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

Article

Critère d'information bayésien fondé sur la pseudo-vraisemblance pour la sélection de variables dans les données d'enquête

par Chen Xu, Jiahua Chen et Harold Mantel

Janvier 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

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

Programme des services de dépôt

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

Comment accéder à ce produit

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

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Note de reconnaissance

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

Signes conventionnels

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

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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 333-354
Statistique Canada, N° 12-001-X au catalogue

Critère d'information bayésien fondé sur la pseudo-vraisemblance pour la sélection de variables dans les données d'enquête

Chen Xu, Jiahua Chen et Harold Mantel[1]

Résumé

Les modèles de régression sont utilisés couramment pour analyser les données d'enquête lorsque l'on souhaite déterminer quels sont les facteurs influents associés à certains indices comportementaux, sociaux ou économiques au sein d'une population cible. Lorsque des données sont recueillies au moyen d'enquêtes complexes, il convient de réexaminer les propriétés des approches classiques de sélection des variables élaborées dans des conditions i.i.d. ne faisant pas appel au sondage. Dans le présent article, nous dérivons un critère BIC fondé sur la pseudo vraisemblance pour la sélection des variables dans l'analyse des données d'enquête et proposons une approche de vraisemblance pénalisée dans des conditions de sondage pour sa mise en œuvre. Les poids de sondage sont attribués comme il convient pour corriger le biais de sélection causé par la distorsion entre l'échantillon et la population cible. Dans un cadre de randomisation conjointe, nous établissons la cohérence de la procédure de sélection proposée. Les propriétés en échantillon fini de l'approche sont évaluées par des analyses et des simulations informatiques en se servant de données provenant de la composante de l'hypertension de l'Enquête sur les personnes ayant une maladie chronique au Canada de 2009.

Mots-clés : Sélection des variables; poids de sondage; inférence sous le modèle et sous le plan; BIC; vraisemblance pénalisée; cohérence de sélection.

1 Introduction

La détermination des facteurs influents associés à certains indices comportementaux, sociaux ou économiques dans une population cible est un sujet d'intérêt commun à de nombreux domaines de recherche scientifique. Par exemple, les sociologues souhaitent cerner les facteurs importants qui ont une incidence sur le taux de chômage dans une région particulière et les épidémiologistes cherchent à découvrir les comportements à risque associés aux maladies. Dans ce genre d'études, les chercheurs commencent souvent par effectuer un sondage auprès de la population cible (par exemple, Rahiala et Teräsvirta 1993; Korn et Graubard 1999; Wolfson 2004). Pour cela, un échantillon représentatif est sélectionné et des mesures des variables d'intérêt sont recueillies auprès des unités échantillonnées. Un modèle de régression est habituellement employé pour résumer l'information contenue dans les données. Ce modèle explique les variations de la variable réponse au moyen d'une fonction simple des variables explicatives (covariables). Lorsqu'ils ne disposent pas d'information a priori, les chercheurs peuvent recueillir des renseignements sur de nombreuses variables explicatives possibles. Ils peuvent ensuite atteindre l'objectif consistant à déterminer quels sont les facteurs influents en appliquant une procédure de sélection de variables.

La sélection des variables est un aspect fondamental de la modélisation statistique. Dans des conditions ne faisant pas appel au sondage, on a élaboré des critères de sélection classiques pour évaluer et sélectionner les variables possibles. La statistique C_p de Mallow (Mallows 1973), la validation croisée

1. Chen Xu et Jiahua Chen, Département de statistique, Université de la Colombie-Britannique, Vancouver (C.-B.), Canada, V6T 1Z4. Courriel : chen.xu@stat.ubc.ca et jhchen.stat.ubc.ca; Harold Mantel, Division de la recherche et de l'innovation en statistique, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6. Courriel : Harold.Mantel@statcan.gc.ca.

(généralisée) (CV/ GCV; Stone 1974; Craven et Wahba 1979), le critère d'information d'Akaike (AIC; Akaike 1973) et le critère d'information bayésien (BIC; Schwarz 1978) en sont des exemples. Tous ces critères sont fort utiles et produisent des inférences significatives en pratique.

Malgré l'abondance de la littérature sur la sélection des variables, peu d'attention a été accordée à ce sujet dans le contexte du sondage. L'application des méthodes de sélection de variables à des données d'enquête peut donner lieu à de nombreuses complications. Nous nous concentrons sur les problèmes qui découlent de caractéristiques particulières des enquêtes. Premièrement, les données recueillies par sondage sont habituellement obtenues auprès d'une population finie sans remise des unités échantillonnées, de sorte qu'elles possèdent une structure de dépendance intrinsèque. Deuxièmement, dans les plans de sondage complexes, les probabilités d'inclusion des unités échantillonnées varient souvent dans la population cible. Par conséquent, la corrélation entre la réponse et les covariables dans l'échantillon peut être faussée comparativement à celle observée dans la population. Cela pourrait être le cas lorsque certaines parties de la population sont échantillonnées de manière plus intensive que d'autres. Ne pas tenir compte du plan de sondage dans le processus de sélection peut donner des résultats biaisés pour la population cible.

Dans la littérature, les poids de sondage sont souvent utilisés pour estimer les paramètres des modèles de régression fondés sur des données d'enquête. Les estimations pondérées des coefficients de régression aident à éviter que l'inférence soit biaisée par l'échantillonnage informatif (Pfeffermann 1993; Fuller 2009, section 6.3; Skinner 2012). Même si l'estimation et la sélection du modèle ont chacune leurs propres objectifs, elles présentent souvent un lien cohérent dans un processus de modélisation. Il est donc naturel de conjecturer que l'utilisation de poids de sondage a un effet positif sur la sélection des variables.

Dans cet esprit, nous étudions l'utilisation de la pseudo-vraisemblance pour tenir compte des poids de sondage, et nous dérivons un critère BIC fondé sur la pseudo-vraisemblance pour la sélection des variables dans les données d'enquête. Nous proposons en outre une procédure fondée sur la pseudo-vraisemblance pénalisée (PVP) pour la mise en œuvre numérique du critère proposé. Dans un cadre de randomisation conjointe, nous prouvons que la nouvelle procédure permet systématiquement de repérer les variables influentes. Nous évaluons la méthode de sélection pondérée au moyen d'études en simulation en utilisant des données provenant de l'Enquête sur les personnes ayant une maladie chronique au Canada réalisée en 2009.

La présentation de l'article est la suivante. À la section 2, nous décrivons le mécanisme de randomisation conjointe et le modèle de superpopulation. À la section 3, nous dérivons le critère BIC fondé sur la pseudo-vraisemblance pour l'analyse des données d'enquête et proposons de l'appliquer au moyen de la procédure PVP. À la section 4, nous étudions le comportement asymptotique de la procédure BIC proposée. À la section 5, nous faisons appel à des études numériques pour évaluer plus en détail les résultats de notre approche et à la section 6, nous présentons nos conclusions. Nous donnons les preuves des théorèmes dans un supplément technique distinct [Xu et Chen (2012)], dans lequel figure également la dérivation du critère BIC proposé.

2 Inférence conjointe et superpopulation

Le comportement aléatoire d'une procédure d'inférence découle principalement du caractère aléatoire des données. Dans le contexte des enquêtes, l'ensemble d'unités échantillonnées est aléatoire en raison du

plan d'échantillonnage probabiliste. Parallèlement, la valeur de chaque unité échantillonnée peut être considérée comme un résultat aléatoire provenant d'une superpopulation conceptuellement infinie (Royall 1976).

Dans une analyse fondée sur le plan de sondage, la population finie est considérée comme non aléatoire et toutes les mesures des unités d'échantillonnage sont constantes. Les paramètres d'intérêt sont les quantités dans la population finie, telles que le total ou la médiane de la population. L'inférence statistique est évaluée en se basant sur le caractère aléatoire découlant du plan de sondage probabiliste.

On peut également considérer le caractère aléatoire induit par le plan de sondage comme un artéfact. Les mesures des unités échantillonnées sont alors des réalisations indépendantes d'une variable aléatoire provenant d'un modèle probabiliste de la superpopulation postulée. Des paramètres d'intérêt sont reliés au modèle hypothétique et les inférences sous le modèle sont évaluées uniquement en se basant sur la randomisation introduite par le modèle.

Une troisième approche, appelée inférence sous le modèle et le plan, incorpore la randomisation venant du plan de sondage ainsi que du modèle. Sous un tel mécanisme de randomisation conjointe, la population finie est considérée comme un échantillon aléatoire tiré d'une superpopulation. L'échantillon d'enquête est considéré comme résultant d'un échantillonnage de deuxième phase de la superpopulation. Les paramètres d'intérêt peuvent être des paramètres du modèle ou des paramètres de population finie. Sous ce mécanisme, les inférences au sujet des paramètres de la population finie sont motivées par le modèle de superpopulation. L'inférence sous le modèle et le plan de sondage peut être plus efficace que les approches fondées purement sur le plan lorsque la population finie est bien décrite par le modèle de superpopulation. Comparativement aux approches fondées purement sur le modèle, elle protège contre la violation du modèle et est par conséquent généralement plus robuste (voir, par exemple, Binder et Roberts 2003; Kalton 1983).

Nous étudions le problème de la sélection des variables sous le mécanisme de randomisation conjointe. Soit $\mathcal{D} = \{1, \ldots, N\}$ une population finie constituée de N unités échantillonnées. Les mesures faites sur la i^{e} unité sont désignées $(y_i, \mathbf{x}_i)$, où y_i est la réponse d'intérêt et $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ est un vecteur de variables explicatives de dimension p (vecteur de covariables). Ces éléments sont considérés comme des réalisations indépendantes de $(Y, \mathbf{X})$ provenant d'une superpopulation. Nous postulons un modèle linéaire généralisé (MLG) sur la superopulation de la façon suivante. Conditionnellement à $\mathbf{X}$, la loi de Y appartient à une famille exponentielle naturelle, dont la densité prend la form

$$f(y;\theta) = c(y)\exp\{\theta y - b(\theta)\}. \tag{2.1}$$

θ est connu comme étant le paramètre naturel de $f(y;\theta)$ tel que $b'(\theta) = E[Y|\mathrm{X}] \equiv \mu$ et $b''(\theta) = \mathrm{Var}[Y|\mathrm{X}] \equiv \sigma^2$, et $c(y)$ est une mesure de base non négative. L'influence de la variable explicative $\mathbf{X}$ sur Y est exprimée par $g(\mu) = \mathbf{X}^T\boldsymbol{\beta}$ pour une certaine fonction de lien supposée $g(.)$, où le vecteur $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$ est le coefficient de régression de dimension p. Si $g(.)$ est le lien canonique, c'est-à-dire $g(\mu) = \theta$, alors nous avons $\theta = \mathbf{X}^T\boldsymbol{\beta}$. Pour simplifier, nous nous concentrons sur le lien canonique dans le présent article.

Sur la base de ce modèle, l'effet de la variable explicative est caractérisé par la taille du coefficient de régression correspondant. Dans les applications, un modèle complexe contenant de nombreuses variables aboutit souvent à un surajustement et à une médiocre capacité d'interprétation. Donc, il est souhaitable d'ajuster les données au moyen d'un modèle parcimonieux dans lequel de nombreux coefficients de régression sont estimés être nuls. Les variables explicatives dont les coefficients ne sont pas nuls sont alors considérées comme influant sur la réponse. À cette fin, nous supposons que $\boldsymbol{\beta}$ est idéalement parcimonieux et nous abordons le problème de sélection des variables en déterminant un modèle parcimonieux formé par les covariables dont les coefficients ne sont pas nuls.

3 Sélection fondée sur la pseudo-vraisemblance avec le BIC

3.1 Le BIC dans les enquêtes

Sous la spécification du modèle décrite à la section 2, il est clair que, si la mesure $\left(y_i, \mathbf{x}_i\right)$ est observée pour chaque unité de la population $\mathcal{D}$, le caractère aléatoire des données introduit par le plan de sondage probabiliste a complètement disparu. Dans cette situation, la sélection de variables influentes est fondée sur la population complète et les critères de sélection classiques élaborés dans des conditions ne faisant pas appel au sondage (fondées purement sur le modèle) demeurent valides pour l'inférence sous le modèle et le plan. En particulier, soit $s \subseteq \{1, \ldots, p\}$ un ensemble arbitraire de $\tau(s)$ covariables, qui correspond à un modèle possible de la forme (2.1). Le BIC fondé sur la population complète (Schwarz 1978) sélectionne le modèle (covariables) qui minimise

$$\mathrm{BIC}_N(s) = -2l_N\left(\breve{\boldsymbol{\beta}}_s\right) + \tau(s)\log N, \tag{3.1}$$

où $l_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log f\left(y_i; \mathbf{x}_i\boldsymbol{\beta}\right)$ est la fonction de vraisemblance pour la population complète et $\breve{\boldsymbol{\beta}}_s$ est le maximiseur de $l_N(\boldsymbol{\beta})$ fondé sur s. On peut constater que le BIC (3.1) est une fonction décroissante de la vraisemblance maximisée et une fonction croissante du nombre de variables incluses dans le modèle. Donc, un plus petit BIC implique un modèle plus simple (moins de variables explicatives), un meilleur ajustement (vraisemblance maximisée plus élevée), ou les deux. La préférence est donnée à un modèle présentant un équilibre entre la complexité et la qualité de l'ajustement.

Nous notons que le BIC sous population complète (3.1) est conceptuel, parce que l'observation de $\left(y_i, \mathbf{x}_i\right)$ pour toutes les unités de $\mathcal{D}$ est habituellement impossible dans les applications. Souvent, on tire plutôt de $\mathcal{D}$ un échantillon représentatif $d = \{i_1, \ldots, i_n\} \subset \{1, \ldots, N\}$ contenant n unités et les mesures sont observées en se basant sur les unités échantillonnées. En raison de la structure de dépendance intrinsèque des unités échantillonnées, il n'est généralement pas possible de calculer une vraisemblance complète sur d. Comme solution de rechange, pour l'inférence sous le modèle et le plan, on utilise fréquemment une fonction de pseudo-log-vraisemblance, qui prend la forme

$$l_n(\boldsymbol{\beta}) = \sum_{i \in d} w_i \log f(y_i; \boldsymbol{\beta}) \tag{3.2}$$

où $w_i = k / P(i \in d)$ désigne le poids de sondage de la i^{e} unité. Le paramètre d'échelle k dans w_i n'a aucune incidence analytique sur l'inférence fondée sur la pseudo-vraisemblance. Pour simplifier l'exposé, nous choisissons $k = n / N$ tel que $n^{-1} l_n(\boldsymbol{\beta})$ est sans biais sous le plan jusqu'à $N^{-1} l_N(\boldsymbol{\beta})$. La maximisation de $l_n(\boldsymbol{\beta})$ sur $\boldsymbol{\beta}$ mène à un estimateur du maximum de pseudo-vraisemblance (EMPV) $\hat{\boldsymbol{\beta}}$ pour $\boldsymbol{\beta}$, c'est-à-dire

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} l_n(\boldsymbol{\beta}).$$

Sous les plans de sondage appropriés, $\hat{\boldsymbol{\beta}}$ est souvent convergent en $n^{-1/2}$ vers $\boldsymbol{\beta}$ dans le cadre de randomisation conjointe. L'idée d'utiliser la pseudo-vraisemblance pour l'inférence sur les paramètres du modèle est largement répandue dans la littérature (voir, par exemple, Binder 1983; Godambe et Thompson 1986; Molina et Skinner 1992).

Dans le présent article, nous tentons d'élaborer un analogue du critère BIC fondé sur la pseudo-vraisemblance. Partant de la formulation de la super-population décrite à la section 2, soit $\boldsymbol{\beta}_s$, le coefficient $\tau(s)$- dimensionnel du modèle s et soit ν_s, la densité a priori de $\boldsymbol{\beta}_s$. Alors, une fonction de pseudo-densité marginale des données est donnée par

$$P_n(\mathbf{y}|s) = \int L_n(\mathbf{y}; \boldsymbol{\beta}_s) \nu_s(\boldsymbol{\beta}_s) d\boldsymbol{\beta}_s$$

avec $L_n(\mathbf{y}; \boldsymbol{\beta}_s) = \exp\{l_n(\mathbf{y}; \boldsymbol{\beta}_s)\}$. Donc, nous pouvons considérer l'expression qui suit comme étant la pseudo-probabilité a posteriori du modèle s :

$$P_n(s|\mathbf{y}) = \frac{P_n(\mathbf{y}|s) P(s)}{\sum_{s \in S} P(s) P_n(\mathbf{y}|s)}, \tag{3.3}$$

où S désigne l'ensemble de tous les modèles possibles. Dans l'esprit de l'analyse bayésienne, le modèle ayant la $P_n(s|\mathbf{y})$ la plus élevée est considéré comme étant celui que les données soutiennent le plus. Puisque $\sum_{s \in S} P(s) P_n(\mathbf{y}|s)$ ne dépend d'aucun modèle particulier, la $P_n(s|\mathbf{y})$ la plus élevée est donnée par le modèle qui maximise la $P_n(\mathbf{y}|s) P(s)$ correspondante. Lorsque l'on utilise le prior uniforme $P(s) = \zeta$ et que l'on choisit le facteur d'échelle de pondération comme étant $k = n / N$, on obtient une approximation de Laplace sous certaines conditions de régularité (voir Xu et Chen 2012) :

$$-2 \log\{P_n(\mathbf{y}|s)\} = -2 l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s) \log n + O_p(1).$$

D'où, nous choisissons le modèle s qui minimise

$$\text{BIC}_n(s) = -2 l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s) \log n. \tag{3.4}$$

Comparativement au BIC sous population complète (3.1), le premier terme du BIC (3.4) est la pseudo-vraisemblance pondérée par les poids de sondage maximale, qui pourrait être utile pour éviter les erreurs dues à l'échantillonnage susceptibles de donner lieu à des inférences biaisées pour la population cible. Nous considérons (3.4) comme une version du BIC fondée sur la pseudo-vraisemblance dans le contexte des sondages. Dans le cadre de randomisation conjointe, nous établissons la cohérence de sélection lorsqu'on utilise le BIC (3.4) par une procédure d'application via la pseudo-vraisemblance pénalisée (PVP), comme nous le verrons à la section 4.

3.2 Application du BIC au moyen de la pseudo-vraisemblance pénalisée

Dans la pratique, un moyen simple d'appliquer le BIC consiste à sélectionner le meilleur sous-ensemble, en évaluant et comparant le BIC pour chaque modèle possible. Cependant, cette procédure peut aboutir à des calculs impossibles quand le nombre de covariables est grand. Pour la remplacer, des méthodes basées sur la vraisemblance pénalisée ont été utilisées récemment comme procédures de calcul efficaces pour appliquer un critère de sélection. Pour exclure des variables du modèle, ces méthodes estiment que les coefficients de ces variables sont nuls et réduisent les autres coefficients en conséquence. En faisant varier la pénalité appliquée à la vraisemblance, nous pouvons obtenir une série de modèles de parcimonie variable. Afin d'éviter une recherche exhaustive sur l'entièreté de l'espace des modèles, on utilise un critère de sélection pour choisir un modèle optimal parmi ces modèles parcimonieux. L'efficacité de cette stratégie a été illustrée dans un contexte ne faisant pas appel au sondage pour le critère BIC (Wang, Li et Tsai 2007; Liu, Wang et Liang 2011) et pour le critère GCV (Fan et Li 2001; Xie, Pan et Shen 2008) entre autres.

Dans le même esprit, nous proposons une procédure fondée sur la pseudo-vraisemblance pénalisée (PVP) pour appliquer le BIC (3.4) aux données d'enquête. En particulier, partant de la pseudo-vraisemblance (3.2) avec $k = n / N$, nous définissons l'estimateur pénalisé pondéré par les poids de sondage $\hat{\boldsymbol{\beta}}_\lambda$, qui minimise la fonction de pseudo-vraisemblance pénalisée.

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n\sum_{j=1}^{p} \phi_\lambda\left(|\beta_j|\right), \tag{3.5}$$

où $\phi_\lambda(.)$ est une fonction de pénalité indicée par un paramètre d'ajustement λ qui contrôle la taille de la pénalité. Moyennant un choix approprié de $\phi_\lambda(.)$, $\hat{\boldsymbol{\beta}}_\lambda$ contient des estimations nulles pour certains coefficients et produit donc automatiquement un modèle parcimonieux. La parcimonie souhaitable de $\hat{\boldsymbol{\beta}}_\lambda$ exige habituellement que la fonction $\phi_\lambda(.)$ correspondante soit singulière à l'origine. Certains choix fréquents de $\phi_\lambda(.)$ comprennent la pénalité L_γ (Frank et Friedman 1993; Tibshirani 1996), c'est-à-dire $\phi_\lambda\left(|\beta|\right) = \lambda|\beta|^\gamma$ avec $\gamma \in (0,1]$, et la pénalité SCAD (Fan et Li 2001), qui est définie par la dérivée suivante :

$$\phi'_\lambda\left(|\beta|\right) = \lambda\left\{I\left(|\beta| \le \lambda\right) + \frac{\left(a\lambda - |\beta|\right)_+}{(a-1)\lambda} I\left(|\beta| > \lambda\right)\right\} \tag{3.6}$$

pour laquelle $a = 3{,}7$ est un choix fréquent.

En utilisant des valeurs différentes de λ pour une fonction $\phi_\lambda(.)$ spécifiée correctement, $\hat{\boldsymbol{\beta}}_\lambda$ produit des modèles de parcimonie variable. Ces modèles parcimonieux (par rapport à λ) forment naturellement une série des modèles possibles. Le BIC (3.4) peut alors être utilisé pour choisir un modèle optimal parmi cette série. Plus précisément, soit Ω l'intervalle de valeurs de λ et soit s_λ un modèle produit par $\hat{\boldsymbol{\beta}}_\lambda$. Nous traitons $S_\Omega = \{s_\lambda : \lambda \in \Omega\}$ comme la série de modèles possibles prise en considération et nous choisissons le modèle $s^* \in S_\Omega$ tel que $\text{BIC}_n(s^*) = \min_{\lambda\in\Omega} \text{BIC}(s_\lambda)$. Nous appelons cette procédure de sélection la méthode du BIC fondée sur la pseudo-vraisemblance pénalisée (BIC-PVP). Comparativement à la sélection classique du meilleur sous-ensemble, la procédure BIC-PVP est axée sur les modèles qui sont produits par les estimateurs analysés pondérés par les poids de sondage et, par conséquent, peut demander nettement moins de calculs.

4 Convergence de la procédure BIC-PVP

Nous examinons maintenant le comportement asymptotique de la procédure BIC-PVP dans le cadre de randomisation conjointe. Nous supposons qu'il existe une série de populations finies, disons $\mathcal{D}_r$ avec $r \to \infty$. Chaque $\mathcal{D}_r$ est un échantillon indépendant et identiquement distribué (i.i.d.) de taille N_r tiré d'une superpopulation modélisée par (2.1) avec la variable aléatoire $\left(Y, \mathbf{X} = \{X_1, \ldots, X_p\}\right)$. Dans chaque population $\mathcal{D}_r$, on tire un échantillon d_r de taille n_r selon un certain plan d'échantillonnage. Nous supposons que N_r ainsi que n_r tendent vers l'infini quand $r \to \infty$, la fraction d'échantillonnage n_r / N_r étant bornée par une constante $C < 1$. Pour simplifier la notation, nous abandonnons l'indice r dans la suite de la discussion.

Sans perte de généralité, nous supposons que les q premiers coefficients ne sont pas nuls et nous désignons la valeur réelle de $\boldsymbol{\beta}$ par $\boldsymbol{\beta}_0 = \{\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}\}$ avec $\boldsymbol{\beta}_{02} = 0$. En outre, nous utilisons s_0 pour désigner le modèle réel $\{1, \ldots, q\}$ qu'il faut identifier. Nous établissons la convergence de sélection de la procédure BIC-PVP en deux étapes. À la première étape, nous montrons que, pour des choix appropriés de $\phi_\lambda(.)$, la PVP peut systématiquement identifier le vrai s_0 de sorte que $s_0 \in S_\Omega$ avec la probabilité tendant vers 1. À la deuxième étape, nous vérifions que le BIC (3.4) sélectionne systématiquement s_0 parmi S_Ω.

Pour l'analyse asymptotique, nous définissons $\varphi_\lambda = \max\left\{\phi'_\lambda\left(|\beta_{0j}|\right) \text{ pour } j \in s_0\right\}$ et associons λ à n pour faire de φ_λ une séquence. Sous le cadre de randomisation conjointe, nous montrons l'allégation de l'étape 1 sous la forme du théorème suivant.

Théorème 1 Sous des conditions de régularité appliquées au modèle (2.1) et d'autres exigences spécifiées dans le supplément en ligne, si $\varphi_\lambda \to 0$ quand $n \to \infty$, il existe alors un maximiseur local $\hat{\boldsymbol{\beta}}_\lambda = \left(\hat{\boldsymbol{\beta}}_{\lambda 1}, \hat{\boldsymbol{\beta}}_{\lambda 2}\right)$ de la fonction de pseudo-vraisemblance pénalisée (3.5) tel que

$$\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-1/2} + \varphi_\lambda\right) \quad \text{et} \quad P\left\{\hat{\boldsymbol{\beta}}_{\lambda 2} = 0\right\} \to 1$$

avec $\|.\|$ désignant la norme euclidienne.

Le résultat de convergence du théorème 1 est vérifié pour les fonctions de pénalité non convexes fréquemment utilisées. Par exemple, pour la pénalité L_γ avec $\gamma \in (0,1)$, la convergence est vérifiée si $\lambda \to 0$; pour la pénalité SCAD, la convergence est vérifiée si $\lambda \to 0$ et $\sqrt{n}\lambda \to \infty$. Cela implique aussi que, si la probabilité tend vers 1, le modèle réel s_0 est inclus dans S_Ω, ce qui sert de condition préalable pour la convergence de sélection du BIC sur S_Ω.

Nous établissons maintenant la convergence en utilisant le BIC sur S_Ω avec une fonction $\phi_\lambda(.)$ spécifiée qui satisfait le théorème 1. En se servant de la notation de la section 3.2, soit s_λ le modèle correspondant à l'estimateur PVP $\hat{\boldsymbol{\beta}}_\lambda$, et soit Ω l'intervalle de valeurs de λ pris en considération. Nous définissons deux séries de modèles possibles comme il suit :

- modèles surajustés : $S_+ = \{s : s_0 \subset s, s \neq s_0\}$;
- modèles sous-ajustés : $S_- = \{s : s_0 \not\subset s\}$.

La notation $\not\subset$ indique ici qu'il existe au moins un élément différent entre deux ensembles, de sorte que S_- est la série de modèles possibles qui ne comprend pas toutes les variables figurant dans le modèle réel. Alors, Ω peut être partitionné en conséquence en

$$\Omega_+ = \{\lambda : s_\lambda \in S_+\}, \quad \Omega_- = \{\lambda : s_\lambda \in S_-\}, \quad \Omega_0 = \{\lambda : s_\lambda = s_0\}. \tag{4.1}$$

Au moyen du théorème 1, nous avons montré que $P(\Omega_0 \neq \varnothing) \to 1$. Par conséquent, la convergence de sélection du BIC sur S_Ω est atteinte si le BIC est capable d'identifier s_0 pour tout modèle s_λ avec $\lambda \in \Omega_+ \cup \Omega_-$. Nous utilisons le théorème qui suit pour établir ce résultat de convergence.

Théorème 2 Sous les mêmes conditions qu'au théorème 1,

$$P\left\{\min_{\lambda \in \Omega_+ \cup \Omega_-} \mathrm{BIC}_n(s_\lambda) \leq \mathrm{BIC}_n(s_0)\right\} \to 0,$$

où Ω_+ et Ω_- sont définis dans (4.1).

5 Études numériques

Afin d'évaluer les résultats de la procédure BIC-PVP sur échantillon fini, nous avons procédé à des études numériques approfondies en nous servant de données tirées de l'Enquête sur les personnes ayant une maladie chronique au Canada (EPMCC, Statistique Canada 2009). En particulier, nous comparons la procédure proposée aux méthodes classiques ne faisant pas appel au sondage en nous basant sur des modèles de régression postulés entre les variables de l'EPMCC et des réponses hypothétiques (simulées).

Nous livrons provisoirement certaines données sur l'utilisation de la sélection fondée sur la pseudo-vraisemblance sous deux scénarios de simulation. Dans le premier scénario, les populations sont générées à partir de modèles présumés et les échantillons sont obtenus en se servant de plans de sondage susceptibles de créer de fausses corrélations entre les variables de l'EPMCC. Dans le deuxième scénario, les populations ne sont pas générées exactement à partir des modèles présumés et les échantillons sont obtenus au moyen d'un plan de sondage en rapport avec la réponse ainsi que les covariables possibles. En outre, nous présentons l'analyse des données originales de l'EPMCC de 2009 à titre d'exemple d'utilisation de la procédure BIC-PVP dans des applications réelles.

5.1 Données de l'EPMCC

L'EPMCC est une étude transversale parrainée par l'Agence de la santé publique du Canada conçue pour recueillir des renseignements concernant les expériences des Canadiens atteints de maladies chroniques. L'un des principaux objectifs de l'EPMCC est de déterminer les comportements influant sur la santé qui ont une incidence sur les résultats de la maladie, afin que le gouvernement puisse planifier des services de santé pour les personnes atteintes de maladie chronique et leur fournir ces services.

L'EPMCC est réalisée tous les deux ans et chaque cycle de l'enquête porte sur deux maladies chroniques. L'enquête de 2009 portait sur l'arthrite et l'hypertension. Nous nous limitons ici aux données sur l'hypertension. La population cible pour l'enquête sur l'hypertension correspond aux Canadiens de 20 ans et plus des dix provinces ayant reçu un diagnostic d'hypertension et vivant dans les logements privés. Pour faciliter le processus d'enquête, les unités d'échantillonnage de l'EPMCC de 2009 sont les personnes atteintes d'hypertension qui avaient participé à l'Enquête sur la santé dans les collectivités canadiennes (ESCC) de 2008. Pour les besoins de l'EPMCC, la population correspondant aux répondants de l'ESCC est d'abord stratifiée selon le sexe et quatre groupes d'âge, à savoir 20 à 44 ans, 45 à 64 ans, 65 à 74 ans et 75 ans et plus. Par conséquent, la population finie formée par les répondants de l'ESCC a été subdivisée en huit catégories d'âge (4 niveaux) selon le sexe (2 niveaux). L'EPMCC est réalisée selon un plan d'échantillonnage stratifié avec répartition proportionnelle à la taille de l'échantillon. Un échantillon global de 9 005 personnes a été sélectionné parmi les 17 437 répondants à l'ESCC, et 6 142 de ces personnes ont participé à l'EPMCC.

Nous avons cerné 40 variables en rapport avec l'hypertension en nous basant sur les données originales de l'EPMCC et, pour 7 de ces variables des données complètes existaient pour les 6 142 répondants. Pour les 33 autres variables, une certaine quantité de valeurs manquaient en raison des cas de non-réponse au questionnaire original (voir le tableau 5.5. en annexe pour la liste des variables et les taux de non-réponse correspondants). Il n'existe aucune raison systématique évidente expliquant la non-réponse totale. La variable pour laquelle il manque le plus de données est INCDRPR (revenu du ménage), le taux de non-réponse étant de 9,6 %, tandis que la quantité de données manquantes est relativement faible pour les autres variables. Afin de faciliter l'analyse, pour remplacer les données manquantes, nous avons utilisé des méthodes d'imputation simples décrites ci-après. Pour une variable catégorique, nous avons imputé une valeur aux cas de non-réponse en nous servant d'une valeur aléatoire provenant du jeu de réponses; pour une variable continue, nous avons imputé une valeur aux cas de non-réponse en nous servant de la valeur moyenne des réponses. Deux exceptions à ces méthodes d'imputation ont été faites pour les variables BMHX_02 et CNHX_05. La première sert de variable de réponse dans l'analyse des données de la seconde, tandis que la seconde présente des contraintes naturelles sur l'intervalle de ces valeurs. Nous

avons supprimé les 274 observations pour lesquelles des valeurs manquaient pour ces deux variables, ce qui donne un jeu de données de travail de base contenant 5 868 observations. La procédure d'imputation/suppression n'a aucun effet sur l'évaluation de la procédure BIC fondée sur la population simulée. Par contre, elle pourrait introduire un biais dans l'analyse des données réelles. Pourtant, étant donné le faible taux de données manquantes, et la plausibilité que les données manquent au hasard dans le cas particulier, il est peu probable que la conclusion soit gravement affectée.

Puisque l'EPMCC est une enquête de suivi à l'ESCC, les poids d'échantillonnage pour l'EPMCC ont été obtenus au départ d'après les poids des données de l'ESCC. Ils ont ensuite été corrigés pour s'assurer que les répondants à l'EPMCC soient représentatifs de la population cible. Par conséquent, les poids corrigés présentent une variation importante d'une unité échantillonnée à l'autre. Après mise à l'échelle au moyen de $k = n / N \approx 10^{-3}$, les poids corrigés varient entre 0,01 et 33,62 avec un intervalle interquartile de 0,76.

5.2 Scénario 1 : Corrélation faussée

Comme il est mentionné plus haut, sous des plans de sondage complexes, la structure de corrélation entre les variables reflétée par l'échantillon peut être faussée comparativement à la population. Dans le premier scénario de simulation, nous évaluons la méthode BIC proposée lorsque les données sont recueillies au moyen de plans de sondage susceptibles de créer des corrélations faussées entre les covariables possibles. En particulier, nous traitons les 40 variables cernées comme des covariables possibles pour une réponse hypothétique Y, et pour simplifier, nous les indiçons de X_1 à X_{40}. Nous considérons des réponses continues ainsi que des réponses binaires dans nos simulations. Pour les cas continus, nous générons les valeurs de Y selon les modèles

- Modèle 1 : $Y = 0,7X_6 + 0,7X_{10} + 0,6X_{18} - 0,6X_{22} + \varepsilon,$
- Modèle 2 : $Y = 0,7X_6 + 0,6X_{10} + 0,6X_{18} - 0,5X_{22} + 0,3X_{30} - 0,3X_{34} + \varepsilon,$

avec $\varepsilon \sim N(0,1)$. Pour les cas binaires, où $Y \in \{0,1\}$, nous générons les valeurs de Y selon les modèles logistiques

- Modèle 3 : $\text{logit}\left(\Pr\{Y = 1 | \mathbf{X}\}\right) = 0,7X_7 - 0,6X_8 + 0,5X_{26},$
- Modèle 4 : $\text{logit}\left(\Pr\{Y = 1 | \mathbf{X}\}\right) = 0,8X_7 - 0,7X_8 + 0,6X_{26} - 0,5X_{28} + 0,4X_{36}.$

Les modèles spécifiés comprennent l'un des identificateurs de strate de l'EPMCC (c'est-à-dire X_6 ou X_7) avec une structure emboîtée pour chaque contexte de modélisation.

La population finie utilisée dans la simulation a été créée comme il suit. Le jeu de données de travail de base de 5 868 répondants a été reproduit 10 fois proportionnellement aux valeurs entières arrondies des poids de sondage de l'EPMCC, ce qui a donné une population pseudo-finie ayant une taille de 55 950 avec information complète sur $X_1, \ldots, X_{40}$. Les valeurs de la réponse Y ont ensuite été générées en se basant sur les modèles 1 à 4 respectivement. Nous considérons le problème de sélection des variables comme consistant à déterminer le modèle postulé qui génère les valeurs de Y.

Nous étudions les résultats de la procédure proposée sous deux plans d'échantillonnage stratifiés. En particulier, nous créons quatre strates en nous basant sur les variables X_6 (âge, moins de 55 ans/55 ans et plus) et X_7 (sexe, Hommes/Femmes), ce qui donne le groupe (Femmes, moins de 55 ans) de taille 7 120, le groupe (Femmes, 55 ans et plus) de taille 19 199, le groupe (Hommes, moins de 55 ans) de taille 6 187 et le groupe (Hommes, 55 ans et plus) de taille 23 458. Sous le premier plan, on tire de chaque strate un échantillon aléatoire simple sans remise (EASSR) avec répartition égale des tailles d'échantillon. L'inférence est basée sur les quatre EASSR regroupés. Sous le deuxième plan, nous construisons en outre trois sous-groupes dans chaque strate en nous basant sur la somme de deux covariables binaires des deux modèles postulés. En particulier, nous construisons les sous-groupes d'après $X_{18} + X_{22}$ pour les données générées par les modèles 1 et 2, et nous construisons de la même façon les sous-groupes fondés sur $X_8 + X_{26}$ pour les données générées par les modèles 3 et 4. Puis, nous effectuons l'inférence en nous basant sur les EASSR tirés de chaque sous-groupe des quatre strates. La taille globale d'échantillon est répartie de manière égale au niveau de la strate avec une proportion de 2 pour 1 pour 2 pour les trois sous-groupes à l'intérieur d'une même strate. Un simple calcul Monte Carlo révèle que la corrélation d'échantillon entre X_{18} et X_{22} (pour les données provenant des modèles 1 et 2) peut être aussi élevé que 0,5, tandis que leur corrélation dans la population est à peine de l'ordre de 0,02. Nous observons un phénomène similaire pour les variables X_8 et X_{26} (des données provenant des modèles 3 et 4). Par conséquent, nous nous attendons à ce que la sélection des variables sous le deuxième plan d'échantillonnage soit plus difficile en raison de cette augmentation systématique. Dans les simulations, nous fixons la taille globale d'échantillon à $n = 500$ pour les modèles 1 et 2 et à $n = 1\,500$ pour les modèles 3 et 4. Un résumé des variables influant sur la réponse et des variables de plan ayant une incidence sur les probabilités d'échantillonnage figure en annexe (tableau A.2).

Nous avons exécuté la procédure de sélection BIC-PVP sur les échantillons probabilistes tirés de la population finie. En particulier, nous avons mis à l'échelle les poids de sondage comme il est mentionné dans (3.2) et nous avons choisi la pénalité SCAD pour la fonction de vraisemblance pénalisée (3.5). Nous avons résolu le maximiseur correspondant de (3.5) en nous servant de l'algorithme de seuillage itératif (She 2011). Aux fins de comparaison, nous utilisons également les critères AIC et GCV comme autres options pour le BIC (3.4) proposé. Partant de la discussion de la section 3, nous définissons les critères AIC et GCV fondés sur la pseudo-vraisemblance comme étant

$$\mathrm{AIC}_n(s) = -2l_n\left(\hat{\beta}_s\right) + 2\tau(s),$$

$$\mathrm{GVC}_n(s) = -\frac{1}{n}\frac{l_n\left(\hat{\beta}_s\right)}{\left(1 - \tau(s)/n\right)^2},$$

qui sont appliqués de la même façon via la procédure fondée sur la PVP. En outre, pour chaque spécification, nous répétons la procédure de sélection en ignorant tous les poids de sondage (en fixant leur valeur à l'unité). Les résultats de la sélection non pondérée correspondent aux inférences fondées purement sur le modèle tel que discuté à la section 2. En particulier, le BIC fondé sur la pseudo-vraisemblance se réduit au BIC classique (3.1) utilisé dans les situations ne faisant pas appel au sondage.

Dans les tableaux 5.1 et 5.2, nous résumons les résultats des simulations fondées sur 1 000 répétitions en ce qui concerne le taux de sélections positives (TSP), le taux de fausses découvertes (TFD), le taux de

sélections correctes (TSC) et la taille moyenne du modèle (TMM). En particulier, soit s_0 un modèle réel qui génère la population finie et s'_j le modèle sélectionné en se basant sur le j^{e} échantillon, $j = 1, \ldots,$ 1 000. Nous estimons les TSP, TFD, TSC et TMM comme il suit

$$\text{TSP} = \frac{\sum_{j=1}^{1\,000} \tau\left(s_0 \cap s'_j\right)}{1\,000\,\tau\left(s_0\right)}, \quad \text{TFD} = \frac{\sum_{j=1}^{1\,000} \tau\left(s'_j \,/\, s_0\right)}{1\,000\,\tau\left(s'_j\right)},$$

$$\text{TSC} = \frac{\sum_{j=1}^{1\,000} I\left(s'_j = s_0\right)}{1\,0}, \quad \text{TMM} = \frac{\sum_{j=1}^{1\,000} \tau\left(s'_j\right)}{1\,0},$$

où $\tau\left(s\right)$ désigne la taille du modèle s et $I\left(\cdot\right)$ est la fonction indicatrice. De plus, nous évaluons comme suit l'exactitude prédictive du modèle sélectionné. Pour chaque spécification, nous générons un échantillon de test de taille 200 par EASSR à partir de la même population finie que celle dont a été tiré l'échantillon d'apprentissage. Pour les modèles 1et 2, nous utilisons la somme des carrés des résidus (SCR) moyenne calculée sur les données de test comme mesure de la capacité prédictive du modèle sélectionné. Pour les modèles 3 et 4, nous calculons les taux de prédictions positives ainsi que négatives. Plus précisément, soit π^* une valeur repère spécifiée et $\hat{\pi}_i$, la probabilité de succès estimée du i^{e} échantillon de test, $i = 1, \ldots, 200$. Nous prédisons alors la i^{e} réponse y_i par $\hat{y}_i = 1$ si $\hat{\pi}_i > \pi^*$ et $\hat{y}_i = 0$ autrement. Les taux de prédictions correctes sont estimés par

$$\text{TPP} = \frac{\sum_{i \in \{i: y_i = 1\}}^{200} I\left(\hat{y}_i = 1\right)}{\sum_{i=1}^{200} I\left(y_i = 1\right)}, \quad \text{TPN} = \frac{\sum_{i \in \{i: y_i = 0\}}^{200} I\left(\hat{y}_i = 0\right)}{\sum_{i=1}^{200} I\left(y_i = 0\right)}.$$

Nous prenons ensuite la moyenne des TPP et TPN finaux sur 1 000 répliques. Notons qu'ici, le TPP et le TPN sont semblables à la sensibilité et à la spécificité dans les études cliniques, qui indiquent la capacité d'une approche de prédiction 0-1 en ce qui concerne les prédictions positives et négatives correctes. En général, une grande valeur de π^* donne lieu à un TPN élevé mais à un TPP faible. La valeur de π^* doit être spécifiée prudemment dans les applications. Dans nos études en simulation, nous fixons la valeur à $\pi^* = 0{,}5$ pour simplifier.

Les résultats sont encourageants pour la méthode BIC proposée. D'après les tableaux 5.1 et 5.2, nous constatons que pour les modèles sélectionnés en appliquant le critère AIC,le TSP et le TFD sont tous deux élevés, ce qui indique l'inclusion d'un nombre excessif de variables non pertinentes. Comparativement, le BIC réduit significativement le TFD des modèles sélectionnés en sacrifiant légèrement le TSP, et donne lieu à la sélection du modèle dont les tailles sont plus proches de la réalité. Bien que le critère GCV se comporte de la même façon que le BIC sous le modèle linéaire, il donne des résultats concordant avec ceux de l'AIC pour les modèles logistiques pour lesquels les réponses binaires fournissent moins d'information.

Sous le premier plan d'échantillonnage, les probabilités d'inclusion ne sont reliées à Y qu'au moyen d'une seule covariable dans le modèle (c'est-à-dire X_6 ou X_7). La structure de corrélation entre la

réponse et les covariables de la population finie est maintenue en grande partie dans l'échantillon. Par conséquent, on n'observe aucune différence importante entre les procédures de sélection pondérées et non pondérées dans le tableau 5.1.

Nous livrons provisoirement les informations concernant l'utilisation des poids de sondage découlant de l'application du deuxième plan de sondage, où la structure de corrélation dans l'échantillon est systématiquement faussée. Clairement, la corrélation faussée entre les covariables pour les unités échantillonnées détériore l'efficacité des méthodes de sélection, comme en témoignent les valeurs plus faibles du TSP et plus élevées du TFD comparativement à celles obtenues pour les procédures non pondérées. L'intégration des poids de sondage dans le processus de sélection aide à corriger le biais résultant. En particulier, nous avons observé des améliorations appréciables pour la sélection fondée sur le BIC. Dans le cas le plus remarquable (c'est-à-dire modèle 3 du tableau 5.2), le BIC fondé sur la pseudo-vraisemblance donne des résultats considérablement meilleurs que ceux fournis par le BIC classique en faisant passer le TSP de 0,65 à 0,89, ce qui réduit le TFD correspondant qui passe de 0,62 à 0,50. Notre observation fait écho à la justification de la pondération voulant que celle-ci élimine le biais dû à l'échantillonnage informatif (section 6.3, Fuller 2009).

Tableau 5.1
Sélection pour le plan de sondage ne produisant pas de corrélation fortement faussée (premier plan). Les résultats sont résumés en fonction du taux de sélections positives (TSP), du taux de fausses découvertes (TFD), du taux de sélections correctes (TSC) et de la taille moyenne du modèle (TMM); les évaluations de la prédiction pour les modèles 1 et 2 sont fondées sur le test de la somme des carrés des résidus (SCR), et pour les modèles 3 et 4, sur le taux de prédictions positives/négatives (TPP, TPN) avec une valeur repère de 0,5.

Pondérations	Critère	TSP	TFD	TSC	TMM	Prédiction
			Modèle 1			
Ignorée	GCV	0,96	0,19	0,28	4,9	1,04
	AIC	0,99	0,48	0,05	8,7	1,08
	BIC	0,96	0,19	0,28	4,9	1,04
Incluse	GCV	0,95	0,24	0,19	5,2	1,05
	AIC	0,99	0,61	0,01	11,4	1,11
	BIC	0,95	0,24	0,20	5,3	1,05
			Modèle 2			
Ignorée	GCV	0,72	0,19	0,02	5,5	1,07
	AIC	0,89	0,44	0,01	10,3	1,09
	BIC	0,73	0,19	0,03	5,6	1,07
Incluse	GCV	0,74	0,24	0,02	6,1	1,08
	AIC	0,89	0,54	0,01	12,5	1,12
	BIC	0,74	0,24	0,03	6,1	1,08
			Modèle 3			
Ignorée	GCV	0,99	0,59	0,00	7,8	(0,71; 0,45)
	AIC	0,99	0,62	0,00	8,4	(0,69; 0,49)
	BIC	0,96	0,43	0,00	5,1	(0,72; 0,44)
Incluse	GCV	0,99	0,67	0,00	9,9	(0,71; 0,47)
	AIC	0,99	0,70	0,00	10,7	(0,68; 0,48)
	BIC	0,94	0,45	0,00	5,3	(0,71; 0,45)
			Modèle 4			
Ignorée	GCV	0,97	0,44	0,01	9,4	(0,66; 0,55)
	AIC	0,98	0,47	0,01	9,8	(0,65; 0,56)
	BIC	0,87	0,26	0,07	6,0	(0,69; 0,53)
Incluse	GCV	0,98	0,54	0,01	11,4	(0,66; 0,54)
	AIC	0,98	0,56	0,00	11,9	(0,66; 0,55)
	BIC	0,86	0,30	0,05	6,2	(0,68; 0,53)

Tableau 5.2
Sélection pour le plan générant des corrélations fortement faussées (2^e plan). Les résultats sont résumés en fonction du taux de sélections positives (TSP), du taux de fausses découvertes (TFD), du taux de sélections correctes (TSC) et de la taille moyenne du modèle (TMM); les évaluations de la prédiction pour les modèles 1 et 2 sont fondées sur le test de la somme des carrés des résidus (SCR), et pour les modèles 3 et 4, sur le taux de prédictions positives/négatives (TPP, TPN) avec une valeur repère de 0,5.

Pondérations	Critère	TSP	TFD	TSC	TMM	Prédiction
			Modèle 1			
Ignorée	GCV	0,83	0,23	0,17	4,6	1,09
	AIC	0,97	0,49	0,04	8,6	1,10
	BIC	0,83	0,23	0,17	4,6	1,09
Incluse	GCV	0,95	0,31	0,13	5,9	1,07
	AIC	0,99	0,65	0,00	12,5	1,12
	BIC	0,95	0,30	0,14	5,9	1,07
			Modèle 2			
Ignorée	GCV	0,62	0,22	0,02	5,0	1,13
	AIC	0,88	0,45	0,01	10,3	1,14
	BIC	0,62	0,22	0,02	5,1	1,12
Incluse	GCV	0,72	0,28	0,01	6,5	1,10
	AIC	0,89	0,59	0,00	13,7	1,12
	BIC	0,72	0,27	0,01	6,5	1,10
			Modèle 3			
Ignorée	GCV	0,87	0,62	0,00	7,3	(0,66; 0,44)
	AIC	0,88	0,63	0,00	7,6	(0,65; 0,45)
	BIC	0,65	0,62	0,00	4,5	(0,68; 0,42)
Incluse	GCV	0,97	0,74	0,00	11,9	(0,70; 0,46)
	AIC	0,97	0,75	0,00	12,4	(0,68; 0,46)
	BIC	0,89	0,50	0,00	5,6	(0,70; 0,44)
			Modèle 4			
Ignorée	GCV	0,94	0,48	0,00	9,5	(0,62; 0,51)
	AIC	0,95	0,50	0,00	10,0	(0,62; 0,52)
	BIC	0,72	0,41	0,00	6,1	(0,64; 0,49)
Incluse	GCV	0,93	0,61	0,00	12,5	(0,64; 0,53)
	AIC	0,94	0,62	0,00	12,9	(0,64; 0,53)
	BIC	0,82	0,34	0,01	6,4	(0,67; 0,54)

5.3 Scénario 2 : Spécification incorrecte du modèle

Une raison bien connue de l'utilisation des poids de sondage est qu'elle protège contre la spécification incorrecte du modèle (Pfeffermann et Holmes 1985; Kott 1991) : les inférences fondées sur les estimations pondérées peuvent demeurer valides pour la population sondée, même si le modèle n'est pas correct. Afin de mieux comprendre le rôle de la pondération dans la sélection des variables, nous comparons la méthode fondée sur le critère BIC proposé aux méthodes non pondérées classiques dans la simulation où le modèle supposé est mal spécifié à partir du modèle qui génère les données. Dans de telles situations, il n'existe pas de modèle « vrai » postulé et l'objectif de la sélection des variables est de trouver un modèle optimal qui décrit bien la population finie. Nous utilisons la population pseudo-finie stratifiée de la section 5.2, mais générerons la variable réponse Y en fonction des strates. Plus précisément, nous avons généré les valeurs de Y pour les unités dans les strates (Hommes, 55 ans et plus) et (Femmes, 55 ans et plus) par

$$Y = 0,6X_6 + 0,4X_{18} + 0,4X_{20} + 0,6X_{38} + \varepsilon,$$

et les valeurs de Y pour les unités dans les strates (Hommes, moins de 55 ans) et (Femmes, moins de 55 ans) par

$$Y = 0,6X_6 + 0,4X_{18} + 0,4X_{20} + \varepsilon$$

avec $\varepsilon \sim N(0,1)$ désignant une erreur aléatoire. Autrement dit, nous supposons que la variable X_{38} est influente seulement pour les personnes de 55 ans et plus, mais non pour les personnes de moins de 55 ans. En outre, nous violons aussi le modèle 1 présumé en excluant X_6 de l'ensemble de covariables possibles, ce qui limite la situation où une caractéristique importante du plan de sondage est omise dans la modélisation. Nous tirons un échantillon EASSR stratifié de taille 500 ou 1 000 en utilisant le premier plan de sondage de la section 5.2. Puis nous testons les procédures pondérées et non pondérées de sélection de variables en nous basant sur les unités échantillonnées.

Nous résumons les résultats de la simulation au tableau 5.3 en estimant les taux de sélection de X_{18}, X_{20} et X_{38} sur la base de 1 000 répliques. Comme pour les simulations précédentes, nous incluons aussi la taille moyenne du modèle (TMM) et la SCR de test des modèles sélectionnés (c'est-à-dire la SCR moyenne fondée sur les données d'un échantillon de test de taille 200) dans le résumé. Le tableau 5.3 nous permet de constater que, si les hypothèses de modélisation sont violées, le BIC fondé sur la pseudo-vraisemblance produit encore une exactitude de prédiction assez élevée en proposant des variables pertinentes avec une forte probabilité. En revanche, le fait d'ignorer les poids de sondage entraîne une perte relative de près de 9 % sur la SCR de test à cause de l'exclusion de X_{38}. Apparemment, l'accroissement de la taille de l'échantillon contribue à l'amélioration de la qualité de l'ajustement des modèles mal spécifiés, mais l'amélioration est obtenue au prix de l'inclusion d'un plus grand nombre de variables.

Tableau 5.3
Fréquence de sélection des variables influentes dans le cas d'un modèle mal spécifié; la taille moyenne du modèle (TMM) et la somme des carrés des résidus (SCR) de test sont également présentées.

Pondération	Critère	X_{18}	X_{20}	X_{38}	TMM	CR de test
			$n = 500$			
Ignorée	GCV	0,78	0,95	0,56	5,9	1,93
	AIC	0,95	0,99	0,73	12,5	1,95
	BIC	0,83	0,97	0,60	6,6	1,93
Incluse	GCV	0,73	0,92	0,84	6,3	1,77
	AIC	0,91	0,99	0,85	12,5	1,79
	BIC	0,78	0,94	0,83	6,9	1,77
			$n = 1\,000$			
Ignorée	GCV	0,96	1,00	0,79	7,6	1,87
	AIC	0,99	1,00	0,87	13,1	1,88
	BIC	0,97	1,00	0,80	7,9	1,87
Incluse	GCV	0,93	1,00	0,94	7,6	1,71
	AIC	0,98	1,00	0,96	13,0	1,72
	BIC	0,94	1,00	0,94	7,7	1,71

5.4 Analyse des données de l'EPMCC

Afin d'illustrer l'application du critère BIC proposé, nous l'utilisons pour déterminer les comportements influant sur la santé qui ont une incidence sur le contrôle de la pression artérielle en utilisant les données de l'EPMCC de 2009. La variable réponse est BMHX_02 provenant du jeu de données de travail obtenu à partir de l'EPMCC, qui comporte deux niveaux indiquant si la pression artérielle du répondant est ou non sous contrôle, selon la dernière mesure faite par un professionnel de la santé. Nous traitons les 39 autres variables du jeu de données de travail comme des covariables possibles et notre objectif est de repérer les covariables influentes qui sont associées au contrôle de la pression artérielle. Nous construisons une régression logistique de BMHX_02 sur les covariables possibles et utilisons la procédure BIC-PVP avec la pénalité SCAD pour sélectionner les covariables influentes (les poids sont rééchelonnés par le facteur $k = 10^{-3}$). En guise d'étape préliminaire, chaque covariable est normalisée de manière à ce que les premier et deuxième moments correspondants dans l'échantillon pondéré soient égaux à 0 et à l'unité, respectivement. À titre de comparaison, les critères AIC et GCV sont également utilisés dans l'analyse.

Dans la figure 5.1, nous représentons les scores du critère en fonction du degré de parcimonie du modèle. Nous voyons que le BIC sélectionne un modèle contenant 11 covariables, tandis que les critères GCV et AIC sélectionnent le modèle contenant 24 covariables. Si l'on ignore les poids de sondage dans la procédure de sélection, des modèles avec 7 ou 21 covariables sont suggérés par le critère BIC standard ou par les critères GCV ou AIC. La distinction entre les résultats des sélections pondérées et non pondérées reflète le faussement possible de la structure de corrélation des unités échantillonnées. Ce genre de distinctions peut également s'expliquer par la spécification incorrecte du modèle pour une partie de la population de l'EPMCC (Lohr et Liu 1994). Étant donné le biais possible des méthodes non pondérées, les résultats de la sélection pondérée sont plus plausibles dans l'analyse.

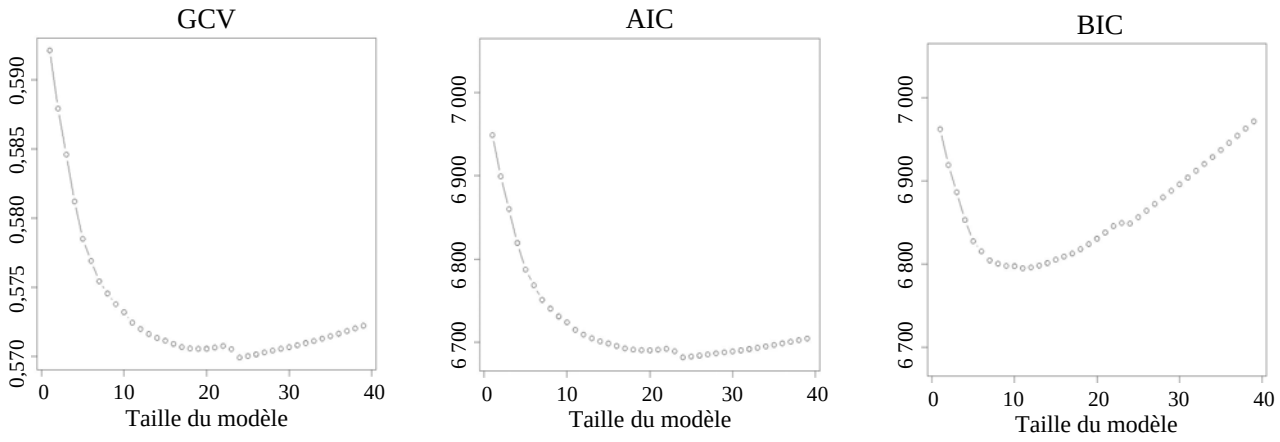

Figure 5.1 Valeurs des critères de sélection fondées sur les modèles possibles

Nous évaluons aussi les modèles sélectionnés en ce qui a trait à l'exactitude de la prédiction comme il suit. Premièrement, nous tirons 500 jeux indépendants de 5 868 échantillons bootstrap (avec remise) du

jeu de données de travail de l'EPMCC. Pour le t^{e} échantillon bootstrap d_t, $t = 1, \ldots, 500$, le poids de sondage w_i pour la i^{e} unité est ajusté selon $\tilde{w}_{ti} = v_{ti} w_i$ avec v_{ti} désignant le nombre de fois que la i^{e} unité est sélectionnée dans d_t. Puis, nous ajustons les modèles sélectionnés à chaque échantillon bootstrap (en tenant compte des poids en conséquence) et nous évaluons les taux pondérés de prédictions positives et de prédictions négatives (TPPP, TPPN) par

$$\text{TPPP} = \frac{\sum_{i \notin d_t} w_i I\left(\hat{y}_i = 1, y_i = 1\right)}{\sum_{i \notin d_t} w_i I\left(y_i = 1\right)}, \quad \text{TPPN} = \frac{\sum_{i \notin d_t} w_i I\left(\hat{y}_i = 0, y_i = 0\right)}{\sum_{i \notin d_t} w_i I\left(y_i = 0\right)},$$

où y_i et $\hat{y}_i$ désigne la i^{e} réponse dans BMHX_02 et sa valeur prédite. Nous résumons les TPPP et TPPN moyens calculés sur 500 échantillons bootstrap au tableau 5.4 pour trois valeurs repères différentes (c'est-à-dire 0,25, 0,35, 0,45).

D'après le tableau 5.4, nous constatons que les modèles sélectionnés d'après l'analyse non pondérée ont généralement un TPPP plus faible, ce qui offre un argument supplémentaire en faveur de l'utilisation des poids de sondage dans la procédure de sélection. Comparativement aux critères GCV/AIC, le critère BIC choisit le modèle ayant un TPPP légèrement plus prudent, mais un TPPN plus élevé. Néanmoins, l'écart n'est pas significatif. La taille du modèle sélectionné en appliquant le BIC est appréciablement plus faible que celle du modèle sélectionné par les critères GCV/AIC, ce qui permet d'interpréter plus facilement la relation entre la réponse BMHX_02 et les covariables.

Tableau 5.4
Exactitude de prédiction des modèles sélectionnés : (TPPP, TPPN) fondés sur différentes valeurs repères

Pondérations	Critère	≥ 0,25	≥ 0,35	≥ 0,45
Ignorée	AIC/GCV	(0,646; 0,525)	(0,460; 0,688)	(0,299; 0,811)
	BIC	(0,649; 0,513)	(0,445; 0,705)	(0,265; 0,818)
Incluse	AIC/GCV	(0,645; 0,523)	(0,488; 0,682)	(0,338; 0,790)
	BIC	(0,654; 0,532)	(0,485; 0,706)	(0,322; 0,830)

Pour évaluer la stabilité de la sélection, nous répétons la procédure de sélection pondérée fondée sur les 500 échantillons bootstrap. Au tableau 5.5, nous donnons le taux de sélection bootstrap pour les sept covariables les plus significatives en fonction de leur EMV dans le jeu de données de travail original de l'EPMCC. Les estimations des coefficients et les erreurs-types correspondantes fondées sur les échantillons bootstrap sont également incluses. D'après le tableau 5.5, nous constatons que quatre variables significatives seulement (c'est-à-dire DHHX_AGE, GENXDHMH, INHX_06, HWTDBMI) sont systématiquement sélectionnées en appliquant le critère BIC, tandis que les critères GCV/AIC ont tendance à sélectionner des variables moins fiables dans le modèle. Les résultats de la sélection fondés sur le critère BIC donnent à penser que le contrôle de la pression artérielle est fortement associé à l'âge, au poids corporel, à la santé mentale et à l'information concernant les médicaments. Nos observations

correspondent à celles de nombreuses études sur l'hypertension publiées (voir, par exemple, Gelber, Gaziano, Manson, Buring et Sesso 2007; Yan, Liu, Matthews, Daviglus, Ferguson et Kiefe 2003).

Tableau 5.5
Résultats de sélection bootstrap pour les variables significatives : (coefficient estimé, erreur-type, taux de sélections)

Variable	GCV	AIC	BIC
GEO_ON	(0,14; 0,09; 0,86)	(0,16; 0,09; 0,92)	(0,09; 0,09; 0,58)
DHHX_AGE	(-0,29; 0,09; 1,0)	(-0,32; 0,09; 1,0)	(-0,27; 0,08; 1,0)
GENXDHMH	(-0,15; 0,05; 0,99)	(-0,15; 0,05; 0,99)	(-0,14; 0,06; 0,92)
SMHXDSLT	(0,11; 0,07; 0,76)	(0,12; 0,07; 0,84)	(0,08; 0,09; 0,47)
MOHXDBPM	(-0,08; 0,07; 0,67)	(-0,09; 0,06; 0,81)	(-0,05; 0,07; 0,35)
INHX_06	(0,18; 0,06; 0,97)	(0,18; 0,06; 0,99)	(0,18; 0,07; 0,91)
HWTDBMI	(0,14; 0,06; 0,95)	(0,14; 0,06; 0,97)	(0,13; 0,06; 0,91)
Taille moyenne du modèle	23,1	27,8	10,3

6 Conclusion

Dans le présent article, nous avons abordé le problème de la sélection des variables dans l'analyse de données d'enquêtes complexes. Lorsque les unités sont sélectionnées selon un plan d'échantillonnage non proportionnel, la structure de corrélation des données reflétée par l'échantillon peut être faussée. L'intégration des poids de sondage dans le processus de sélection protège contre l'obtention de résultats de sélection biaisés. Dans cet esprit, nous avons dérivé un critère BIC pondéré par les poids de sondage fondé sur la pseudo-vraisemblance et proposé en outre une procédure efficace (pseudo-vraisemblance pénalisée) pour son application. Sous certaines conditions de régularité, nous avons montré que notre critère repère systématiquement les variables influentes sous un cadre de randomisation conjoint modèle-plan. Les résultats acceptables de la méthode proposée ont été confirmés par des études numériques.

Remerciements

Les auteurs remercient le rédacteur associé et les deux examinateurs anonymes de leurs commentaires judicieux et de leurs suggestions utiles. Les auteurs remercient le professeur J.N.K. Rao de l'Université Carleton de ses commentaires constructifs concernant un manuscrit antérieur. Les présents travaux ont été financés par Statistique Canada et par le MITACS.

Annexe

Tableau A.1
Variables pour l'analyse des données de l'EPMCC avec ajustement de la non-réponse : A : affectée à d'autres catégories; S : supprimée des données; M : imputée par les valeurs moyennes; NA : non ajustée pour la non-réponse

	Variable	Description	Niveaux	Manquante	Ajustement
1	BMHX_02	État de contrôle de la pression artérielle	2	1,6 %	S
2	GEO_QB	Provinces groupées par région – Québec	2	- -	NA
3	GEO_ON	Provinces groupées par région – Ontario	2	- -	NA
4	GEO_BC	Provinces groupées par région – Colombie-Britannique	2	- -	NA
5	GEO_PR	Provinces groupées par région – Prairies	2	- -	NA
6	DHHX_AGE	Âge	Cont.	- -	NA
7	DHHX_SEX	Sexe	2	- -	NA
8	GENXDHMH	Santé mentale perçue	2	0,2 %	A
9	CNHX_05	Hypertension – âge au diagnostic	Cont.	2,7 %	S
10	MEHX_02	Nombre de médicaments pris	Cont.	0,3 %	M
11	MEHX_03	Nombre de fois par jour que les médicaments sont pris	Cont.	0,1 %	M
12	MEHXGMED	Nombre de médicaments pour l'hypertension	Cont.	2,0 %	M
13	MEHX_06	Nombre de fois par jour que des médicaments pour l'hypertension sont pris	Cont.	1,0 %	M
14	MEHXDMCO	Respect des prescriptions concernant la prise de médicaments – global	2	0,2 %	A
15	HUHXDHP	A consulté un médecin de famille au sujet de l'hypertension	2	0,1 %	A
16	SMHX_11A	A fumé à n'importe quel moment depuis le diagnostic	2	0,1 %	A
17	SMHX_13A	A bu de l'alcool depuis le diagnostic	2	0,2 %	A
18	SMHXDSLT	Apport quotidien en sel	2	0,2 %	A
19	SMHXDFDC	Aliments de régime	2	0,1 %	A
20	SMHXDPAC	Exercice/activité physique	2	0,1 %	A
21	SMHXDBW	Contrôle du poids corporel	2	0,2 %	A
22	MOHXDBPM	Autosurveillance de la pression artérielle	2	0,3 %	A
23	MOHX_02	Usage correct de l'appareil de mesure de la pression artérielle	2	0,5 %	A
24	INHX_01A	Information fournie par le médecin de famille	2	2,4 %	A
25	INHX_01F	Information fournie par un membre de la famille/ami	2	2,4 %	A
26	INHX_02A	Information tirée de livres, brochures, dépliants	2	1,5 %	A
27	INHX_02C	Information tirée de la notice figurant dans l'emballage du médicament	2	1,5 %	A
28	INHX_02G	Information provenant des médias	2	1,5 %	A
29	INHX_02H	Information provenant d'Internet	2	1,5 %	A
30	INHX_04	Information reçue – effet émotionnel de l'hypertension	2	0,8 %	A
31	INHX_06	Information reçue – usage correct des médicaments	2	0,6 %	A
32	INHX_07	Information reçue – information supplémentaire	2	0,9 %	A
33	CPGFGAM	Activités de jeux de hasard	2	0,5 %	A
34	DHHDECF	Type de logement	2	0,2 %	A
35	EDUDH04	Niveau d'études le plus élevé dans le ménage	2	3,4 %	A
36	FGVCTOT	Consommation quotidienne – fruits et légumes	2	5,2 %	A
37	GEODUR2	Régions urbaines et rurales	2	- -	NA
38	HWTDBMI	Indice de masse corporelle (IMC), données autodéclarées	Cont.	2,1 %	M
39	INCDRPR	Revenu du ménage – niveau provincial	10	9,6 %	A
40	SACDTOT	Nombre total d'heures – activités sédentaires	Cont.	1,5 %	M

Tableau A.2
Variables influentes et variables du plan de sondage dans les simulations : ∗ - variable influant sur la réponse; • - variable du plan affectant les probabilités d'échantillonnage dans le premier plan; ◊ - variable du plan affectant les probabilités d'échantillonnage dans le deuxième plan.

	Variable	Modèle 1	Modèle 2	Modèle 3	Modèle 4
6	DHHX_AGE	* • ◊	* • ◊	• ◊	• ◊
7	DHHX_SEX	• ◊	• ◊	* • ◊	* • ◊
8	GENXDHMH			* ◊	* ◊
10	MEHX_02	*	*		
18	SMHXDSLT	* ◊	* ◊		
22	MOHXDBPM	* ◊	* ◊		
26	INHX_02A			* ◊	* ◊
28	INHX_02G				*
30	INHX_04		*		
34	DHHDECF		*		
36	FVCGTOT				*

Bibliographie

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Dans le *2nd International Symposium on Information Theory*, (Éds., B.N. Petrox et F. Caski), 267-281.

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

Binder, D., et Roberts, G. (2003). *Analysis of Survey Data*, Chapter: Design-based and model-based methods for estimating model parameters. Wiley Series in Survey Methodology, Chichester.

Craven, P., et Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 377-403.

Fan, J., et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Frank, I.E., et Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.

Fuller, W.A. (2009). *Sampling Statistics*. Wiley, Hoboken.

Gelber, R.P., Gaziano, J.M., Manson, J.E., Buring, J.E. et Sesso, H.D. (2007). A prospective study of body mass index and the risk of developing hypertension in men. *American Journal of Hypertension*, 20, 370-377.

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.

Kalton, G. (1983). Models in the practice of survey sampling. *Revue Internationale de Statistique*, 51, 175-188.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc.

Kott, P.S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, 45, 107-112.

Liu, X., Wang, L. et Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21, 1225-1248.

Lohr, S.L., et Liu, J. (1994). A comparison of weighted and unweighted analyses in the NCVS. *Journal of Quantitative Criminology*, 10, 343-360.

Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, 15, 661-675.

Molina, E.A., et Skinner, C.J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis*, 13, 395-405.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Pfeffermann, D., et Holmes, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, Séries A, 148, 268-278.

Rahiala, M., et Teräsvirta, T. (1993). Business survey data in forecasting the output of Swedish and Finnish metal and engineering industries: A Kalman filter approach. *Journal of Forecasting*, 12, 255-271.

Royall, M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.

She, Y. (2011). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, in press.

Skinner, C. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics*, manuscript.

Statistics Canada (2009). Enquête sur les personnes ayant une maladie chronique au Canada – Guide de l'utilisateur 2009. Documentation supplémentaire.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (avec discussion). *Journal of the Royal Statistical Society,* Séries B, 39, 111-147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Séries B, 58, 267-288.

Wang, H., Li, R. et Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Wolfson, W.G. (2004). Analysis of labour force survey data for the information technology occupations 2000-2003. *Report for the Software Human Resource Council*, WGW Services Ltd., Ottawa, Ontario.

Xie, B., Pan, W. et Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64, 921-930.

Xu, C., et Chen, J. (2012). Technical supplement to "Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data". Disponible auprès du premier auteur.

Yan, L.L., Liu, K., Matthews, K.A., Daviglus, M., Ferguson, T.F. et Kiefe, C.I. (2003). Psychosocial factors and risk of hypertension: The coronary artery risk development in young adults (CARDIA) study. *The Journal of the American Medical Association*, 290, 2138-2148.