No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Actes du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Inférence bayésienne prédictive robuste pour les quantiles de population finie d'un petit domaine

Balgobin Nandram et Jai Won Choi[1]

Résumé

Nous suivons une méthode bayésienne robuste pour analyser des données pouvant présenter un biais de non-réponse et un biais de sélection *non ignorables*. Nous utilisons un modèle de régression logistique *robuste* pour établir le lien entre les indicateurs de réponse (variable aléatoire de Bernoulli) et les covariables, dont nous disposons pour tous les membres de la population finie. Ce lien permet d'expliquer l'écart entre les répondants et les non-répondants de l'échantillon. Nous obtenons ce modèle robuste en élargissant le modèle de régression logistique conventionnel à un mélange de lois t de Student, ce qui nous fournit des scores de propension (probabilité de sélection) que nous utilisons pour construire des *cellules d'ajustement*. Nous introduisons les valeurs des non-répondants en tirant un échantillon aléatoire à partir d'un estimateur à noyau de la densité, formé d'après les valeurs des répondants à l'intérieur des cellules d'ajustement. La prédiction fait appel à une régression linéaire spline, fondée sur les rangs, de la variable de réponse sur les covariables selon le domaine, en échantillonnant les erreurs à partir d'un autre estimateur à noyau de la densité, ce qui rend notre méthode encore plus robuste. Nous utilisons des méthodes de Monte-Carlo par chaînes de Markov (MCMC) pour ajuster notre modèle. Dans chaque sous-domaine, nous obtenons la loi a posteriori d'un quantile de la variable de réponse à l'intérieur de chaque sous-domaine en utilisant les statistiques d'ordre sur l'ensemble des individus (échantillonnés et non échantillonnés). Nous comparons notre méthode robuste à des méthodes paramétriques proposées récemment.

MOTS-CLÉS : Régression logistique; échantillonneur de Metropolis-Hastings; statistiques d'ordre; scores de propension; méthode fondée sur les rangs; loi t de Student.

1. Introduction

L'objet de l'étude est de prédire le centile de l'indice de masse corporelle (IMC) pour la population finie des enfants et des adolescents, stratifiée a posteriori selon le comté pour chaque domaine formé par l'âge, la race et le sexe, et de déterminer quels ajustements sont nécessaires pour tenir compte du biais de non-réponse et du biais de sélection, à l'aide des données de la troisième National Health and Nutrition Examination Survey (NHANES III). Nandram et Choi (2005, 2006) ajustent des modèles hiérarchiques bayésiens de façon à montrer le mécanisme de non-réponse. Nous cherchons des modèles plus robustes en présence d'hypothèses quant à la distribution et de valeurs aberrantes.

Greenlees, Reece et Zieschang (1982) ont élaboré un modèle de régression logistique-normale, soit un modèle de non-réponse non ignorable dans le cadre de la méthode de sélection, pour imputer les valeurs manquantes de la Current Population Survey lorsque la probabilité de réponse dépend de la variable imputée. Nandram et Choi (2005) étendent ce modèle pour prendre en compte de petits domaines dans les données de la NHANES III. Le principal apport de Nandram et Choi (2005) est une inférence prédictive bayésienne de la moyenne de la population finie utilisant un modèle de régression spline dans lequel on modélise le logarithme des valeurs de l'IMC. Nandram et Choi (2006) apportent quatre nouveaux éléments. Premièrement, ils font une inférence au sujet des centiles de la population finie, qui sont plus appropriés. Deuxièmement, ils montrent que la transformation logarithmique est la meilleure dans un ensemble sélectionné à l'intérieur de la famille de Box-Cox. Troisièmement, ils montrent l'incidence minime de la mise en grappes. Quatrièmement, ils montrent comment tenir compte des probabilités de sélection.

[1]Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280; Jai Won Choi, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. Les opinions exprimées dans le présent document sont celles des auteurs et ne reflètent pas nécessairement le point de vue du National Center for Health Statistics.

Dans le présent document, nous cherchons surtout à rendre plus robuste le modèle de régression logistique fondé sur la relation entre la loi logistique et la loi t de Student. Les quatre premiers moments d'une loi logistique standard (pour une position nulle et une échelle unitaire) sont les mêmes que ceux d'une loi t de Student avec position nulle, échelle $\gamma = \pi\sqrt{7/27}$ et neuf degrés de liberté. Effectivement, si x/γ présente une loi t de Student standard, alors x présente à peu près une loi logistique standard; on trouve l'exposé de ce principe dans Mudholkar et George (1978) et une application dans Albert et Chib (1993). Ce résultat ouvre la voie à notre analyse robuste de la non-réponse. Soit $I_i = 1$ si la i^e personne répond et $I_i = 0$ dans le cas contraire, pour un échantillon de taille n, et soit $\underline{z}_i$ qui représente des covariables. Puis, le modèle de régression logistique standard avec $I_i \mid \underline{\beta} \overset{ind}{:} \; Bernoulli\{e^{\underline{z}_i'\underline{\beta}}/(1+e^{\underline{z}_i'\underline{\beta}})\}$, est à peu près le même que $I_i \mid \underline{\beta} \overset{ind}{:} \; Bernoulli\{\mathrm{T}_\eta(\underline{z}_i'\underline{\beta}/\gamma)\}$, où $\mathrm{T}_\eta(\cdot)$ est la fonction de distribution cumulative de la variable aléatoire t de Student sur $\eta = 9$ degrés de liberté (ainsi, les coefficients de régression $\underline{\beta}$ dans les deux modèles sont à peu près les mêmes). On obtient donc la « robustification » en assignant des poids à différentes valeurs de η, dont un poids substantiel à $\eta = 9$ degrés de liberté, ce qui permet former une catégorie plus souple de modèles.

En outre, nous rendons notre méthode plus robuste à l'égard de la prédiction. Au lieu de supposer la normalité, nous utilisons une méthode linéaire fondée sur les rangs (Hettmansperger, 1984) pour ajuster un modèle linéaire du logarithme de l'IMC sur les covariables, ce qui ne nécessite pas d'hypothèses quant à la distribution; voir à ce sujet les observations intéressantes de Potvin et Roff (1993) au sujet de la normalité. Cette nouvelle méthode constitue manifestement un progrès important par rapport au modèle logistique-normal de Greenlees, Reece et Zieschang (1982) et de Nandram et Choi (2005, 2006).

La grande différence avec nos travaux antérieurs est que nous utilisons des scores de propension pour étudier la non-réponse « non ignorable ». Ici, notre démarche diffère même de notre méthode non paramétrique antérieure fondée sur la loi a priori de Dirichlet (voir Nandram et Choi, 2004). Les scores de propension intègrent les écarts entre les répondants et les non-répondants par le biais de covariables; Little et Rubin (2002) présentent un exposé sur la non-réponse et Rosenbaum (2002), un exposé semblable sur les études d'observation. Le score de propension décrit simplement la probabilité de réponse d'une personne en fonction des covariables. Selon une hypothèse raisonnable courante, les indicateurs de réponse suivent un modèle de régression logistique dans lequel la probabilité de réponse (score de propension) est une fonction des covariables. Nous utilisons les scores de propension pour former des cellules d'ajustement et nous introduisons les valeurs des non-répondants en tirant des échantillons à partir d'un estimateur à noyau de la densité (Silverman, 1986).

La notion de non ignorabilité diffère quelque peu de celle de Little et Rubin (2002). Les scores de propension estimatifs sont une fonction des indicateurs de réponse et les covariables ne sont pas liées aux réponses observées. Par contre, les réponses non observées sont une fonction des scores de propension estimatifs et des covariables; elles sont donc une fonction des indicateurs de réponse. Toutefois, les covariables ne peuvent pas discriminer entre les répondants et les non-répondants. Nos modèles sont donc très souples puisqu'ils peuvent refléter un certain degré de non ignorabilité.

Le présent document a pour objet de décrire une méthode robuste servant à obtenir des scores de propension afin d'introduire les valeurs des non-répondants et de prédire les centiles de la population finie (à risque d'embonpoint et avec embonpoint) pour de petits domaines formés par l'âge, la race et le sexe. Dans la section 2, nous présentons les données de la NHANES III que nous étudions. Dans la section 3, nous passons en revue le modèle hiérarchique bayésien antérieur (Nandram et Choi, 2006) pour la non-réponse non ignorable par le biais de la méthode de sélection. Dans la section 4, nous proposons notre méthode bayésienne robuste. Dans la section 5, nous présentons l'analyse des données.

2. Principales caractéristiques des données de la NHANES III

La non-réponse peut se produire dans les volets interview et examen physique de la NHANES III, menée d'octobre 1988 à septembre 1994 (pour plus de détails, voir National Center for Health Statistics, 1994). La non-réponse à l'interview se produit lorsque les personnes échantillonnées ne participent pas à l'interview. Certaines personnes,

déjà interviewées et qui participent à l'évaluation de la santé, n'ont pas subi l'examen physique à la maison ou au centre d'examen mobile, et ont donc manqué la totalité ou une partie des examens physiques. Dans tous nos travaux antérieurs, on entend par « non-réponse » une valeur de l'IMC manquante pour les personnes échantillonnées dont on dispose de renseignements sur l'âge, le sexe et la race. Nous notons aussi que, pour les enfants et les adolescents (de 2 à 19 ans), le taux observé de non-réponse est de $100 \times (1606/6791) \approx 24$ %.

Nous étudions les données sur l'IMC dans quatre groupes d'âge (2 à 4 ans, 5 à 9 ans, 10 à 14 ans et 15 à 19 ans). Si l'on se souvient qu'il existe 560 ($35 \times 2 \times 2 \times 4$) domaines, la taille d'échantillon par domaine est, en moyenne, très faible, par ex., $6791/560 \approx 12$ pour l'échantillon et $5185/560 = 9$ pour les répondants. Bon nombre de domaines par comté sont trop petits (ils ne comptent pas suffisamment de personnes comprises dans l'échantillon) pour permettre d'effectuer une analyse significative. Nous formons donc les petits domaines en recoupant l'âge, la race et le sexe dans les comtés. Comme dans Nandram et Choi (2005, 2006), nous construisons nos modèles au niveau du comté et nous représentons l'âge, la race et le sexe comme des covariables. Nous procédons à l'inférence pour chaque domaine formé par le recoupement de l'âge, de la race et du sexe dans le comté. Il existe des probabilités de sélection différentielles selon l'âge, la race et le sexe, d'après une méthode de sélection, d'où le suréchantillonnage d'un groupe d'âge et d'une race.

3. Résumé du modèle bayésien de régression spline

Nous présentons un résumé du modèle de régression spline de Nandram et Choi (2005) et décrivons comment Nandram et Choi (2006) intègrent au modèle la corrélation et les probabilités de sélection.

Nous disposons de données provenant de $\ell = 35$ comtés et chaque comté comprend N_i personnes (connues). Nous supposons qu'un échantillon probabiliste de n_i personnes est tiré dans le i^e comté. Soit s l'ensemble d'unités échantillonnées et ns l'ensemble d'unités non échantillonnées. Soit I_{ij}, $i = 1,2,..., \ell$ et $j = 1,2,..., N_i$ l'indicateur de réponse pour la j^e personne dans le i^e comté dans la population. En outre, soit x_{ij} la valeur de l'IMC, qui peut être transformée (p. ex., la transformation logarithmique). Il convient de souligner que les valeurs de r_{ij} et x_{ij} sont toutes observées dans l'échantillon s ; celles de x_{ij} sont inconnues et celles de r_{ij} sont inutiles dans ns. Soit $r_i = \sum_{j=1}^{n_i} r_{ij}$ (autrement dit, r_i est le nombre de personnes échantillonnées qui ont répondu dans le i^e comté). Par souci de commodité, nous exprimons le logarithme de l'IMC x_{ij} sous la forme $x_{i1}, x_{i2}, \ldots, x_{ir_i}, x_{ir_i+1}, \ldots, x_{in_i}$ dans s, $x_{in_i+1}, \ldots, x_{iN_i}$ dans ns et $(N_i - n_i)$ personnes non échantillonnées pour le i^e comté.

Un point important que nous tenons à souligner pour la suite est que les r_i personnes ne sont pas nécessairement des répondants aléatoires provenant des n_i répondants choisis. Il s'agit là du biais de non-réponse dont nous devons tenir compte. Il est évident que nous devons prédire les valeurs de l'IMC, x_{ij}, pour a) les non-répondants dans s et b) les personnes dans ns. Donc, pour la population finie de N_i personnes, nous avons besoin d'une inférence prédictive bayésienne pour le $100\eta, 0 < \eta < 1$, centile de la population finie des valeurs de l'IMC pour chaque domaine âge-race-sexe dans le i^e comté. Par exemple, pour le i^e comté, soit $\underline{x}_i = (\underline{x}_i^{(s,r)}, \underline{x}_i^{(s,nr)}, \underline{x}_i^{(ns)})'$, où $\underline{x}_i^{(s,r)}$ représente les valeurs observées de l'IMC des répondants échantillonnés, alors que $\underline{x}_i^{(s,nr)}$ (les valeurs de l'IMC des non-répondants échantillonnés) et $\underline{x}_i^{(ns)}$ (celles des personnes non échantillonnées) ne sont pas observées. Puis, le 100η centile du i^e comté est la statistique d'ordre $[\eta N_i]^e$ ($[\cdot]$ est le nombre entier le plus proche de ηN_i) parmi

les N_i composantes de $\underline{x}_i$. Comme seul $\underline{x}_i^{(s,r)}$ est observé, Nandram et Choi (2005) élaborent un modèle bayésien de sélection et un modèle bayésien de mélange de schémas d'observation pour prédire l'IMC moyen de la population finie pour chaque domaine. Il convient de souligner que seuls les $\underline{x}_i^{(s,r)}$ sont observés; $\underline{x}_i^{(s,nr)}$ et $\underline{x}_i^{(ns)}$ sont à prédire. La tâche est de taille pour trois raisons : la population finie est considérable, il existe des covariables et l'on effectue une transformation.

Pour tenir compte des valeurs de l'IMC et des indicateurs de non-réponse, Nandram et Choi (2006) utilisent le modèle de sélection de non-réponse, qui comprend deux parties. Pour la partie 1 du modèle de sélection, la réponse dépend de l'IMC comme suit : $I_{ij} \mid x_{ij}, \underline{\beta}_i \overset{ind}{:} Bernoulli\left\{e^{\beta_{0i}+\beta_{1i}x_{ij}}/(1+e^{\beta_{0i}+\beta_{1i}x_{ij}})\right\}$, où $\underline{\beta}$ est une loi normale bivariée avec des lois a priori appropriées pour la moyenne, la variance et la corrélation. Pour la partie 2 du modèle de sélection, le prédicteur de loin le plus important de l'IMC est l'âge, le rôle de la race et du sexe étant relativement mineur. En outre, il est nécessaire de comprendre la relation entre l'IMC et l'âge, la race et le sexe. Pour $i = 1,\ldots,\ell,\ j = 1,\ldots,N_i$, soit $z_{ij0} = 1$ pour une coordonnée à l'origine, $z_{ij1} = 1$ pour non-noir et $z_{ij1} = 0$ pour noir, $z_{ij2} = 1$ pour masculin et $z_{ij2} = 0$ pour féminin, $z_{ij3} = z_{ij1} \times z_{ij2}$ pour l'interaction entre la race et le sexe, et soit $\underline{z}_{ij}^{'} = \left(z_{ij0}, z_{ij1}, z_{ij2}, z_{ij3}\right)$. En outre, soit a_{ij} l'âge de la j^e personne dans le i^e comté. De façon générique, en posant que $c^+ = 0$ si $c \le 0$ et $c^+ = c$ si $c > 0$, $w_{ij1} = 1, w_{ij2} = (a_{ij} - 8)^+, w_{ij3} = (a_{ij} - 13)^+$, pour une régression spline de l'IMC sur l'âge, en corrigeant pour la race et le sexe, nous prenons $x_{ij} = \sum_{t=1}^{3}(\underline{z}_{ij}^{'}\underline{\alpha}_t + \nu_{ti})w_{ijt} + e_{ij},\ e_{ij} \mid \sigma_3^2 \overset{iid}{:} Normal(0, \sigma_3^2)$ avec des lois a priori et des hyper-lois a priori pertinentes. Voir l'annexe B de Nandram et Choi (2005) pour connaître les détails de l'ajustement du modèle de sélection à l'aide des méthodes de Monte-Carlo par chaînes de Markov.

La prédiction des centiles de la population finie est simple. Nous prédisons d'abord $\underline{x}^{(ns)} = (x_{ij},\ j = n_i + 1,\ldots,N_i)$. Puis, $f(\underline{x}^{(ns)} \mid \underline{x}^{(obs)}) = \int\{\prod_{j=n_i+1}^{N_i} f(x_{ij} \mid \Omega)\}\,\pi(\Omega \mid \underline{x}^{(obs)})d\Omega$, où $x_{ij} \mid \Omega:\ Normal\{\sum_{t=1}^{3}(\underline{z}_{ij}^{'}\underline{\alpha}_t + \nu_{ti})w_{ijt}, \sigma_3^2\}$ et Ω est l'ensemble de tous les paramètres comprenant $\underline{x}_i^{(s,nr)}$. Donc, il est simple de prédire $x_{in_i+1},\ldots,x_{iN_i}$. Premièrement, nous prenons un échantillon de taille M dans la loi a posteriori, $\{\Omega^{(h)} : h = 1,\ldots,M\}$; deuxièmement, nous introduisons $x_{i,n_i+1},\ldots,x_{iN_i}$. (Après avoir prédit les valeurs de l'IMC transformées, on peut facilement les retransformer à l'échelle initiale.) En posant $\underline{x}_i^{(h)}$ comme vecteur de toutes les N_i valeurs itérées, nous ordonnons ces composantes pour obtenir la $[\eta N_i]^e$ valeur $x_{[\eta N_i]}^{(h)}$. Nous obtenons donc un « échantillon aléatoire » $x_{[\eta N_i]}^{(h)}, h = 1,\ldots,M$, d'après la loi a posteriori du $[100\eta]$ centile.

4. Méthode bayésienne robuste

Nous allons maintenant décrire la méthode robuste servant à prédire les centiles de la population finie. Premièrement, nous utilisons un modèle robuste de régression logistique pour obtenir des scores de propension. Deuxièmement, en utilisant ces scores de propension pour former des cellules d'ajustement, nous introduisons les valeurs des non-répondants. Troisièmement, nous ajustons un modèle de régression linéaire spline fondé sur les rangs des valeurs logarithmiques de l'IMC (répondants et non-répondants) en fonction des covariables pertinentes et nous utilisons ce modèle pour prédire les valeurs de l'IMC des personnes non échantillonnées. Nous incluons également les probabilités de sélection. Nous appellerons le modèle décrit dans la présente section le *modèle de mélange* ou *modèle de régression logistique robustifié*. Il convient de souligner que le modèle de régression

logistique à effets aléatoires constitue un cas particulier de ce modèle lorsqu'il y a neuf degrés de liberté.

4.1 Régression logistique robuste

Nous supposons que

$$I_{ij} \mid \{\underline{\beta}, \nu_i, \eta = a_r\} \overset{ind}{:} Bernoulli[\mathrm{T}_{a_r}\{(\underline{z}_{ij}'\underline{\beta} + \nu_i)/\gamma\}], j = 1,\ldots,n_i, \tag{1}$$

où $\mathrm{T}_{a_r}(\cdot)$ est la fonction de distribution cumulative de la loi t de Student sur a_r degrés de liberté,

$$\nu_i \mid \sigma^2 \overset{iid}{:} Normal(0, \sigma^2), i = 1,\ldots,\ell, \tag{2}$$

$$Pr(\eta = a_r) = \omega_r, r = 1,\ldots,R, \sum_{r=1}^{R}\omega_r = 1, \tag{3}$$

où $\gamma = \pi\sqrt{7/27}$ et l'on spécifie $\{(a_r, \omega_r), r = 1,\ldots,R\}$ où R est le nombre de valeurs de η. [Il convient de souligner qu'il est pratique d'utiliser R ici; plus loin, nous l'utiliserons pour représenter des rangs; en outre, nous avons utilisé η dans la section 3 pour les centiles.] Dans notre application, nous prenons $a_1 = 3, a_2 = 6, a_3 = 9, a_4 = 12, a_5 = 25, a_6 = 50$. Ici, $a_1 = 3$ est proche d'une loi de Cauchy, $a_3 = 9$ est la loi logistique approximative et $a_6 = 50$ est proche d'une loi normale. Dans (1), (2) et (3), la construction est une prescription de R modèles, le r^e modèle ayant la probabilité ω_r. Il convient de souligner que les ν_i sont des effets de domaine et qu'ils forment un processus stochastique courant. Il convient aussi de souligner que les coefficients de régression ne varient pas d'un modèle à l'autre, ce qui permet l'emprunt d'information entre les comtés; ce phénomène est atténué par les ν_i. En outre, il convient de déterminer un seul ensemble de paramètres de régression comme dans (1); on évite ainsi d'avoir à calculer la moyenne par modèle, opération complexe et peu souhaitable. Nous faisons en sorte que seul le degré de liberté a_r dépende du modèle.

Enfin, nous supposons a priori que $\underline{\beta}$ et σ^2 sont indépendants et que

$$\underline{\beta} \sim Normal(\underline{\theta}_0, \Delta_0) et\ p(\sigma^2) = 1/(1+\sigma^2)^2,\ \sigma^2 > 0, \tag{4}$$

où l'on spécifie $\underline{\theta}_0$ et Δ_0, et nous évitons l'encombrante loi a priori σ^{-2}: $Gamma(.001, .001)$ (voir Gelman, 2006). En effectuant une régression logistique standard des indicateurs de réponse sur les covariables, nous spécifions $\underline{\theta}_0$ et Δ_0 en prenant $\underline{\theta}_0$ comme estimateur ponctuel de $\underline{\beta}$ et Δ_0 comme matrice de covariance grossie cent fois.

Il convient de souligner que les équivalences $\omega_1 = 1\ et\ a_1 = 9$ donnent le modèle de régression logistique et qu'à la limite, lorsque σ^2 devient nul, les ν_i deviennent des masses ponctuelles à 0 (il n'y a donc pas d'effets de comté). Les deux opérations simultanées produisent le modèle de régression logistique sans effets de comté. Nous appellerons le modèle spécifié par les équations (1) à (4) le modèle de régression logistique robustifié, qui est à la base de la prédiction robuste des centiles de la population finie.

Nous ajustons le modèle à l'aide de l'échantillonneur de Metropolis-Hastings (voir l'annexe B de Nandram et Choi, 2006, au sujet du modèle de sélection). Nous obtenons un échantillon $(\underline{\beta}^{(h)}, \eta^{(h)}, \underline{\nu}^{(h)}), h = 1,\ldots,M = 1000$ que nous utilisons pour formuler une inférence au sujet des centiles de la population finie. La loi a posteriori conjointe est $p(\underline{\beta}, \underline{\nu}, \sigma^2 \mid \underline{I})$. Pour exécuter l'échantillonneur de Metropolis-Hastings, nous avons besoin des lois a posteriori conditionnelles pour les lois a priori, $p(\underline{\beta} \mid \underline{\nu}, \sigma^2, \underline{I})$, $p(\underline{\nu}_i \mid \underline{\beta}, \underline{\nu}_{(i)}, \sigma^2, \underline{I})$, et $p(\sigma^2 \mid \underline{\beta}, \underline{\nu}, \underline{I})$ (voir Nandram et Choi, 2006).

Nous voulons prédire les valeurs de l'IMC des non-répondants et les valeurs de l'IMC de la population non échantillonnée. Les méthodes servant à introduire les valeurs manquantes pour les non-répondants et pour la population non échantillonnée sont toutes deux robustes. Prenons d'abord les non-répondants. À la h^e itération, les scores de propension (probabilité de sélection) sont $\mathrm{T}_{\eta^{(h)}}\{(\underline{z}_{ij}\underline{\beta}^{(h)} + \nu_i^{(h)})/\gamma\}$, $j = 1,\ldots,n_i, i = 1,\ldots,\ell$, où $\eta^{(h)}$ est l'un des $a_r, r = 1,\ldots,R$. Pour la h^e itération, nous avons divisé tous les scores de propension (pour les répondants et les non-répondants de tous les comtés) en cinq strates (cellules d'ajustement). Nous avons calculé les cinq quintiles (.2, .4, .6, .8 pour former les cinq strates) des scores de propension, et nous avons réparti les r_i répondants et les $(n_i - r_i)$ non-répondants dans ces cinq strates. Nous savons donc maintenant à quelle strate appartient chaque personne (répondant ou non-répondant), mais avec une part d'incertitude.

Pour chaque strate, nous construisons un estimateur à noyau de la densité des n valeurs observées. Soit L_s le nombre de répondants compris dans la s^e cellule d'ajustement et l'ordre du logarithme des valeurs de l'IMC des répondants compris dans la s^e cellule d'ajustement sous la forme $\tilde{x}_{sj}, j = 1,\ldots,L_s$. Pour estimer la densité des $\tilde{x}_{sj}$, nous utilisons alors l'estimateur à noyau de la densité,

$$g(\tilde{x}) = \sum_{j=1}^{L_s} \frac{1}{L_s}\frac{1}{h_{opt}}\phi\left(\frac{\tilde{x} - \tilde{x}_{sj}}{h_{opt}}\right),$$

où $\phi(\cdot)$ est la loi normale standard et $h_{opt} = \frac{1{,}06}{L_s^{1/5}} min(STD, IQR/1{,}34)$ et STD et IQR sont respectivement l'écart-type et l'intervalle interquartile des $\tilde{x}_{sj}, j = 1,\ldots,L_s$, s=1,...,5 (voir Silverman, 1986, p. 47).

Puis, pour introduire la valeur de l'IMC d'un non-répondant, nous tirons une valeur aléatoire à partir de $g(\tilde{x})$. Plus précisément, nous tirons aléatoirement un échantillon de taille un à partir des étiquettes $j = 1,\ldots,L_s$, par exemple j', puis nous tirons le logarithme de la valeur de l'IMC à partir de la loi normale, la moyenne étant $\tilde{x}_{sj'}$ et l'écart-type étant h_{opt}. Nous obtenons la valeur de l'IMC x_{sj} en retransformant $\tilde{x}_{sj}$ (soit $x_{sj} = exp(\tilde{x}_{sj})$. Nous répétons indépendamment tout le processus pour tous les non-répondants de chacune des cinq cellules d'ajustement. À la h^e itération, nous l'avons fait pour tous les non-répondants de toutes les strates. Il convient de souligner que ce processus comprend aussi une part d'incertitude au sujet de la formation des strates, car les cellules d'ajustement sont formées à chaque itération.

Deuxièmement, nous pouvons appliquer une méthode semblable, mais sans utiliser les cellules d'ajustement, aux valeurs de l'IMC des personnes non échantillonnées, dont nous avons besoin pour obtenir les centiles de la population finie. Plus précisément, nous utilisons une méthode d'estimation linéaire spline fondée sur les rangs. Mentionnons en passant que la méthode des moindres carrés constitue une solution de rechange, moins robuste. Toutefois, le nombre de calculs à effectuer selon cette méthode est moins élevé et nous avons trouvé de légers écarts entre la méthode linéaire spline fondée sur les rangs et celle des moindres carrés; nous préférons donc la méthode linéaire spline fondée sur les rangs.

Nous utilisons une notation semblable à celle de la partie 2 du modèle de sélection pour la régression spline exécutée dans la section 3. Pour $i = 1,\ldots,\ell, j = 1,\ldots,N_i$, soit $z_{ij0} = 1$ pour une coordonnée à l'origine, $z_{ij3} = 1$ pour non-noir et $z_{ij3} = 0$ pour noir, $z_{ij4} = 1$ pour masculin et $z_{ij4} = 0$ pour féminin, $z_{ij5} = z_{ij3} \times z_{ij4}$ pour l'interaction entre la race et le sexe. Ici, $z_{ij1} = (a_{ij} - 8)^+$ et $z_{ij2} = (a_{ij} - 13)^+$ avec, encore une fois, $c^+ = 0$ si $c \le 0$ et $c^+ = c$ si $c > 0$. Il convient de signaler la légère variation de la définition de

$\underline{z}_{ij}^{'} = \left(z_{ij0}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}, z_{ij5},\right)$, soit les deuxième et troisième composantes pour les deux splines non constantes de l'âge.

Nous spécifions le modèle ci-dessous pour toutes les personnes comprises dans la population finie (les répondants et les non-répondants échantillonnés et les personnes non échantillonnées). À la h^e itération, $\widetilde{x}_{ij}$ étant le logarithme des valeurs de l'IMC, nous utilisons le modèle

$$\widetilde{x}_{ij}^{(h)} = \underline{z}_{ij}^{'} \underline{\widetilde{\beta}}_i + e_{ij}, i = 1,\ldots,\ell, j = 1,\ldots,N_i,$$

où e_{ij} sont indépendants et répartis de façon identique. Il convient de souligner que $\widetilde{x}_{ij}^{(h)}, j = 1,\ldots,r_i$ ne varient pas avec h parce qu'ils sont observés et que les $\widetilde{x}_{ij}^{(h)}, j = r_i + 1,\ldots,n_i$ non observés pour les non-répondants de l'échantillon sont obtenus à l'aide de l'estimateur à noyau de la densité comme dans le cas des non-répondants, sans hypothèses précises quant à la distribution. Ainsi, nous utilisons les valeurs de l'IMC des répondants et des non-répondants pour estimer les $\underline{\widetilde{\beta}}_i$. Il convient également de souligner que les paramètres de ce modèle n'ont aucun lien avec ceux du modèle de régression logistique.

Nous obtenons les estimateurs fondés sur les rangs $\underline{\widetilde{\beta}}_i^{(h)}$ (voir l'annexe B de Nandram et Choi, 2006). Puis, à l'aide des estimateurs fondés sur les rangs, nous obtenons les résidus $\widetilde{e}_{ij}^{(h)} = \widetilde{x}_{ij}^{(h)} - \underline{z}_{ij}^{'} \underline{\widetilde{\beta}}_i^{(h)}, i = 1,\ldots,\ell, j = 1,\ldots,n_i$. Comme dans le cas des non-répondants, dans le i^e comté et à la h^e itération, nous avons construit un estimateur à noyau de la densité à l'aide des $\widetilde{e}_{ij}^{(h)}, j = 1,\ldots,n_i$. Pour introduire les valeurs de l'IMC de la population non échantillonnée, nous tirons donc un échantillon aléatoire de taille $N_i - n_i$ à l'aide de l'estimateur à noyau de la densité, formé par les $\widetilde{e}_{ij}^{(h)}$ du i^e comté, $i = 1,\ldots,\ell$. Puis, $x_{ij}^{(h)} = \exp\{\widetilde{x}_{ij}^{(h)}\}$, où nous obtenons $\widetilde{x}_{ij}^{(h)} = \widetilde{e}_{ij}^{(h)} + \underline{z}_{ij}^{'} \underline{\widetilde{\beta}}_i^{(h)}, i = 1,\ldots,\ell, j = n_i + 1,\ldots,N_i$.

Ainsi, compte tenu des valeurs observées de l'IMC, nous avons obtenu les valeurs de l'IMC des non-répondants et des personnes non échantillonnées. Par conséquent, il est maintenant aisé d'obtenir les quantiles de la population finie comme dans le modèle de régression spline (voir la section 3).

Enfin, nous avons aussi inclus les probabilités de sélection dans ce cadre robuste. Nous reconnaissons que l'inclusion des probabilités de sélection est pose des problèmes, car il n'existe pas de relation linéaire simple entre les probabilités de sélection et les valeurs de l'IMC, mais une étude préliminaire révèle que ces probabilités de sélection renferment des renseignements importants au sujet de l'IMC. Voir l'annexe B de Nandram et Choi, 2006, au sujet de l'inclusion des probabilités de sélection.

5. Analyse des données

Dans la présente section, nous présentons une analyse des données sur l'IMC à l'aide de cinq modèles. Les deux premiers sont le modèle de sélection de non-réponse avec régression spline (1) (Nandram et Choi, 2005) et le modèle de sélection de non-réponse avec régression spline et probabilités de sélection (2) (Nandram et Choi, 2006). Il s'agit de modèles paramétriques. Les trois autres sont le modèle de régression logistique (3), le modèle de régression logistique robustifié (4) et le modèle de régression logistique robustifié avec probabilités de sélection (5). Il s'agit de modèles non paramétriques puisqu'il n'existe pas d'hypothèses quant à la distribution des valeurs de l'IMC des non-répondants et de la population non échantillonnée. Pour l'inférence, nous avons présenté une moyenne a posteriori (MP), un écart-type a posteriori (ÉTP), une erreur-type numérique (ETN) et un intervalle crédible à 95 %. Dans les tableaux 1 et 2, nous avons étudié l'inférence des coefficients de régression; dans le tableau 3, nous avons prédit les valeurs de l'IMC aux 85^e et 95^e centiles selon le modèle et l'âge pour les hommes blancs. Dans le tableau 1, nous

avons étudié comment les coefficients de régression du modèle de régression logistique robustifié (4) variaient selon les poids, mais non les probabilités de sélection. Comme on peut le constater, les variations selon les poids sont minimes. Ces écarts n'ont pas d'incidence marquée sur l'inférence. Pour tous nos calculs, nous avons donc utilisé $\omega_3 = 0{,}5$ et $\omega_k = (1-\omega_3)/5, k \neq 3$.

Nous avons comparé les coefficients de régression du modèle de régression logistique (3), du modèle de régression logistique robustifié (4) et du modèle de régression logistique robustifié avec probabilités de sélection (5). Dans le tableau 2, on observe des écarts infimes entre ces trois modèles lorsqu'on considère les intervalles crédibles à 95 %. Dans le modèle de régression logistique robustifié avec et sans probabilités de sélection, les valeurs a posteriori devraient être exactement les mêmes; les légers écarts sont attribuables aux erreurs de Monte-Carlo. Les écarts entre le modèle de régression logistique (3) et les modèles logistiques robustifiés (4 et 5) avec et sans probabilités de sélection devraient être supérieurs aux écarts entre ces deux derniers modèles (4 et 5). Mais surtout, l'âge et la race constituent des discriminateurs des répondants et des non-répondants. Par exemple, selon le modèle robuste (5) avec probabilités de sélection, les intervalles crédibles à 95 % pour le coefficient de régression de l'âge et de la race sont $(0{,}932, 1{,}054)$ et $(-0{,}305, -0{,}201)$.

Nous présentons ensuite nos résultats quant à la prédiction des 85^e et 95^e centiles de la population finie. Bien qu'il existe des écarts importants entre les comtés (Nandram et Choi, 2006), nous prenons le comté 11 comme exemple de nos résultats. Toutefois, nous comparons maintenant les cinq modèles pour constater les effets de la robustification et de l'inclusion des probabilités de sélection. Dans le tableau 3, nous présentons le détail de nos résultats et de nos prévisions en ce qui concerne les hommes blancs. Enfin, nous résumons les similitudes et les différences entre les deux modèles paramétriques (les deux premiers) et les trois modèles non paramétriques (les trois derniers). Nous avons relevé trois similitudes : [a.] Les moyennes a posteriori du 85^e centile et du 95^e centile sont respectivement très semblables pour les cinq modèles et la hausse prévue entre le groupe d'âge 1 et le groupe d'âge 4 est semblable. Comme il se doit, les moyennes a posteriori du 95^e centile sont supérieures à celles du 85^e centile. [b.] Les probabilités de sélection ont un double effet. Les moyennes a posteriori et les écarts-types a posteriori sont moins élevés sous les modèles avec probabilités de sélection que sous les modèles sans probabilités de sélection. Ainsi, les probabilités de sélection corrigent en partie le biais de sélection et permettent d'améliorer la précision. [c.] Comme on pouvait s'y attendre, les écarts-types a posteriori pour les 85^e et 95^e centiles sont généralement plus élevés selon les modèles robustes que selon les modèles paramétriques, à quelques exceptions près. Selon les modèles paramétriques, on observe les plus grands écarts-types a posteriori au groupe d'âge 4, alors qu'on les observe au groupe d'âge 3 selon tous les modèles robustes, sauf le modèle de régression logistique robustifié lorsqu'on estime le 95^e centile.

Références

Albert, J. H. et Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, **88**, 669-679.

Greenlees, J. S., Reece, W. S., et Zieschang, K. D.(1982), "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, **77**, 251-261.

Hettmansperger, T. P. (1984), *Statistical Inference Based on ranks*. New York: Wiley.

Little, R. J. A. et Rubin, D. B. (2002), *Statistical Analysis with Missing Data*. New York: Wiley.

Mudholkar, G. S. et George, E. O. (1978), "A remark on the shape of the logistic distribution," *Biometrika*, 65, 667-668.

Nandram, B. et Choi, J.W. (2006), "A Bayesian Assessment of Obesity Among Children and Adolescents from Small Domains Under Nonignorable Nonresponse," *Technical Report*, Worcester Polytechnic Institute.

Nandram, B. et Choi, J.W. (2005), "Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines: Une application aux données de la NHANES", *Techniques d'enquête,* **31**, 79-92.

Nandram, B. et Choi, J. W. (2004), "A Nonparametric Bayesian Analysis of a Proportion for a Small Area Under Nonignorable Nonresponse," *Journal of Nonparametric Statistics*, **16**, 821-839.

National Center for Health Statistics (1994), "Plan and Operation of the Third National Health and Nutrition Examination Survey," *Vital and Health Statistics Series 1*, **32**.

Potvin, C. et Roff, D. A. (1993), "Distribution-free and robust statistical methods: Viable alternatives to parametric statistics?" *Ecology*, **74**, 1617-1628.

Rosenbaum, P. R. (2002), *Observational Studies*. New York: Springer-Verlag.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Tableau 1 : Sensibilité des moyennes a posteriori (MP), des écarts-types a posteriori (ÉTP), des erreurs-types numériques (ETN) et des intervalles crédibles à 95 % pour les coefficients de régression (coeff.) du modèle logistique robustifié (4) sans probabilités de sélection à l'égard du choix des poids

ω_3	Coeff.	MP	ÉTP	ETN	Intervalles
0,3	β_0	1,445	0,034	0,002	(1,378, 1,509)
	β_1	0,990	0,031	0,001	(0,931, 1,053)
	β_2	-0,252	0,025	0,001	(-0,300, -0,201)
	β_3	-0,007	0,024	0,001	(-0,055, 0,039)
	β_4	-0,014	0,026	0,001	(-0,063, 0,040)
0,5	β_0	1,448	0,033	0,002	(1,387, 1,515)
	β_1	0,993	0,030	0,001	(0,931, 1,052)
	β_2	-0,253	0,027	0,001	(-0,303, -0,201)
	β_3	-0,007	0,025	0,001	(-0,059, 0,040)
	β_4	-0,015	0,026	0,001	(-0,064, 0,034)
0,7	β_0	1,448	0,037	0,006	(1,399, 1,524)
	β_1	1,000	0,030	0,003	(0,943, 1,067)
	β_2	-0,246	0,025	0,003	(-0,311, -0,211)
	β_3	-0,004	0,025	0,002	(-0,054, 0,042)
	β_4	-0,024	0,022	0,002	(-0,065, 0,020)
0,9	β_0	1,454	0,035	0,004	(1,401, 1,521)
	β_1	1,005	0,030	0,003	(0,956, 1,065)
	β_2	-0,254	0,028	0,003	(-0,313, -0,207)
	β_3	-0,008	0,026	0,002	(-0,061, 0,043)
	β_4	-0,017	0,024	0,002	(-0,056, 0,035)

NOTA. Les coefficients de régression logistique sont l'ordonnée à l'origine (β_0), l'âge (β_1), la race (β_2), le sexe (β_3) et l'interaction entre la race et le sexe (β_4). Dans le modèle de régression logistique robustifié (4) sans probabilités de sélection, mais avec effets de comté, les poids sont $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6)$; $\omega_k = (1-\omega_3)/5, k \neq 3$ et l'on spécifie ω_3 .

Tableau 2 : Moyennes a posteriori (MP), écarts-types a posteriori (ÉTP), erreurs-types numériques (ETN) et intervalles crédibles à 95 % pour les coefficients de régression (coeff.) selon le modèle

Modèle	Coeff.	MP	ÉTP	ETN	Intervalles
3	β_0	1,474	0,033	0,003	(1,413, 1,536)
	β_1	1,014	0,031	0,002	(0,952, 1,070)
	β_2	-0,260	0,029	0,002	(-0,316, -0,204)
	β_3	-0,006	0,027	0,002	(-0,060, 0,040)
	β_4	-0,017	0,024	0,002	(-0,055, 0,037)
4	β_0	1,448	0,033	0,001	(1,387, 1,515)
	β_1	0,993	0,030	0,001	(0,931, 1,052)
	β_2	-0,253	0,027	0,001	(-0,303, -0,201)
	β_3	-0,007	0,025	0,001	(-0,059, 0,040)
	β_4	-0,015	0,026	0,001	(-0,064, 0,034)
5	β_0	1,452	0,032	0,001	(1,389, 1,518)
	β_1	0,995	0,031	0,001	(0,932, 1,054)
	β_2	-0,252	0,027	0,001	(-0,305, -0,201)
	β_3	-0,007	0,026	0,001	(-0,059, 0,041)
	β_4	-0,015	0,026	0,001	(-0,064, 0,036)

NOTA. Les coefficients de régression sont l'ordonnée à l'origine (β_0), l'âge (β_1), la race (β_2), le sexe (β_3) et l'interaction entre la race et le sexe (β_4). Le sexe et l'interaction entre la race et le sexe ne comptent pas dans la discrimination entre les répondants et les non-répondants. Modèle : 1 (modèle de régression spline); 2 (modèle de régression spline avec probabilités de sélection); 3 (modèle de régression logistique); 4 (modèle de régression logistique robustifié); 5 (modèle de régression logistique robustifié avec probabilités de sélection). Pour les modèles de régression logistique robustifiés (4 et 5), $\omega_3 = 0{,}5$. Les coefficients de régression des modèles 1 et 2 ne sont pas comparables à ceux des modèles 3 à 5; voir Nandram et Choi (2005).

Tableau 3 : Moyennes a posteriori (MP), écarts-types a posteriori (ÉTP) et intervalles crédibles à 95 % pour les 85^{e} et 95^{e} centiles de l'IMC pour les *hommes blancs* du comté 11 selon le modèle et l'âge

		85e centile			95e centile		
Modèle	Âge	MP	ÉTP	Intervalles	MP	ÉTP	Intervalles
1	1	18,0	0,23	(17,5, 18,4)	19,7	0,25	(19,2, 20,2)
	2	18,2	0,23	(17,8, 18,7)	19,9	0,25	(19,4, 20,4)
	3	22,6	0,43	(21,7, 23,4)	24,8	0,48	(23,8, 25,7)
	4	25,1	0,84	(23,4, 26,7)	27,5	0,92	(25,6, 29,3)
2	1	18,1	0,15	(17,8, 18,3)	19,4	0,16	(19,1, 19,8)
	2	18,2	0,15	(17,9, 18,5)	19,7	0,17	(19,3, 20,0)
	3	22,1	0,39	(21,3, 22,8)	23,8	0,42	(22,9, 24,6)
	4	24,3	0,78	(22,7, 26,0)	26,2	0,85	(24,5, 28,0)
3	1	17,9	0,35	(17,2, 18,6)	20,2	0,61	(19,2, 21,5)
	2	18,0	0,37	(17,3, 18,7)	20,4	0,64	(19,4, 21,8)
	3	20,8	0,70	(19,5, 22,3)	23,7	0,92	(22,1, 25,6)
	4	25,5	0,41	(24,8, 26,4)	29,0	0,76	(27,7, 30,6)
4	1	17,9	0,35	(17,2, 18,6)	20,3	0,62	(19,3, 21,6)
	2	18,0	0,35	(17,4, 18,8)	20,5	0,62	(19,4, 21,9)
	3	20,9	0,69	(19,7, 22,3)	23,7	0,95	(22,2, 25,9)
	4	25,6	0,40	(24,8, 26,4)	29,0	0,81	(27,7, 30,9)
5	1	17,1	0,22	(16,8, 17,6)	19,4	0,44	(18,6, 20,4)
	2	17,3	0,23	(16,9, 17,7)	19,5	0,45	(18,8, 20,5)
	3	20,1	0,39	(19,4, 20,9)	22,7	0,63	(21,7, 24,1)
	4	25,3	0,35	(24,7, 26,0)	28,6	0,74	(27,4, 30,3)

NOTA. Âge : 1 (2 à 4 ans), 2 (5 à 9 ans), 3 (10 à 14 ans) et 4 (15 à 19 ans); Modèle : 1 (modèle de régression spline), 2 (modèle de régression spline avec probabilités de sélection), 3 (modèle de régression logistique), 4 (modèle de régression logistique robustifié), 5 (modèle de régression logistique robustifié avec probabilités de sélection).