valeur distincte ou la i^e classe (ou fenêtre) et supposons que les valeurs de z_i soient séparées par des intervalles égaux de longueur $z_i - z_{i-1}$. La proportion d'observations dans la population finie dont la valeur est z_i est représentée par p_i. Soit $\bar{\mathbf{y}} = (\bar{y}_1, \cdots, \bar{y}_m)$ le vecteur des moyennes de population finie pour la variable réponse *y* calculées pour des valeurs distinctes de *z* et $\bar{\mathbf{x}}_j = (\bar{x}_{1j}, \cdots, \bar{x}_{mj})$ le vecteur des moyennes de population finie pour la j^e variable indépendante $\mathbf{x}_j$ pour $j = 1, \ldots, p$. $\hat{\bar{y}}_i$, $\hat{\bar{x}}_{ij}$ et $\hat{\bar{p}}_i$ sont respectivement les estimateurs d'enquête de $\bar{y}_i$, $\bar{x}_{ij}$ et $\bar{p}_i$ pour tout $i = 1, \ldots, m$.

D'après les moyennes et les proportions d'échantillon calculées par classe, l'estimateur d'après les données d'enquête de $m_y(z)$ et $m_{x_j}(z)$ à la valeur z_i pour tout $i = 1, \ldots, m$ est de la forme,

$$\hat{E}(\mathbf{x}_j \mid z = z_i) = \hat{m}_{\mathbf{x}_j}(z_i) = \mathbf{e}^T (\mathbf{Z}_{z_i}^T \hat{\mathbf{KW}}_{z_i} \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^T \hat{\mathbf{KW}}_{z_i} \hat{\bar{\mathbf{x}}}_j , \tag{6}$$

et

$$\hat{E}(\mathbf{y} \mid z = z_i) = \hat{m}_{\mathbf{y}}(z_i) = \mathbf{e}^{\mathbf{T}} (\mathbf{Z}_{z_i}^{\mathbf{T}} \hat{\mathbf{KW}}_{z_i} \mathbf{Z}_{z_i})^{-1} \mathbf{Z}_{z_i}^{\mathbf{T}} \hat{\mathbf{KW}}_{z_i} \hat{\bar{\mathbf{y}}} , \tag{7}$$

où

$$\mathbf{Z}_{z_i} = \begin{pmatrix} 1 & z_1 - z_i & \cdots & (z_1 - z_i)^q \\ 1 & z_2 - z_i & \ldots & (z_2 - z_i)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_m - z_i & \cdots & (z_m - z_i)^q \end{pmatrix}$$

et

$$\hat{\mathbf{KW}}_{z_i} = \frac{1}{h} \mathbf{diag}\left(\hat{p}_1 K\left(\frac{z_1 - z_i}{h}\right), \ldots, \hat{p}_m K\left(\frac{z_m - z_i}{h}\right)\right) .$$

$K(\cdot)$ est une fonction noyau telle que $\int K(t)\, dt = 1$ et $\int K(t)^2\, dt < \infty$ et *q* est le degré du polynôme auquel est ajusté le modèle. *h* est la largeur de bande qui contrôle la taille du voisinage. Le vecteur **e** est le vecteur de dimension $m \times 1$ de la forme $(1,0, \ldots, 0)^T$. Le vecteur $\hat{\bar{\mathbf{y}}}$ est le vecteur de dimension $m \times 1$ de la forme $(\hat{\bar{y}}_1, \ldots, \hat{\bar{y}}_m)^T$, et $\hat{\bar{\mathbf{x}}}_j$ est la matrice de dimensions $m \times 1$ de la forme $(\hat{\bar{x}}_{1j}, \cdots, \hat{\bar{x}}_{mj})^T$.

Afin d'estimer la population finie **B**, nous devons reconstruire les données de façon à ce que le modèle de travail représenté par l'équation (3) puisse être utilisé. Soit N_i le nombre d'observations qui tombent dans la i^e classe et $\sum_{i=1}^{m} N_i = N$. $\mathbf{M}_{\mathbf{x}}$ est une matrice de dimensions $N \times p$ contenant toutes les espérances conditionnelles de population de $\mathbf{X}$ et dont la forme est

$$\mathbf{M}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} \mathrm{m}_{\mathbf{x}_1}(z_1) & \mathrm{m}_{\mathbf{x}_2}(z_1) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{m}_{\mathbf{x}_1}(z_1) & \mathrm{m}_{\mathbf{x}_2}(z_1) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_1) \end{pmatrix}_{N_1 \times p} \\ \vdots \quad\quad\quad\quad \vdots \\ \begin{pmatrix} \mathrm{m}_{\mathbf{x}_1}(z_m) & \mathrm{m}_{\mathbf{x}_2}(z_m) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_m) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{m}_{\mathbf{x}_1}(z_m) & \mathrm{m}_{\mathbf{x}_2}(z_m) & \cdots & \mathrm{m}_{\mathbf{x}_p}(z_m) \end{pmatrix}_{N_m \times p} \end{pmatrix} \tag{8}$$

De la même façon, nous établissons un vecteur de dimension $N \times 1$, $\mathbf{M}_{\mathbf{y}}$, tel que $\mathrm{m}_{\mathbf{y}}(z_i)$ est répété N_i fois dans la i^e classe,

$$\mathbf{M}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \mathrm{m}_{\mathbf{y}}(z_1) \\ \vdots \\ \mathrm{m}_{\mathbf{y}}(z_1) \end{pmatrix}_{N_1 \times 1} \\ \vdots \\ \begin{pmatrix} \mathrm{m}_{\mathbf{y}}(z_m) \\ \vdots \\ \mathrm{m}_{\mathbf{y}}(z_m) \end{pmatrix}_{N_m \times 1} \end{pmatrix} \tag{9}$$

Soit $\mathbf{y} = (y_1, \ldots, y_N)$, $\mathbf{x}_j = (x_{j1}, \ldots, x_{jN})$ et $\mathbf{X}=(\mathbf{x}_1, \ldots, \mathbf{x}_p)$. En utilisant $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$, nous avons $\boldsymbol{Y} \equiv \mathbf{y} - \mathbf{M}_{\mathbf{y}}$ et $\boldsymbol{X} \equiv \mathbf{X} - \mathbf{M}_{\mathbf{X}}$. Au moyen de ces données transformées et du modèle de travail représenté par l'équation (3)**,** nous obtenons les estimations dans des conditions de recensement par régression multiple par la méthode des moindres carrés sans terme constant :

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

De façon analogue au modèle de la population, nous représentons les données de l'échantillon par (**X**, **y, z,** ***w***) avec une taille d'échantillon de n et n_i observations dans chaque classe, de sorte que $\sum_{i=1}^{m} n_i = n$. **y** est un vecteur de dimension $n \times 1$ et de la forme $(y_1, \ldots, y_n)$ et **x** est une matrice de dimensions $n \times p$ et de la forme $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Nous pouvons construire $\hat{\mathbf{M}}_{\mathbf{X}}$ et $\hat{\mathbf{M}}_{\mathbf{y}}$ de la même façon que nous avons construit $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$ dans les équations (8) et (9). Autrement dit, nous utilisons les estimations de l'échantillon $\hat{\mathrm{m}}_{\mathbf{x}_j}(z_i)$ et $\hat{\mathrm{m}}_{\mathbf{y}}(z_i)$ données par les équations (6) et (7) pour obtenir

$$\hat{\mathbf{M}}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{x}_1}(z_1) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_1) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\mathrm{m}}_{\mathbf{x}_1}(z_1) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_1) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_1) \end{pmatrix}_{N_1 \times \boldsymbol{p}} \\ \vdots \quad\quad\quad\quad \vdots \\ \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{x}_1}(z_m) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_m) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_m) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\mathrm{m}}_{\mathbf{x}_1}(z_m) & \hat{\mathrm{m}}_{\mathbf{x}_2}(z_m) & \cdots & \hat{\mathrm{m}}_{\mathbf{x}_p}(z_m) \end{pmatrix}_{N_m \times \boldsymbol{p}} \end{pmatrix} \text{ et } \hat{\mathbf{M}}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{y}}(z_1) \\ \vdots \\ \hat{\mathrm{m}}_{\mathbf{y}}(z_1) \end{pmatrix}_{N_1 \times 1} \\ \vdots \\ \begin{pmatrix} \hat{\mathrm{m}}_{\mathbf{y}}(z_m) \\ \vdots \\ \hat{\mathrm{m}}_{\mathbf{y}}(z_m) \end{pmatrix}_{N_m \times 1} \end{pmatrix}$$

En définissant $\hat{\boldsymbol{Y}} = \mathbf{y} - \hat{\mathbf{M}}_{\mathbf{y}}$ et $\hat{\boldsymbol{X}} = \mathbf{X} - \hat{\mathbf{M}}_{\mathbf{X}}$, nous obtenons l'estimateur de **B** dans le contexte de l'enquête complexe :

$$\hat{\mathbf{B}} = (\hat{\boldsymbol{X}}^{\mathrm{T}}\,\mathbf{W}\,\hat{\boldsymbol{X}})^{-1}\,\hat{\boldsymbol{X}}^{\mathrm{T}}\,\mathbf{W}\hat{\boldsymbol{Y}}.$$

où **W** est la matrice de poids de dimensions $n \times n$ avec le poids de sondage, w_k, sur la diagonale.

Une fois que nous obtenons l'estimateur d'échantillon, $\hat{\mathbf{B}}$, nous pouvons estimer le paramètre de population g(·) en prenant l'équation (5) comme modèle de travail. En appliquant de nouveau la technique de régression polynomiale locale et en utilisant les estimations de l'échantillon $\hat{\mathbf{B}}$, nous obtenons

$$\hat{\mathrm{g}}(z_i) = \mathbf{e}^{\boldsymbol{T}}(\mathbf{Z}_{z_i}^{\boldsymbol{T}}\,\hat{\mathbf{K}}\mathbf{W}_{z_i}\,\mathbf{Z}_{z_i})^{-1}\mathbf{Z}_{z_i}^{\boldsymbol{T}}\,\hat{\mathbf{K}}\mathbf{W}_{z_i}\,\hat{\bar{\mathbf{R}}}, \tag{10}$$

où $\hat{\mathbf{R}} = \mathbf{y} - \mathbf{x}\hat{\mathbf{B}}$ et $\hat{\bar{\mathbf{R}}}$ est le vecteur des moyennes calculées par classe de l'estimation d'après les données d'enquête $\hat{\mathbf{R}}$.

3. PROPRIÉTÉS ASYMPTOTIQUES

Soit $\boldsymbol{\theta}^T = (\mathbf{B}^T{}_{1\times p}, \mathbf{m}_{\mathbf{x}}(\mathbf{z})^T{}_{1\times pm}, \mathbf{m}_{\mathbf{y}}(\mathbf{z})^T{}_{1\times m})$ un vecteur de taille 1× (p+(p+1)m) contenant tous les paramètres de population finie. En nous inspirant de Binder (1983) et en fondant le modèle de travail sur l'équation (3), nous pouvons exprimer les paramètres de population finie au moyen d'une équation normale de la façon suivante

$$\mathbf{u}(\boldsymbol{\theta}) \equiv \sum_{k=1}^{N}(\mathbf{x}_k - \mathbf{M}_{\mathbf{x}k})^{\boldsymbol{T}}(y_k - \mathrm{M}_{\mathbf{y}k}) - \sum_{i=1}^{N}(\mathbf{x}_k - \mathbf{M}_{\mathbf{x}\boldsymbol{k}})^{\boldsymbol{T}}(\mathbf{x}_k - \mathbf{M}_{\mathbf{x}k})\mathbf{B} = \mathbf{0}_{p\times 1} \tag{11}$$

où $\mathbf{M}_{\mathbf{X}k}$ est la k^e ligne de la matrice $\mathbf{M}_{\mathbf{X}}$ de dimensions $N \times p$ et $\mathrm{M}_{\mathbf{y}k}$ est le k^e élément du vecteur $\mathbf{M}_{\mathbf{y}}$ de dimension $N \times 1$. $\mathbf{M}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{y}}$ sont tous deux définis par les équations (8) et (9), dont les matrices sont constituées de toutes les espérances conditionnelles estimées sur **z**.
L'établissement de l'équation (11) a pour objectif d'obtenir la solution de **B**, qui est l'estimateur par les moindres carrés du modèle de régression de superpopulation de l'équation (2).

De façon analogue à l'équation normale de population, l'estimation d'après les données d'enquête de $\mathbf{u}(\boldsymbol{\theta})$ est

$$\hat{\mathbf{u}}(\mathbf{B}, \hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z})) \equiv \sum_{k\in\mathbf{s}}(\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k})^{\boldsymbol{T}}(y_k - \hat{\mathrm{M}}_{\mathbf{y}k})w_k - \sum_{\boldsymbol{k}\in\mathbf{s}}(\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}\boldsymbol{k}})^{\boldsymbol{T}}(\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}\boldsymbol{k}})\mathbf{B}w_k,$$

où $\hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z})$ est un vecteur dont les éléments sont les espérances conditionnelles estimées de **y** sur tous les points distincts de **z** et $\hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z})$ est un vecteur de la forme $(\hat{\mathbf{m}}_{\mathbf{x}_1}(\mathbf{z}), \ldots, \hat{\mathbf{m}}_{\mathbf{x}_p}(\mathbf{z}))$, où chaque $\hat{\mathbf{m}}_{\mathbf{x}_j}(\mathbf{z})$ est composé des espérances conditionnelles estimées de $\mathbf{x}_j$ sur tous les points distincts de **z.** Sachant que $\hat{\mathbf{B}}$ est l'estimateur par les moindres carrés d'après les données d'enquête de **B** et que $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{B}}, \hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z}))$, nous avons

$$\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) \equiv \sum_{k\in\mathbf{s}}(\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}k})^{\boldsymbol{T}}(y_k - \hat{\mathrm{M}}_{\mathbf{y}k})w_k - \sum_{\boldsymbol{k}\in\mathbf{s}}(\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}\boldsymbol{k}})^{\boldsymbol{T}}(\mathbf{x}_k - \hat{\mathbf{M}}_{\mathbf{x}\boldsymbol{k}})\hat{\mathbf{B}}w_k = \mathbf{0}_{\boldsymbol{p}\times 1}$$

Par développement en série de Taylor de $\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}})$ à $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, nous obtenons

$$\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) \equiv \mathbf{0}_{p\times 1} \approx \hat{\mathbf{u}}(\boldsymbol{\theta}) + \frac{\partial\hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial\mathbf{B}}(\hat{\mathbf{B}} - \mathbf{B}) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})) \tag{12}$$

où $\hat{\mathbf{m}}_{\xi}(\mathbf{z})$ et $\mathbf{m}_{\xi}(\mathbf{z})$ sont deux vecteurs de forme respective $(\mathbf{m}_{\mathbf{y}}(\mathbf{z}), \mathbf{m}_{\mathbf{X}}(\mathbf{z}))$ et $(\hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{X}}(\mathbf{z}))$. $\hat{\mathbf{U}}_{\xi}(\boldsymbol{\theta})$ est une matrice de dimensions $p \times (p+1)m$ dont les composantes sont les dérivées premières de $\hat{\mathbf{u}}(\boldsymbol{\theta})$ par rapport à $m_y(z_i)$ et $m_{x_j}(z_i)$ pour tout $j = 1, \ldots, p$ et $i = 1, \ldots, m$. Notons que, dans le modèle d'intérêt, ξ représente $\mathbf{y}$ ou une covariable $\mathbf{x}_j$. En réarrangeant l'équation (12), nous avons

$$\hat{\mathbf{u}}(\boldsymbol{\theta}) + \hat{\mathbf{U}}_{\xi}(\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})) \approx -\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}(\hat{\mathbf{B}} - \mathbf{B}) .$$

En prenant les variances des deux membres, nous obtenons à la limite,

$$\Omega = \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right) V(\hat{\mathbf{B}}) \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right)^{T}$$

où

$$\Omega = \mathrm{V}(\hat{\mathrm{u}}(\boldsymbol{\theta})) + \mathbf{U}_{\xi}\, \mathrm{V}(\hat{\mathbf{m}}_{\xi}(\mathbf{z}))\mathbf{U}_{\xi}^{T} + 2\,\mathrm{COV}(\hat{\mathrm{u}}(\boldsymbol{\theta}), \hat{\mathbf{m}}_{\xi}(\mathbf{z}))\mathbf{U}_{\xi}^{T} .$$

À condition que la matrice $\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}$ soit de plein rang, nous obtenons l'expression de la variance de l'estimateur d'échantillon des coefficients linéaires qui suit

$$V(\hat{\mathbf{B}}) = \left(\frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right)^{-1} \Omega \left(\frac{\partial \mathbf{u}(\boldsymbol{\theta})^{T}}{\partial \mathbf{B}}\right)^{-1} . \tag{13}$$

Intuitivement, l'équation (13) donne à penser que la variabilité de $\hat{\mathbf{B}}$ est causée par les espérances conditionnelles estimées, le total d'enquête tiré de l'équation d'estimation et la covariance entre le total d'enquête estimé et les espérances conditionnelles estimées. Sachant que

$$\hat{\Omega} = \hat{\mathrm{V}}(\hat{\mathrm{u}}(\boldsymbol{\theta})) + \hat{\mathbf{U}}_{\xi}\, \hat{\mathrm{V}}(\hat{\mathbf{m}}_{\xi}(\mathbf{z}))\hat{\mathbf{U}}_{\xi}^{T} + 2\,\hat{\mathrm{COV}}(\hat{\mathrm{u}}(\boldsymbol{\theta}), \hat{\mathbf{m}}_{\xi}(\mathbf{z}))\hat{\mathbf{U}}_{\xi}^{T} ,$$

un estimateur de $\mathrm{V}(\hat{\mathbf{B}})$ est

$$\hat{\mathrm{V}}(\hat{\mathbf{B}}) = \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})}{\partial \mathbf{B}}\right)^{-1} \hat{\Omega} \left(\frac{\partial \hat{\mathbf{u}}(\boldsymbol{\theta})^{T}}{\partial \mathbf{B}}\right)^{-1} .$$

Notons que toutes les estimations des dérivées dans l'équation (12) sont de la forme d'un total d'enquête et de totaux de domaine. En outre, les estimations d'après les données d'enquête des espérances conditionnelles sont dépendantes des moyennes de domaine et des proportions de domaine. Par conséquent, **B** défini dans l'équation (12) est simplement une fonction des moyennes d'enquête et des moyennes de domaine. Donc, dans le cadre de la superpopulation 2 et sous certaines conditions de régularité énoncées dans la littérature (par exemple, Shao (1998) et Krewski & Rao (1981)), nous énonçons les propriétés asymptotiques de $\hat{\mathbf{B}}$ qui suivent.

Théorème 1 :
Sous les conditions de régularité et un plan d'échantillonnage à plusieurs degrés, la distribution asymptotique de $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B})$ est normale de moyenne zéro et de variance $\mathrm{V}(\hat{\mathbf{B}})$, et $\hat{\mathrm{V}}(\hat{\mathbf{B}})$ converge en probabilité vers $\mathrm{V}(\hat{\mathbf{B}})$.

Si nous définissons $\hat{\mathbf{A}}_i = \mathbf{e}^T(\mathbf{Z}_{z_i}^T \hat{\mathbf{K}}\mathbf{W}_{z_i} \mathbf{Z}_{z_i})^{-1}\mathbf{Z}_{z_i}^T \hat{\mathbf{K}}\mathbf{W}_{z_i}$, l'équation (10) peut être réécrite sous la forme $\hat{g}(z_i) = \hat{\mathbf{A}}_i\hat{\bar{\mathbf{R}}}$. Naturellement, $\hat{\mathbf{A}}_i$ et $\hat{\bar{\mathbf{R}}}$ sont les estimateurs d'après les données d'enquête de $\mathbf{A_i} \equiv \mathbf{e}^T(\mathbf{Z}_{z_i}^T\mathbf{K}\mathbf{W}_{z_i}\mathbf{Z}_{z_i})^{-1}\mathbf{Z}_{z_i}^T\mathbf{K}\mathbf{W}_{z_i}$ et $\bar{\mathbf{R}}$, où $\bar{\mathbf{R}}$ est le vecteur des moyennes de population calculées par classe de $\mathbf{R} = \mathbf{Y} - \mathbf{XB}$. Par le développement en série de Taylor de $\hat{g}(z_i)$ à $\mathbf{A}_i$ et $\bar{\mathbf{R}}$, nous obtenons

$$\hat{g}(z_i) = g(z_i) + (\hat{\mathbf{A}}_i - \mathbf{A}_i)\bar{\mathbf{R}} + \mathbf{A}_i(\hat{\bar{\mathbf{R}}} - \bar{\mathbf{R}}) .$$

En définissant $\hat{\mathbf{C}} = \mathbf{K}\hat{\mathbf{W}} - \mathbf{K}\mathbf{W}$ et en développant l'inverse de la matrice dans $\hat{\mathbf{A}}_i$ par le développement de l'inverse de deux matrices, nous pouvons montrer que $\hat{\mathbf{A}}_i$ est asymptotiquement sans biais pour $\mathbf{A}_i$. Puisque $\hat{\bar{\mathbf{R}}}$ est un vecteur de moyennes d'échantillon calculées par classe, il s'agit d'un estimateur asymptotiquement sans biais de $\bar{\mathbf{R}}$. Donc, l'espérance asymptotique de $\hat{g}(z_i)$ est $g(z_i)$ et la variance fondée sur le plan de sondage de $\hat{g}(z_i)$ est

$$V(\hat{g}(z_i)) = \mathbf{A}_i\, V(\hat{\bar{\mathbf{R}}})\mathbf{A}_i^T \tag{14}$$

où $V(\hat{\bar{\mathbf{R}}}) \approx (\mathbf{Q} \otimes \mathbf{I_m})\mathrm{COV}(\hat{\bar{\mathbf{X}}}, \hat{\bar{\mathbf{y}}})(\mathbf{Q} \otimes \mathbf{I_m})^T + \bar{\mathbf{X}}\, V(\hat{\mathbf{B}})\bar{\mathbf{X}}^{\mathbf{T}}$, sachant que $\bar{\mathbf{X}}$ est la matrice de dimensions $m \times p$ de la forme ($\bar{\mathbf{x}}_1^T$, …, $\bar{\mathbf{x}}_p^T$) et que $\hat{\bar{\mathbf{X}}}$ est la matrice de dimensions $m \times p$ de la forme ($\hat{\bar{\mathbf{x}}}_1^T$, …, $\hat{\bar{\mathbf{x}}}_p^T$) et **Q=** (1, -$B_1$, …, - B_p). En remplaçant toutes les variances et tous les paramètres de population de l'équation (14), nous obtenons l'estimateur de la variance de $V(\hat{g}(z_i))$ sous la forme

$$V(\hat{g}(z_i)) = \hat{\mathbf{A}}_i\, \hat{V}(\hat{\bar{\mathbf{R}}})\hat{\mathbf{A}}_i^T$$

où $\hat{V}(\hat{\bar{\mathbf{R}}}) \approx (\hat{\mathbf{Q}} \otimes \mathbf{I_m})\hat{\mathrm{COV}}(\hat{\bar{\mathbf{X}}}, \hat{\bar{\mathbf{y}}})(\hat{\mathbf{Q}} \otimes \mathbf{I_m})^T + \hat{\bar{\mathbf{X}}}\, \hat{V}(\hat{\mathbf{B}})\hat{\bar{\mathbf{X}}}^{\mathbf{T}}$.

En outre, nous établissons la normalité asymptotique de $\hat{g}(z_i)$, autrement dit,

Théorème 2 :

$\sqrt{n}(\hat{g}(z_i) - g(z_i))$ converge vers une loi normale de moyenne 0 et de variance $\mathbf{V}(\hat{g}(z_i))$ et $\mathbf{V}(\hat{g}(z_i))$ converge vers $\mathbf{V}(\hat{g}(z_i))$ en probabilité.

4. ANALYSE DES DONNÉES

Dans cette analyse, nous illustrons le modèle de régression semiparamétrique partiellement linéaire au moyen de données provenant de l'Enquête sur la santé en Ontario (ESO), qui a été réalisée selon un plan de sondage stratifié par grappes à deux degrés. Les strates étaient les bureaux de santé publics de la province de l'Ontario et, dans chacune des strates, des quartiers ont été sélectionnés aléatoirement, comme l'ont été les ménages dans chaque quartier. L'objectif de cette enquête est de mesurer l'état de santé des résidents de l'Ontario et de recueillir des données sur les facteurs de risque associés aux causes principales de mortalité en Ontario.

Afin d'illustrer le modèle partiellement linéaire, 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 désirée (IMCD). L'IMC est une mesure représentative du poids réel et l'IMCD est une mesure représentative du poids souhaité. L'IMC et l'IMCD se calculent comme suit

$$\text{IMC} = \frac{\text{poids en kilos}}{(\text{taille en mètre})^2}$$

$$\text{IMCD} = \frac{\text{poids souhaité en kilos}}{(\text{taille en mètre})^2}$$

Nous utilisons l'âge comme variable continue et traitons les autres facteurs comme des variables discrètes. Le modèle de travail est le suivant :

$$\text{IMC} = g_1(\hat{a}ge) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1$$
$$\text{IMCD} = g_2(\hat{a}ge) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2$$

où **X** est la matrice du plan d'expérience comprenant toutes les variables indicatrices provenant des facteurs.

Parmi l'ensemble de variables explicatives, nous nous concentrons sur la variable continue d'âge. Puisque l'IMC ne s'applique pas aux adolescents, nous retenons uniquement les répondants de 18 à 64 ans. Après avoir éliminé toutes les observations pour lesquelles des valeurs manquaient ou pour lesquelles la réponse était « non précisé », le fichier de données contient en tout 21 968 observations. Puisque la variable d'âge ne compte que 46 points distincts, nous regroupons l'ensemble de données par classe en fonction de l'âge. Nous fixons la taille des classes à l'unité, de sorte que nous obtenons 46 classes dont les points milieu sont 18, 19,…, 64.

Le **tableau 1** donne les estimations d'après les données d'enquête des coefficients linéaires du premier modèle. La comparaison de l'IMC selon le sexe nous porte à conclure que l'IMC des hommes est plus élevé que celui des femmes. En prenant comme catégorie de référence les anciens fumeurs, nous obtenons des coefficients négatifs et significatifs pour toutes les catégories d'usage du tabac, ce qui donne à penser que les anciens fumeurs ont tendance à peser plus que les personnes appartenant à d'autres catégories d'usage du tabac.

Tableau 1 : Estimations des coefficients linéaires

Facteurs	$\hat{\mathbf{B}}$	**E.-T.**$(\hat{\mathbf{B}})$	**Valeur *t***
Sexe	1,45	0,052	27,90
N'a jamais fumé	-1,45	0,065	-22,27
Fume à l'occasion	-1,72	0,12	-14,41
Fume tous les jours	-1,48	0,072	-20,63
Moyennement actif(ve)	0,66	0,095	6,96
Inactif(ve)	1,43	0,078	18,45

Aux **figures 1** et **2**, les fonctions d'âge estimées, $\hat{g}_1(\hat{a}ge)$ et $\hat{g}_2(\hat{a}ge)$, et leurs intervalles de confiance sont représentés graphiquement en fonction de l'âge. Nous constatons que, dans les deux cas, l'IMC et l'IMCD sont des fonctions non linéaires croissantes de l'âge. La **figure 3** présente une comparaison de $\hat{g}_1(\hat{a}ge)$ à $\hat{g}_2(\hat{a}ge)$. Nous constatons qu'en moyenne, pour chaque individu actif ou moyennement actif, l'IMCD est plus faible que l'IMC.

Figure 1 : Tendance estimée de l'IMC selon l'âge avec intervalles de confiance

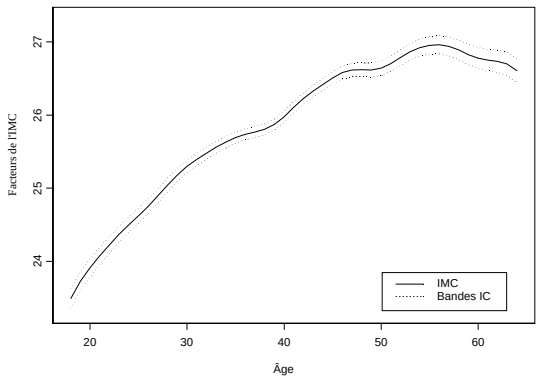


Figure 2: Tendance estimée de l'IMCD selon l'âge avec intervalles de confiance

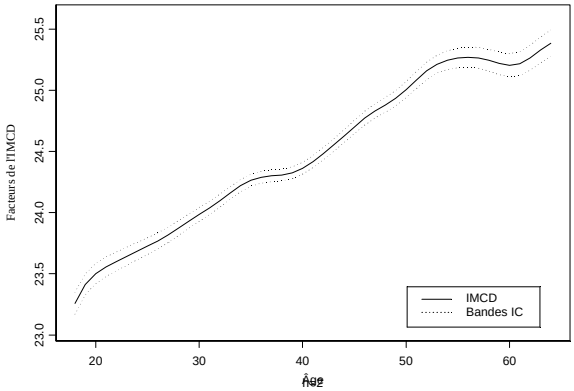
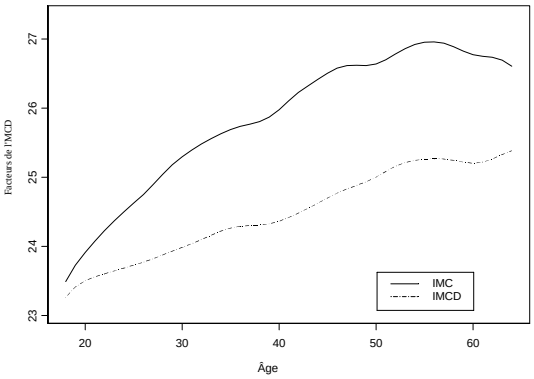


Figure 3: Tendance estimée de l'IMC et de l'IMCD selon l'âge



5. CONCLUSION

À l'aide d'un modèle partiellement linéaire, nous étendons les techniques de régression semiparamétrique aux données d'enquête complexes. Nous établissons les propriétés asymptotiques des estimateurs fondés sur les données d'enquête. Le calcul des estimations de la variance des coefficients linéaires ainsi que de la fonction de régression s'appuie sur la variance des totaux et des moyennes d'enquête. À condition d'obtenir les estimations de variance requises, nous pouvons appliquer cette méthode en nous servant de progiciels statistiques standards. Dans le cas du modèle de travail partiellement linéaire, nous supposons qu'il n'existe aucune interaction entre les composantes paramétrique et non paramétrique. Cette hypothèse peut être atténuée de façon telle que la composante non paramétrique apparaisse linéairement dans le terme d'interaction. Pour estimer les espérances conditionnelles des composantes non paramétriques pour des variables indicatrices aléatoires discrètes, nous proposons d'utiliser des modèles linéaires ou additifs généralisés pour procéder à l'estimation.

RÉFÉRENCES

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

Bellhouse D. R. et Stafford, J. E. (2001), Local polynomial regression in complex survey. *Survey methodology*, 27(2), pp. 197 - 203.

Binder, D. A.(1983), On the variance of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, pp. 279-292.

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

Konijn, H. S. (1962), Regression analysis in the sample surveys, *Journal of the American Statistical Association*, 57, pp. 590-606.

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

Ministère ontarien de la santé (1996), *Ontario Health Survey: User's Guide, Volumes I and II*, Queen's Printer for Ontario.

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

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

Speckman, P. (1988), Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society, Series B (Methodological)*, 50(3), pp. 413-436.