Nº 12-001-XIF au catalogue

Techniques d'enquête

Juin 2007

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1-800-263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web à www.statcan.ca.

Service national de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Renseignements concernant le Programme des services de dépôt	1-800-700-1033
Télécopieur pour le Programme des services de dépôt	1-800-889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder ou commander le produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Publications.

Ce produit nº 12-001-XPF au catalogue est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.ca
- Poste Statistique Canada
  Division des finances
  Immeuble R.-H.-Coats, 6e étage
  100, promenade Tunney's Pasture
  Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

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 qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. 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 dans le site www.statcan.ca sous À propos de nous > Offrir des services aux Canadiens.

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2007

Publication autorisée par le ministre responsable de Statistique Canada

Juin 2007

Nº 12-001-XIF au catalogue, vol. 33, nº 1
ISSN 1712-5685

No 12-001-XPB au catalogue, vol. 33, nº 1
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in english upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

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

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 20 $ CA (10 $ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.ca.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 33, numéro 1, juin 2007

Table des matières

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 1-2
Statistique Canada, N° 12-001 au catalogue

Dans ce numéro

Ce numéro de *Techniques d'enquête* comprend des articles portant sur divers sujets méthodologiques tels que la modélisation et l'estimation, la pondération et l'estimation de la variance, la non-réponse et l'échantillonnage.

Dans le premier article, Skinner et Vieira étudient l'effet de l'échantillonnage en grappes sur l'estimation de la variance dans les enquêtes longitudinales. Ils présentent des arguments théoriques et des données empiriques démontrant les effets de la non-prise en compte de la mise en grappes dans les analyses longitudinales et constatent qu'en général, ces effets ont tendance à être plus importants que dans le cas des analyses transversales correspondantes. Ils comparent aussi les méthodes basées sur le plan de sondage classiques pour tenir compte de la mise en grappes dans l'estimation de la variance à une approche de modélisation multiniveaux.

Kovačević et Roberts comparent trois modèles conçus pour l'analyse des périodes multiples émanant de données recueillies au moyen d'enquêtes longitudinales à plan de sondage complexe pouvant comprendre une stratification et une mise en grappes. Ces modèles sont des variantes du modèle à risques proportionnels de Cox du même genre que celles proposés dans la littérature par Lin et Wei (1989), Binder (1992) et Lin (2000). Ces trois modèles sont comparés à l'aide de données provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée par Statistique Canada. L'article fournit de nouveaux éclaircissements concernant l'ajustement des modèles de Cox à des données d'enquête représentant plusieurs périodes par individu, situation qui survient assez fréquemment. L'article illustre aussi certains défis que pose l'ajustement des modèles de Cox aux données d'enquête.

Elliott présente dans son article un moyen de réaliser un compromis entre la variance élevée due à des poids de valeur extrême et le biais éventuel à l'aide d'une méthode bayésienne de réduction des poids dans des modèles linéaires généralisés. Le compromis est obtenu en utilisant un modèle hiérarchique bayésien stratifié dans lequel les strates sont déterminées par les probabilités d'inclusion ou par les poids de sondage. Il illustre et évalue l'approche à l'aide de simulations fondées sur des modèles de régression linéaire et de régression logistique, ainsi qu'une application portant sur des données provenant de la Partners for Child Passenger Surveillance Survey.

L'article de Breidt, Opsomer, Johnson et Ranalli explore l'utilisation de méthodes semiparamétriques pour l'estimation des moyennes de population. Dans l'estimation semiparamétrique, il est supposé que certaines variables sont reliées linéairement à la variable d'intérêt, tandis que d'autres peuvent l'être de façon plus compliquée, non spécifiée. Les auteurs étudient théoriquement les propriétés sous le plan de sondage des estimateurs résultants. En particulier, ils montrent la convergence sous le plan et la normalité asymptotique de leur estimateur. Puis, ils appliquent leur méthode à des données provenant d'une enquête sur les lacs du Nord-Est des États-Unis.

Tanguay et Lavallée abordent la question de l'estimation de la dépréciation des actifs à l'aide d'une base de données sur les ratios de prix. Dans leur article, le problème est dû au fait que les ratios ne proviennent pas d'un échantillon aléatoire tiré de la population de ratios. Les auteurs soutiennent que la distribution des ratios devrait converger vers une loi uniforme et proposent un scénario de pondération qui rendra la fonction de répartition empirique pondérée approximativement uniforme. Ils illustrent la méthode proposée à l'aide de données sur la dépréciation des automobiles.

Steel et Clark présentent une comparaison empirique et théorique des pondérations produites par la régression généralisée au niveau de la personne et des pondérations intégrées au niveau du ménage dans le cas d'un échantillon aléatoire simple de ménages à partir duquel tous les membres de chaque ménage sont sélectionnés. Ils concluent que l'utilisation de la pondération intégrée est associée à une perte faible, voire nulle, d'efficacité.

Dans son article, Saigo propose une méthode bootstrap d'estimation de la variance pour les plans de sondage à deux phases avec fractions de sondage élevées. La méthode s'appuie sur les techniques du bootstrap courantes, mais comporte un ajustement des valeurs des variables auxiliaires pour les unités qui sont sélectionnées à la première phase seulement. La méthode proposée est illustrée à l'aide de plusieurs estimateurs utilisés couramment, comme l'estimateur par le ratio et les estimateurs de la fonction de répartition et des quantiles. Les résultats d'une étude par simulation comparant la méthode proposée à plusieurs autres sont présentés.

L'article de Longford traite du problème de l'estimation de l'EQM pour les estimations pour petits domaines. L'auteur obtient un estimateur composite de l'EQM des moyennes de petits domaines en combinant un estimateur de la variance sous un modèle et un estimateur naïf de l'EQM. Le coefficient qui combine les deux estimateurs minimise l'EQM prévue de l'estimateur composite de l'EQM résultant. L'estimateur proposé est comparé aux estimateurs existants dans plusieurs études par simulation.

Shao considère le problème de l'imputation pour remplacer les valeurs manquantes en cas de non-réponse non ignorable. Dans la situation où la non-réponse dépend d'un effet aléatoire au niveau de la grappe, il montre que l'estimateur imputé par la moyenne est biaisé, à moins que l'on utilise la moyenne de la grappe. Pour l'estimation de la variance, il fournit une méthode d'estimation de la variance par le jackknife pour l'estimateur proposé. Il compare ce dernier à l'estimateur imputé par la moyenne à l'aide d'une étude par simulation.

Dans le dernier article du numéro, Tiwari, Nigam et Pant utilisent le concept de plan d'échantillonnage proportionnel à la taille le plus proche pour obtenir un plan d'échantillonnage contrôlé optimal assurant que les probabilités de sélection des échantillons non privilégiés soient nulles. Le plan d'échantillonnage contrôlé optimal est obtenu en combinant un plan d'échantillonnage avec probabilité d'inclusion proportionnelle à la taille et des techniques de programmation quadratique pour assurer que les échantillons non privilégiés aient une probabilité de sélection nulle. Les auteurs illustrent leur méthode à l'aide de plusieurs exemples.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 3-13
Statistique Canada, N° 12-001 au catalogue

Estimation de la variance dans l'analyse de données d'enquête longitudinale en grappes

Chris Skinner et Marcel de Toledo Vieira [1]

Résumé

Nous étudions l'effet de l'échantillonnage en grappes sur les erreurs-types dans l'analyse des données d'enquête longitudinale. Nous considérons une classe de modèles de régression pour données longitudinales d'usage très répandu et une classe standard d'estimateurs ponctuels de type moindres carrés généralisés. Nous soutenons théoriquement que l'effet de la non-prise en compte de la mise en grappes dans l'estimation de l'erreur-type a tendance à augmenter avec le nombre de vagues de l'enquête incluses dans l'analyse, sous certains scénarios de mise en grappes raisonnables pour de nombreuses enquêtes sociales. La conséquence est qu'en général, il est au moins aussi important de tenir compte de la mise en grappes dans le calcul des erreurs-types dans le cas des analyses longitudinales que dans celui des analyses transversales. Nous illustrons cet argument théorique à l'aide des résultats empiriques d'une analyse par régression de données longitudinales sur les attitudes à l'égard des rôles de l'homme et de la femme provenant de l'enquête par panel menée auprès des ménages au Royaume-Uni (*British Household Panel Survey*). Nous comparons aussi deux approches d'estimation de la variance dans l'analyse des données d'enquête longitudinale, à savoir une approche par plan de sondage basée sur la linéarisation et une approche par modélisation multiniveaux. Nous concluons que l'effet de la mise en grappes peut être sérieusement sous-estimé si l'on se contente, en vue d'en tenir compte, d'inclure un effet aléatoire additif pour représenter la mise en grappes dans un modèle multiniveaux.

Mots clés : Mise en grappes; effet de plan; effet d'erreur de spécification; modèle multiniveaux.

1. Introduction

Il est bien connu qu'il importe de tenir compte de la mise en grappes de l'échantillon lors de l'estimation des erreurs-types dans l'analyse des données d'enquête. Sinon, les estimateurs des erreurs-types risquent d'être gravement biaisés. Dans le présent article, nous étudions l'effet de la mise en grappes dans l'analyse de données d'enquête longitudinale par régression et le comparons à celui observé dans l'analyse transversale correspondante. Kish et Frankel (1974) ont présenté des travaux empiriques donnant à penser que l'effet d'un plan de sondage complexe sur la variance diminue lorsque les statistiques analytiques deviennent plus complexes et l'on pourrait donc conjecturer que l'effet est susceptible de diminuer aussi dans le cas des analyses longitudinales. Nous soutenons qu'en fait, l'effet de la mise en grappes tend parfois à être plus important dans les analyses longitudinales, du moins pour plusieurs types courants d'analyse et certaines conditions pratiques courantes. Une explication intuitive serait que certaines formes courantes d'analyse longitudinale de données individuelles d'enquête « regroupent » ces données au fil du temps et permettent d'« extraire » de l'estimation des coefficients de régression une grande part de la variation temporelle « aléatoire » présente dans les réponses individuelles. En revanche, il se peut qu'il soit impossible d'extraire autant de variation des effets de la mise en grappes, puisque cette dernière, représentant la géographie par exemple, a souvent tendance à produire des effets plus stables que les mesures répétées de comportements individuels. Par conséquent, l'importance relative de la mise en grappes dans les erreurs-types peut croître à mesure qu'augmente le nombre de vagues de sondage incluses dans l'analyse.

En plus de considérer l'effet de la mise en grappes sur l'estimation de la variance, nous étudions la question de savoir comment entreprendre l'estimation proprement dite de la variance. Il est naturel pour de nombreux analystes de représenter la mise en grappes à l'aide de modèles multiniveaux (Goldstein 2003, chapitre 9; Renard et Molenberghs 2002) et nous comparerons les méthodes d'estimation de la variance fondées sur ce genre de modèle à celles fondées sur le plan de sondage dans le cas de l'échantillonnage en grappes.

Il existe une abondante littérature sur les méthodes permettant de tenir compte des plans d'échantillonnage complexes dans l'analyse de données d'enquête par la régression. Voir, par exemple, Kish et Frankel (1974), Fuller (1975), Binder (1983), Skinner, Holt et Smith (1989), ainsi que Chambers et Skinner (2003). Nous ne nous intéressons ici qu'aux analyses par la régression « agrégée » (Skinner et coll. 1989), où les coefficients de régression au « niveau de la population » sont les paramètres d'intérêt, où les estimations appropriées de ces coefficients peuvent être obtenues en adaptant des méthodes basées sur un modèle standard avec l'utilisation de poids de sondage et où les variances de ces coefficients de régression estimés peuvent être estimées par des méthodes de linéarisation (Kish et Frankel 1974; Fuller 1975). Dans le présent article, nous

1. Chris Skinner, Université de Southampton, Royaume-Uni; Marcel de Toledo Vieira, Université fédérale de Juiz de Fora, Brésil.

étendons les travaux de ces auteurs à la situation où l'on obtient des observations d'enquête longitudinale basées sur un échantillon initial tiré selon un plan d'échantillonnage complexe, en nous concentrant de nouveau sur le cas d'un plan de sondage en grappes. Nous considérons pour ces données longitudinales une classe standard de modèles de régression linéaire qui est décrite dans la littérature biostatistique (par exemple, Diggle, Heagerty, Liang et Zeger 2002), la littérature sur la modélisation multiniveaux (par exemple, Goldstein 2003) et la littérature économétrique (par exemple, Baltagi 2001). Nous considérons une classe établie d'estimateurs ponctuels de type moindres carrés généralisés modifiés par les poids de sondage. Pour certaines applications de ces méthodes à des données d'enquête, voir Lavange, Koch et Schwartz (2001), ainsi que Lavange, Stearns, Lafata, Koch et Shah (1996).

Nous mesurons l'effet d'un plan d'échantillonnage complexe sur l'estimation de la variance par l'« effet d'erreur de spécification », dénoté meff (pour *mispecification effect* en anglais) (Skinner 1989a), qui est la variance de l'estimateur ponctuel d'intérêt sous le plan d'échantillonnage réel divisée par l'espérance d'un estimateur spécifié de la variance. Il s'agit d'une mesure du biais relatif de l'estimateur spécifié de la variance. Si celui-ci est sans biais, le meff sera égal à un. Ce concept est étroitement lié à celui de l'« effet de plan » ou deff de Kish (1965), défini comme étant la variance de l'estimateur ponctuel sous le plan donné divisée par sa variance sous échantillonnage aléatoire simple avec la même taille d'échantillon, notion qui est plus en rapport avec le choix du plan qu'avec celui de l'estimateur de l'erreur-type.

Nous illustrerons nos arguments théoriques grâce à des analyses de données provenant de la British Household Panel Survey (BHPS) sur les attitudes à l'égard des rôles de l'homme et de la femme, où les principales unités d'intérêt analytique sont les femmes prises individuellement et où les grappes correspondent aux secteurs postaux (*postcode sectors* en anglais) utilisés comme unités primaires d'échantillonnage lors de la sélection de l'échantillon de première vague à partir d'un registre d'adresses.

Le cadre, y compris les modèles et les méthodes d'estimation, est décrit à la section 2. Les propriétés théoriques des méthodes d'estimation de la variance sont exposées à la section 3. La section 4 illustre numériquement ces propriétés, à l'aide d'une analyse des données de la BHPS. Enfin, certaines conclusions sont présentées à la section 5.

2. Modèle de régression, données et méthodes d'inférence

Considérons une population finie $U = \{1, ..., N\}$ de N unités, que nous supposons fixe au cours d'une série d'éditions, ou vagues, $t = 1, ..., T$ d'une enquête. Nous parlerons d'individus pour les unités, quoique notre discussion soit d'application plus générale. Soit y_{it} la valeur d'une variable résultat pour l'individu $i \in U$ à la vague t et soit $y_i = (y_{i1}, ..., y_{iT})'$ le vecteur de mesures répétées. Soit x_{it} un vecteur $1 \times q$ correspondant de valeurs des covariables pour l'individu i à la vague t et soit $x_i = (x_{i1}', ..., x_{iT}')'$. Nous supposons que le modèle linéaire qui suit est vérifié pour l'espérance de y_i sachant $(x_1, ..., x_N)$:

$$E(y_i) = x_i\beta, \qquad (1)$$

où β est un vecteur $q \times 1$ de coefficients de régression et l'espérance est calculée sous le modèle. Nous supposons que β est la cible de l'inférence, autrement dit que les coefficients de régression sont les paramètres qui intéressent principalement l'analyste. Nous examinerons également d'autres caractéristiques du modèle, comme la matrice de covariance de y_i, mais nous supposons qu'elles sont d'un intérêt secondaire pour l'analyste.

Les données disponibles pour l'inférence au sujet de β proviennent d'une enquête longitudinale dans laquelle les valeurs de y_{it} et de x_{it} sont observées lors de chaque édition (vague) $t = 1, ..., T$ pour les individus i dans un échantillon, s, tiré à partir de U à la vague 1 selon un plan d'échantillonnage spécifié. Pour simplifier, nous supposons ici qu'il n'y a pas de non-réponse, mais nous examinons cette possibilité à la section 4.

Afin de formuler un estimateur ponctuel de β, nous étendons la spécification de (1) au modèle « de travail » suivant :

$$y_{it} = x_{it}\beta + u_i + v_{it}, \qquad (2)$$

où u_i et v_{it} sont des effets aléatoires indépendants de moyenne nulle et de variances $\sigma_u^2 = \rho\sigma^2$ et $\sigma_v^2 = (1 - \rho)\sigma^2$ respectivement, sachant $(x_1, ..., x_N)$. Ce modèle peut être appelé un modèle de corrélation uniforme (Diggle et coll. 2002, page 55) ou un modèle à deux niveaux (Goldstein 2003). Le paramètre ρ est la corrélation intra-individu.

L'estimateur ponctuel de base de β que nous considérons est

$$\hat{\beta} = \left(\sum_{i \in s} w_i\, x_i' V^{-1} x_i \right)^{-1} \sum_{i \in s} w_i\, x_i' V^{-1}\, y_i, \qquad (3)$$

où w_i est un poids de sondage et V est une matrice de covariance $T \times T$ estimée de y_i sous le modèle de travail (2), c'est-à-dire qu'elle contient des éléments diagonaux $\hat{\sigma}^2$ et des éléments non diagonaux $\hat{\rho}\hat{\sigma}^2$, où $(\hat{\rho}, \hat{\sigma}^2)$ est un estimateur de (ρ, σ^2). (Notons qu'en fait, $\hat{\sigma}^2$ s'annule dans (3) et que σ^2 ne doit donc pas être estimé pour $\hat{\beta}$). En l'absence des termes de pondération et des considérations

relatives à l'enquête, la forme de $\hat{\beta}$ est motivée par l'approche des équations d'estimation généralisées (EEG) de Liang et Zeger (1986). L'idée ici est que $\hat{\beta}$, en tant qu'estimateur par les moindres carrés généralisés de β, serait entièrement efficace si le modèle de travail (2) tenait. Cependant, $\hat{\beta}$ demeure convergent sous (1) et l'on peut encore s'attendre à ce qu'il combine l'information intra-individu et inter-individus de manière raisonnablement efficace, même si le modèle de travail de la structure de l'erreur n'est pas exactement vérifié.

Les poids de sondage sont inclus dans (3) suivant l'approche de la pseudo-vraisemblance (Skinner 1989b) pour s'assurer que $\hat{\beta}$ est approximativement sans biais pour β sous le modèle et sous le plan, à condition que (1) tienne.

Il existe plusieurs moyens d'estimer ρ. Dans d'autres conditions que celles d'une enquête, Liang et Zeger (1986) donnent une approche itérative avec alternance entre les estimations de β et ρ. Shah, Barnwell et Bieler (1997) décrivent comment les poids de sondage peuvent être intégrés dans cette approche et implémentent cette méthode dans la procédure REGRESS du logiciel SUDAAN. Par défaut, SUDAAN n'implémente qu'une seule étape de cette méthode itérative et, dans des conditions autres qu'un sondage, Lipsitz, Fitzmaurice, Orav et Laird (1994) concluent qu'il y a peu à perdre à n'utiliser qu'une seule étape. Pour le modèle de travail (2), l'approche de Liang et Zeger (1986) d'estimation de β et ρ est presque identique à l'approche d'estimation par les moindres carrés généralisés itérés (MCGI) de Goldstein (1986). Les deux méthodes comportent une itération entre les estimations de β et de ρ, et toutes deux estiment β par les MCGI, sachant l'estimation courante de ρ. La seule légère différence tient à la méthode utilisée pour estimer ρ. Pfeffermann, Skinner, Holmes, Goldstein et Rasbash (1998) montrent comment intégrer les poids de sondage dans l'approche des MCGI et l'on peut s'attendre à ce que leur méthode produise des estimations de ρ fort semblables à celles de la procédure REGRESS de SUDAAN. Pour le besoin du présent article, la forme précise de $\hat{\rho}$ ne sera pas critique et nous pouvons considérer $\hat{\beta}$ comme étant un estimateur par EEG pondérées ou un estimateur par MCGI pondérés.

Nous nous penchons maintenant sur l'estimation de la matrice de covariance de $\hat{\beta}$ sous le plan d'échantillonnage complexe. Nous supposerons, de manière générale, qu'un plan d'échantillonnage stratifié à plusieurs degrés est utilisé. Nous considérons deux grandes approches d'estimation de la variance.

Notre première approche est la méthode classique de linéarisation (Skinner 1989b page 78). L'estimateur de la matrice de covariance de $\hat{\beta}$ est

$$v(\hat{\beta}) = \left[\sum_{i\in s} w_i\, x_i' V^{-1} x_i\right]^{-1} \times \left[\sum_h n_h/(n_h-1)\sum_a (z_{ha}-\bar{z}_h)(z_{ha}-\bar{z}_h)'\right] \times \left[\sum_{i\in s} w_i\, x_i' V^{-1} x_i\right]^{-1} \quad (4)$$

où h dénote la strate, a dénote l'unité primaire d'échantillonnage (UPE), n_h est le nombre d'UPE dans la strate h, $z_{ha} = \sum_i w_i\, x_i' V^{-1} e_i$, $\bar{z}_h = \sum_a z_{ha}/n_h$ et $e_i = y_i - x_i\hat{\beta}$. Des estimateurs similaires sont considérés par Shah et coll. (1997, pages 8-9) et par Lavange et coll. (2001). Si l'on ignore les pondérations, le plan d'échantillonnage et l'écart entre $n/(n-1)$ et 1, cet estimateur se réduit à l'estimateur de variance « robuste » présenté par Liang et Zeger (1986).

Notre deuxième approche est basée plus directement sur un modèle. Nous commençons par étendre le modèle afin de représenter la population complexe qui sous-tend le plan d'échantillonnage, puis nous procédons à l'inférence sous le modèle étendu. Nous considérons uniquement le cas de l'échantillonnage à deux degrés à partir d'une population en grappes, où le modèle à deux niveaux donné par (2) est étendu au modèle à trois niveaux (Goldstein 2003):

$$y_{ait} = x_{ait}\beta + \eta_a + u_{ai} + v_{ait}. \quad (5)$$

L'indice a supplémentaire dénote la grappe et le terme aléatoire supplémentaire η_a de variance σ_η^2 représente l'effet de grappe (que l'on suppose indépendant de u_{ai} et v_{ait}). Soit σ_u^2 et σ_v^2 les variances de u_{ai} et v_{ait}, respectivement. L'inférence a alors lieu en utilisant les MCGI, qui peuvent être pondérés pour éviter le biais de sélection. Cette approche génère directement une matrice de covariance estimée de l'estimateur de β. Il convient toutefois de souligner que l'estimateur de β dérivé en utilisant les MCGI pondérés sous le modèle (5) peut différer légèrement de l'estimateur (3) (quoique, pour des estimations données des trois composantes de la variance dans (5), il sera le même qu'un estimateur par EEG pondérées avec une matrice de covariance de travail basée sur ce modèle à trois niveaux). Néanmoins, d'après notre expérience des applications aux enquêtes sociales, comme il est décrit à la section 4, et d'après la théorie (Scott et Holt 1982), l'écart entre ces deux estimateurs ponctuels de rechange est souvent négligeable.

Nous disposons de deux approches générales pour dériver les estimateurs de la variance en (5). Premièrement, en ne tenant pas compte des poids de sondage, nous pouvons employer la méthode des MCGI standard (Goldstein 1986), en supposant que chaque effet aléatoire suit une loi normale. Deuxièmement, pour éviter l'hypothèse d'effets aléatoires

homoscédastiques normaux, nous pouvons employer une méthode d'estimation « robuste » de la variance (Goldstein 2003, page 80). Cette approche est étendue à l'utilisation des poids de sondage dans Pfeffermann et coll. (1998). À part la stratification, l'estimateur de la variance est identique à l'estimateur par linéarisation (4) pour une valeur donnée de $\hat{\rho}$.

3. Propriétés des estimateurs de variance

À la présente section, nous considérons les propriétés des estimateurs de la matrice de covariance de $\hat{\beta}$ décrite à la section précédente. Nous examinons d'abord l'estimateur par linéarisation $v(\hat{\beta})$ donné par (4).

La convergence de $v(\hat{\beta})$ pour la matrice de covariance de $\hat{\beta}$ suit les arguments établis dans un cadre asymptotique approprié (par exemple, Fuller 1975; Binder 1983). La seule caractéristique non standard est la présence de V^{-1} dans $\hat{\beta}$ et $v(\hat{\beta})$ et la dépendance de V à l'égard de $\hat{\rho}$. En fait, dans les grands échantillons, la matrice de covariance de $\hat{\beta}$ ne dépend de $\hat{\rho}$ que par la voie de sa valeur limite ρ^* (dans un cadre asymptotique donné). Pour le voir, écrivons $\hat{\beta} - \beta = (\sum_s u_i)^{-1} \sum_s \tilde{z}_i$, où $u_i = w_i x_i' V^{-1} x_i$, $\tilde{z}_i = w_i x_i' V^{-1} \tilde{e}_i$ et $\tilde{e}_i = y_i - x_i \beta$. Notons que, sous des conditions de régularité faibles (Fuller et Battese 1973, corollaire 3), la distribution asymptotique de $\hat{\beta} - \beta$ est la même que celle de $\beta^* - \beta = (\sum_s u_i^*)^{-1} \sum_s z_i^*$, où $u_i^* = w_i x_i' V^{*-1} x_i$, $z_i^* = w_i x_i' V^{*-1} \tilde{e}_i$ et V^* prend la même forme que V avec $\hat{\rho}$ remplacé par $\rho^* = p\lim(\hat{\rho})$, la limite de probabilité de $\hat{\rho}$ dans le cadre asymptotique. En écrivant $\bar{z}^* = \sum_s z_i^* / n$ et $\bar{U} = p\lim(\sum_s u_i^* / n)$, nous pouvons donc approximer la matrice de covariance de $\hat{\beta}$ asymptotiquement par $\text{var}(\hat{\beta}) \approx \bar{U}^{-1} \text{var}(\bar{z}^*) \bar{U}^{-1}$. Si le modèle de travail (2) tient, alors $\rho^* = \rho$ et cette matrice de covariance sera la même pour toute méthode convergente d'estimation de ρ. Même si le modèle de travail ne tient pas, $v(\hat{\beta})$ sera convergent pour $\bar{U}^{-1} \text{var}(\bar{z}^*) \bar{U}^{-1}$ dans le genre de cadre asymptotique considéré par Fuller (1975) et Binder (1983), ainsi que sous les genres de conditions de régularité que ces auteurs et Fuller et Battese (1973) établissent.

Ensuite, nous examinons l'effet sur la méthode de linéarisation de la non-prise en compte du plan d'échantillonnage complexe. Nous dénotons par $v_0(\hat{\beta})$ l'estimateur par linéarisation obtenu d'après l'expression (4) en ignorant le plan, c'est-à-dire en supposant qu'il n'existe qu'une seule strate contenant des UPE identiques aux individus, de sorte que $n_h = n$ est la taille globale d'échantillon et que z_{ha} est remplacé par $z_i = w_i x_i' V^{-1} e_i$. Nous nous intéresserons au biais de $v_0(\hat{\beta})$ quand le plan de sondage est, en fait, complexe. Soit $\hat{\beta}_k$ le k^{e} élément de $\hat{\beta}$ et soit $v_0(\hat{\beta}_k)$ le k^{e} élément de $v_0(\hat{\beta})$. Alors, à l'exemple de Skinner (1989a, page 24), nous mesurons le biais relatif de l'estimateur de la variance « incorrectement spécifiée » $v_0(\hat{\beta}_k)$ en tant qu'estimateur de $\text{var}(\hat{\beta}_k)$ par l'*effet d'erreur de spécification*, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \text{var}(\hat{\beta}_k) / E[v_0(\hat{\beta}_k)]$. Puisque $v(\hat{\beta}_k)$ est un estimateur convergent de $\text{var}(\hat{\beta}_k)$, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ peut être estimé par $v(\hat{\beta}_k) / v_0(\hat{\beta}_k)$ et est étroitement relié à la notion d'effet de plan.

Pour étudier la nature de $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$, nous commençons par écrire :

$$v_0(\hat{\beta}) = \left(\sum_s u_i\right)^{-1} [n/(n-1)] \times \left[\sum_s (z_i - \bar{z})(z_i - \bar{z})'\right] \left(\sum_s u_i\right)^{-1}, \qquad (6)$$

où $\bar{z} = \sum_s z_i / n$. Alors, en tant qu'approximation asymptotique, nous avons $E[v_0(\hat{\beta})] \approx \bar{U}^{-1}[n^{-1} S_z^*] \bar{U}^{-1}$, où S_z^* est la limite de probabilité de la matrice de covariance en population finie de z_i^*. Étant donné que le numérateur de $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ peut être approximé par $\bar{U}^{-1} \text{var}(\bar{z}^*) \bar{U}^{-1}$, nous pouvons écrire :

$$\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \frac{(\bar{U}^{-1})_k \text{var}(\bar{z}^*) (\bar{U}^{-1})_k'}{(\bar{U}^{-1})_k [n^{-1} S_z^*] (\bar{U}^{-1})_k'}, \qquad (7)$$

où $(\bar{U}^{-1})_k$ est la k^{e} ligne de $\bar{U}^{-1}$. Si $q = 1$, cette expression se simplifie en :

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = \text{var}(\bar{z}^*) / [n^{-1} S_z^*]. \qquad (8)$$

Nous pouvons explorer des formes plus spécifiques de ces expressions sous divers modèles et hypothèses au sujet des pondérations et du plan d'échantillonnage. Nous nous concentrons ici sur l'effet de la mise en grappes, en supposant que les pondérations sont égales et qu'il n'existe pas de stratification. Considérons le modèle à trois niveaux (5) et, pour simplifier les choses, supposons que $q = 1$ et $x_{ait} \equiv 1$ et que β est la moyenne de y_{ait}. Alors, des calculs algébriques simples montrent que la valeur de z_i^* pour l'individu i dans la grappe a est $[1 + \rho^*(T-1)]^{-1} \sum_t (\eta_a + u_{ai} + v_{ait})$. Supposons maintenant que l'on recourt à l'échantillonnage à deux degrés avec une taille d'échantillon commune m par grappe. Alors, en évaluant la variance $\text{var}(\bar{z}^*)$ et la limite de probabilité S_z^* dans (8) sous le modèle (5), nous trouvons, à l'instar de Skinner (1989a, page 38) que :

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = 1 + (m-1)\tau, \qquad (9)$$

où $\tau = \sigma_\eta^2 / (\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2 / T)$ est la corrélation intra-grappe de z_i^*. Nous voyons que, sous ce modèle, le meff augmente à mesure que T augmente (à condition que $\sigma_v^2 > 0$) et, donc, que l'effet de la mise en grappes sur l'estimation de la variance est plus important dans le cas longitudinal que dans le problème transversal (où $T = 1$).

Cette constatation dépend de l'hypothèse assez forte selon laquelle les effets de grappe η_a sont constants au

cours du temps. En fait, l'équation (9) est encore vérifiée si nous remplaçons η_a par un effet variable en fonction du temps η_{at} à condition que nous remplacions τ par $\tau = \text{var}(\bar{\eta}_a)/[\text{var}(\bar{\eta}_a) + \sigma_u^2 + \sigma_v^2/T]$, où $\bar{\eta}_a = \sum_t \eta_{at}/T$. Dans ces conditions, le meff augmentera à mesure que T augmente si (et uniquement si) $\sigma_u^2 + \sigma_v^2/T$ diminue plus rapidement avec T que $\text{var}(\bar{\eta}_a)$. Qu'il en soit ainsi ou non dépend de l'application particulière. Cependant, nous soutenons que, pour de nombreuses enquêtes longitudinales auprès d'individus avec les grappes fondées sur les régions (le genre de contexte auquel nous pensons), cette condition est plausible. Dans de telles applications, nous pouvons souvent nous attendre à ce que la valeur de σ_v^2 soit grande relativement à celle σ_u^2 (c'est-à-dire que la corrélation intra-grappe transversale soit faible), en particulier à cause d'une erreur de mesure propre à la vague d'enquête et, donc, à ce que $\sigma_u^2 + \sigma_v^2/T$ diminue assez rapidement à mesure que T augmente. En principe, les caractéristiques socio-économiques des régions sont souvent plus stables et dans des situations inhabituelles seulement devrait-on s'attendre à ce qu'une erreur de mesure donne lieu à une variance propre à la vague d'enquête importante dans η_{at}. Donc, selon nous, dans de telles applications, on pourrait habituellement s'attendre à ce que le ratio de $\text{var}(\bar{\eta}_a)$ pour $T = 5$, disons, comparativement à $T = 1$ soit plus grand que $(\sigma_u^2 + \sigma_v^2/5)/(\sigma_u^2 + \sigma_v^2)$ qui s'approchera de 1/5 à mesure que σ_u^2/σ_v^2 tend vers 0. Nous sommes donc d'avis que, dans de nombreuses circonstances pratiques, tenir compte de la mise en grappes sera plus important dans les analyses longitudinales que dans les analyses transversales correspondantes. Nous présentons un exemple empirique à la section 4.

Considérons maintenant les propriétés des estimateurs de variance basés sur le modèle à trois niveaux (5). Nous n'examinons que l'approche fondée sur l'hypothèse d'effets aléatoires homoscédastiques suivant une loi normale, en ignorant les poids de sondage, étant donné l'équivalence (virtuelle) de l'approche multiniveaux « robuste » et de la linéarisation.

Si le modèle (5) est correct et que nous ignorons effectivement les poids de sondage, alors l'estimateur de la variance basé sur le modèle sera convergent (Goldstein 1986). Cependant, comme il est discuté dans Skinner (1989b, page 68) et corroboré par la théorie dans Skinner (1986), la principale caractéristique de la mise en grappes susceptible d'avoir une incidence sur les erreurs-types des coefficients de régression estimés est la variation inter-grappes des coefficients de régression, ce qui n'est pas pris en compte dans le modèle (5).

Afin de voir comment le modèle (5) pourrait ne pas refléter correctement les effets de la mise en grappes, considérons le cas transversal $(T = 1)$, où x est un scalaire. Alors, si le modèle à trois niveaux (5) tient, une expression approximative du meff de l'estimateur de la variance de $\hat{\beta}$ basé sur le modèle à deux niveaux (2) est :

$$\text{meff} = 1 + (m-1)\tau_1\tau_x, \qquad (10)$$

où $\tau_1 = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2)$ et τ_x est la corrélation intra-grappe pour x (Scott et Holt 1982; Skinner 1989b, page 68). Ce résultat s'étend, dans le cas longitudinal, à :

$$1 \le \text{meff} \le 1 + (m-1)\tilde{\tau}\tau_z, \qquad (11)$$

où $\tilde{\tau}$ est la version de long terme $(T = \infty)$ de τ (voir l'annexe) et τ_z est un coefficient de corrélation intra-grappe pour $z_{ai} = \sum_t x_{ait}/T$. La preuve de ce résultat et les hypothèses simplificatrices requises sont esquissées à l'annexe. Le point principal est que $\tilde{\tau}$ et τ_z seront souvent faibles, auquel cas $\tilde{\tau}\tau_z$ sera très faible, et donc, la valeur de meff pourrait être invraisemblablement proche de un, l'estimateur de la variance basé sur le modèle présentant un biais par défaut. Nous explorons cet aspect empiriquement à la section 4. Naturellement, nous pourrions introduire des coefficients aléatoires dans le modèle (5), et nous examinons cela également à la section 4. Toutefois, étant donné la difficulté qu'il y a à spécifier correctement un modèle à coefficients aléatoires, il semble peu probable que cette approche soit très robuste.

À la présente section, nous nous sommes concentrés jusqu'à présent sur le biais (ou non-convergence) éventuel des méthodes d'estimation de la variance. Mais il est également souhaitable d'examiner leur efficacité. En particulier, nous pourrions nous attendre à ce que la méthode de linéarisation soit moins efficace que l'estimation de la variance basée sur le modèle, si ce dernier est correct. En principe, l'importance relative de l'efficacité par rapport au biais devrait augmenter à mesure que le nombre de grappes diminue. Wolter (1985, chapitre 8) résume un certain nombre d'études par simulation conçues pour examiner le biais et la variance de l'estimateur de variance par linéarisation qui laissent entendre que la méthode de linéarisation donne de bons résultats même si le nombre de grappes est faible. Fuller (1984) discute de corrections éventuelles des intervalles de confiance en fonction du nombre de degrés de liberté pour les coefficients de régression basés sur la méthode de linéarisation lorsque le nombre de grappes est faible. Une étude par simulation des estimateurs pour les modèles multiniveaux décrite dans Maas et Hox (2004) ne permet pas de conclure que la méthode de linéarisation donne de nettement moins bons résultats que l'approche basée sur une modèle, en ce qui a trait à la couverture des intervalles de confiance pour les coefficients de β, même si le nombre de grappes est aussi faible que 30.

4. Exemple : analyse par la régression des données de la BHPS sur les attitudes à l'égard des rôles de l'homme et de la femme

Nous présentons maintenant une application aux données de la BHPS afin d'illustrer certaines propriétés théoriques discutées à la section précédente.

Ces dernières décennies, nous avons été témoins d'une évolution importante des rôles de l'homme et de la femme au sein de la famille dans de nombreux pays. Les spécialistes des sciences sociales s'intéressent à la relation entre l'évolution des attitudes à l'égard des rôles de l'homme et de la femme, d'une part, et les changements de comportement, comme ceux relatifs à la condition parentale et à la participation au marché du travail, d'autre part (par exemple, Morgan et Waite 1987; Fan et Marini 2000). Diverses formes d'analyse statistique sont utilisées pour fournir des preuves de ces relations. Ici, nous considérons l'estimation d'un modèle linéaire de la forme (1) contenant comme variable de résultat, y, une mesure de l'attitude à l'égard des rôles de l'homme et de la femme, à la suite d'une analyse de Berrington (2002).

Les données proviennent des vagues 1, 3, 5, 7 et 9 (recueillies en 1991, 1993, 1995, 1997 et 1999, respectivement) de la BHPS et ces vagues sont codées $t = 1, ..., T = 5$ respectivement. On a demandé aux répondants s'ils étaient « tout à fait d'accord », « d'accord », « ni d'accord ni en désaccord », « en désaccord » ou « tout à fait en désaccord » avec une série d'énoncés concernant la famille, ainsi que les rôles de la femme et le travail à l'extérieur du ménage. Les réponses ont été cotées de 1 à 5. On a recouru à l'analyse factorielle pour évaluer quels énoncés pourraient être combinés en une mesure de l'attitude à l'égard des rôles de l'homme et de la femme. La cote d'attitude , y_{it}, considérée ici est la cote totale pour les six énoncés choisis pour la femme i lors de la vague t. Plus la cote est élevée, plus les attitudes à l'égard des rôles de l'homme et de la femme sont égalitaires. Berrington (2002) présente une discussion détaillée de cette variable. Une analyse plus complexe pourrait inclure un modèle d'erreur de mesure pour les attitudes (par exemple, Fan et Marini 2000), chacune des réponses sur l'échelle de cinq points aux six énoncés étant traitée comme une variable ordinaire. Ici, nous adoptons une approche plus simple consistant à traiter la cote agrégée y_{it} et le vecteur de coefficients connexe β comme présentant un intérêt scientifique et à inclure l'erreur de mesure dans le terme d'erreur du modèle.

Nous nous sommes fondés sur la discussion de Berrington (2002) en vue de choisir les covariables pour l'analyse par la régression, mais avons réduit leur nombre afin de nous concentrer plus facilement sur les questions méthodologiques d'intérêt. La covariable présentant le principal intérêt scientifique est l'activité économique, qui permet de faire la distinction, en particulier, entre les femmes qui restent au foyer pour s'occuper des enfants (dénoté « soin de la famille ») et celles qui poursuivent d'autres formes d'activité reliée au marché du travail. Les variables reflétant l'âge et le niveau de scolarité sont également incluses, puisqu'il a été souvent démontré qu'elles sont fortement corrélées aux attitudes à l'égard des rôles de l'homme et de la femme (par exemple, Fan et Marini 2000). Les valeurs de toutes ces covariables peuvent varier d'une vague à l'autre de l'enquête. Une variable d'année (prenant les valeurs 1, 3, …, 9) est également incluse. Elle peut refléter à la fois les changements chronologiques et le vieillissement général des femmes comprises dans l'échantillon.

La BHPS est une enquête-ménage par panel réalisée auprès des membres de la population à domicile de la Grande-Bretagne (Taylor, Brice, Buck et Prentice-Lane 2001). L'échantillon initial (vague 1) a été sélectionné en 1991 selon un plan stratifié à plusieurs degrés dans lequel les probabilités d'inclusion des ménages étaient à peu près égales. Les ménages ont été regroupés en 250 unités primaires d'échantillonnage (UPE) correspondant aux secteurs postaux. Tous les membres résidents de 16 ans et plus ont été sélectionnés dans les ménages échantillonnés. Tous les adultes sélectionnés durant la première vague ont été suivis lors de la deuxième vague et ainsi de suite, et représentent le panel longitudinal. L'enquête est sujette à une érosion de l'échantillon et à d'autres formes de non-réponse à une vague. Pour traiter cette non-réponse, nous avons simplement remplacé s dans (3) par l'« échantillon longitudinal » d'individus pour lesquels les observations étaient disponibles pour chacune des vagues $t = 1, ..., T$ et nous avons choisi de n'appliquer aucun poids de sondage, puisque notre but est d'étudier les effets d'erreur de spécification éventuellement associés à la mise en grappes et nous voulons éviter de confondre ces effets avec ceux de la pondération. Nous ignorons également l'effet de la stratification dans le travail numérique de la présente section (mais nous présentons à la section 5 certains commentaires sur l'effet de la pondération et de la stratification).

Puisque nous nous intéressons dans l'analyse à la question de savoir si l'activité principale des femmes consiste à « prendre soin de la famille », nous définissons notre population étudiée comme étant les femmes de 16 à 39 ans en 1991. Donc, nos données correspondent à l'échantillon longitudinal de femmes dont l'âge se situe dans la fourchette admissible et ayant répondu à toutes les questions de l'interview (enregistrements complets) lors de chacune des cinq vagues, ce qui donne un échantillon de $n = 1\,340$ femmes. Ces femmes sont réparties de manière relativement

uniforme entre 248 secteurs postaux. La petite taille moyenne d'échantillon d'environ cinq par secteur postal combinée à la corrélation intra-secteur postal relativement faible pour la variable d'attitude étudiée se traduit par un effet assez faible du plan, tel qu'il est mesuré par le meff. Puisque nos objectifs sont de nature méthodologique, nous avons choisi de grouper les secteurs postaux en 47 grappes géographiquement contigües, pour permettre des comparaisons plus précises, moins brouillées par les erreurs d'échantillonnage qui peuvent être appréciables dans l'estimation de la variance. Les valeurs du meff présentées dans les tableaux ont donc tendance à être plus élevées que pour le plan de sondage réel. Les seconds résultats ont tendance à suivre un profil comparable, quoique celui-ci soit moins précis, à cause de l'erreur d'échantillonnage.

Nous commençons par estimer le meff pour l'estimateur par la linéarisation, comme il est discuté au début de la section 3. À l'aide de données provenant uniquement de la première vague et en fixant $x_{ait} \equiv 1$, le meff estimé pour cette moyenne transversale donné dans le tableau 1 est de l'ordre de 1,5. Cette valeur est plausible car, si nous procédons à l'approximation habituelle de (9) pour des tailles d'échantillon de grappe inégales en remplaçant m par $\bar{m}$, la taille moyenne d'échantillon par grappe, nous trouvons que $1 + (\bar{m} - 1)\tau = 1{,}5$ et $\bar{m} = 1\,340/47 \approx 29$ impliquent une valeur de τ d'environ 0,02 et cette faible valeur est en accord avec d'autres valeurs estimées de τ obtenues pour des variables attitudinales dans les enquêtes britanniques (Lynn et Lievesley 1991, annexe D).

Tableau 1 Estimations pour les moyennes longitudinales

	$\hat{\beta}$	e.-t.	meff				
Vagues	1-9	1-9	1	1,3	1,3,5	1-7	1-9
	19,83	0,12	1,51	1,50	1,68	1,81	1,84

Pour évaluer l'effet de l'aspect longitudinal des données, nous estimons une série de meffs en utilisant les données provenant des vagues 1, ..., t pour $t = 2, 3, ..., 5$. Bien que ces meffs estimés soient sujets à une erreur d'échantillonnage, il semble évident, si l'on examine le tableau 1, que le meff tend à augmenter avec le nombre de vagues. Cette tendance pourrait être escomptée, compte tenu de la discussion théorique de la section 3, si le niveau moyen des attitudes égalitaires dans une région varie moins d'année en année que les cotes d'attitude individuelles des femmes. Cela paraît vraisemblable, puisque les secondes seront affectées à la fois par l'erreur de mesure et par les changements réels d'attitude, si bien qu'on pourrait prévoir que $\text{var}(\bar{\eta}_a)$ diminue plus lentement avec T que $\text{var}(\bar{u}_a + \bar{v}_a)$. Nous pouvons donc nous attendre à ce que τ, et par conséquent le meff, augmente à mesure que T croît, comme nous l'observons au tableau 1.

Nous étoffons ensuite l'analyse en introduisant des variables indicatrices d'activité économique comme covariables. Le modèle de régression résultant comprend un terme d'ordonnée à l'origine et quatre covariables représentant les contrastes entre les femmes occupées à temps plein et celles appartenant à d'autres catégories d'activité économique. Les valeurs estimées du meff sont présentées au tableau 2. L'ordonnée à l'origine est une moyenne de domaine et, selon la théorie classique du meff d'une moyenne dans un domaine recoupant les grappes (Skinner 1989b, page 60), sa valeur sera passablement plus faible que celle du meff pour la moyenne d'échantillon complet, ce que nous observons effectivement, le meff pour la moyenne de domaine transversale de 1,13 du tableau 2 étant inférieur à la valeur de 1,51 du tableau 1. Comme auparavant, le tableau 2 semble indiquer une tendance du meff à augmenter, de 1,13 dans le cas d'une seule vague à 1,50 dans le cas de cinq vagues, quoique ces valeurs soient plus faibles qu'au tableau 1. La taille du meff pour les contrastes du tableau 2 varie, certaines valeurs étant supérieures et d'autres inférieures à un. Ces valeurs du meff peuvent être considérées comme une combinaison de l'effet classique d'augmentation de la variance due à la mise en grappes dans les enquêtes et de l'effet de réduction de la variance due à la mise en blocs dans une expérience. Cette réduction de la variance a lieu si les domaines comparés ont un effet de grappe commun (de la forme η_a dans le modèle (5)) qui a tendance à s'annuler dans les contrastes, ce qui sous-entend que la variance réelle du contraste est plus faible que l'espérance de l'estimateur de variance fondé sur l'hypothèse d'indépendance entre domaines. Cette dernière espérance sera augmentée par les effets communs. La caractéristique de ces résultats qui présente le plus d'intérêt ici est que, de nouveau, le meff n'a pas tendance à converger vers l'unité à mesure que le nombre de vagues augmente. Si tant est qu'il y ait une tendance, celle-ci est de direction opposée. Pour le contraste d'intérêt scientifique ici, c'est-à-dire celui entre les femmes occupées à temps plein et celles qui « restent à la maison pour s'occuper de la famille », le meff est systématiquement nettement inférieur à un.

Nous perfectionnons ensuite davantage le modèle en incluant, comme covariables supplémentaires, le groupe d'âge, l'année et les qualifications. Les valeurs estimées du meff sont données au tableau 3. Le meff pour les coefficients de régression correspondant aux catégories d'activité économique varie de nouveau, certaines valeurs étant supérieures et d'autres inférieures à l'unité, pour les mêmes raisons que pour les contrastes (qui pourraient également être interprétés comme des coefficients de régression) du tableau 2. De nouveau, il semble que le meff ait tendance à s'écarter de l'unité à mesure que le nombre de vagues augmente. Une comparaison des tableaux 1 et 3 confirme

l'observation de Kish et Frankel (1974), à savoir que les valeurs du meff n'ont pas tendance à être plus élevées pour les coefficients de régression que pour les moyennes de la variable dépendante.

Tableau 2 Estimations pour la régression avec des covariables définies selon l'activité économique

	$\hat{\beta}$	e.-t.	meff				
Vagues	1-9	1-9	1	1,3	1,3, 5	1-7	1-9
Ordonnée à l'origine	20,58	0,11	1,13	1,01	1,09	1,38	1,50
Contrastes pour							
Occupée temps partiel	-1,03	0,10	0,93	0,91	0,93	1,00	0,89
Autre inactive	-0,80	0,15	0,60	0,96	0,68	0,76	0,81
Étudiante temps plein	0,41	0,24	1,10	1,32	1,14	1,48	1,44
Soin de la famille	-2,18	0,10	0,72	0,49	0,58	0,66	0,60

Nota : a) L'ordonnée à l'origine est la moyenne pour les femmes occupées à temps plein
b) Les contrastes sont calculés pour les autres catégories d'activité économique relativement au travail à temps plein.

Tableau 3 Estimations pour les coefficients de régression avec covariables supplémentaires dans le modèle

	$\hat{\beta}$	e.-t.	Meff				
Vagues	1-9	1-9	1	1,3	1,3, 5	1-7	1-9
Ordonnée à l'origine	20,20	0,30	0,95	0,87	0,87	1,04	1,07
Année, *t*	-0,04	0,01	-	0,86	0,69	0,59	0,96
Groupe d'âge							
16 à 21 ans	0,00	-					
22 à 27 ans	-0,71	0,25	1,22	1,37	1,44	1,73	1,64
28 à 33 ans	-0,89	0,27	1,38	1,40	1,46	1,68	1,59
34 ans et plus	-1,03	0,27	0,94	1,10	1,13	1,26	1,34
Activité économique							
Occupée temps plein	0,00	-					
Occupée temps partiel	-0,93	0,10	0,97	0,95	0,96	1,06	0,91
Autre inactive	-0,75	0,15	0,60	0,96	0,68	0,77	0,81
Étudiante temps plein	0,17	0,24	0,93	1,32	1,23	1,39	1,32
Soin de la famille	-2,09	0,10	0,77	0,59	0,70	0,78	0,67
Qualification							
Diplôme	0,00	-					
Qualif.	-0,52	0,21	0,77	0,64	0,75	0,87	0,85
Niveau A	-0,61	0,24	0,98	0,87	0,94	0,94	1,01
Niveau O	-0,44	0,20	0,62	0,62	0,59	0,69	0,73
Autre	-1,16	0,22	0,83	0,83	0,78	0,80	0,82

Nous considérons ensuite les erreurs-types fondées sur le modèle obtenues à partir du modèle à trois niveaux (5), comme il est discuté à la section 2. Les résultats sont présentés au tableau 4 dans la colonne intitulée « fondée sur le modèle à trois niveaux ». Aux fins de comparaison, nous estimons aussi les erreurs-types sous le modèle à deux niveaux (2) et présentons les résultats dans la colonne intitulée « fondée sur le modèle à deux niveaux ». Les estimations qui figurent dans ces deux colonnes sont presque identiques. Il existe un écart d'un chiffre au niveau de la troisième décimale pour certains coefficients et un écart un peu plus important pour l'ordonnée à l'origine. Nous pensons qu'il s'agit d'une preuve qu'ajouter simplement un terme d'effet régional aléatoire peut donner lieu à une sous-estimation importante de l'effet de la mise en grappes sur les erreurs-types estimées des coefficients de régression. Ces données sont en accord avec la borne supérieure théorique du meff donnée en (11). La valeur estimée de $\tilde{\tau}$ dans l'expression (11) est 0,019 et aucune des covariables ne devrait, en principe, présenter une corrélation intra-régionale importante, de sorte que les valeurs prévues des estimateurs de la variance pour les modèles à deux et à trois niveaux devraient être très proches.

Nous avançons à la section 3 que la caractéristique de la mise en grappes principalement susceptible d'avoir une incidence sur la matrice de covariance de $\hat{\beta}$ est la variation inter-grappes des coefficients de régression. Nous avons exploré cette idée en introduisant des coefficients aléatoires dans le modèle. En traitant alors les éléments de β comme les valeurs prévues des coefficients aléatoires, nous avons constaté que les estimations de β avaient à peine changé. Nous avons trouvé que les erreurs-types estimées de ces estimations étaient, en effet, exagérées, et ce, bien davantage que par introduction de l'effet aléatoire supplémentaire de grappe dans le modèle (5), et que l'accroissement était du même ordre de grandeur que ceux des meffs des tableaux 2 et 3. Néanmoins, la méthode des MCGI a produit plusieurs estimations négatives des variances des coefficients aléatoires, ce qui oblige à s'interroger sur les coefficients qu'il faut laisser varier ou, plus généralement, sur la spécification du modèle. Ce problème s'accentue à mesure qu'augmente le nombre de covariables, car le nombre de paramètres dans la matrice de covariance du vecteur de coefficients augmente en fonction du carré du nombre de covariables. Dans l'ensemble, l'introduction de coefficients aléatoires semble créer au moins autant de problèmes qu'elle n'en résout si la mise en grappes ne présente pas d'intérêt scientifique intrinsèque et ne semble pas être un moyen très satisfaisant de tenir compte de la mise en grappes dans l'estimation de la variance. Il est plus simple de changer de méthode d'estimation.

Comme nous l'avons mentionné à la fin de la section 2, une option est l'utilisation d'une méthode d'estimation de la variance « robuste » basée sur le modèle (5) (Goldstein 2003, page 80). Les valeurs de ces estimations robustes de l'erreur-type sont également incluses dans le tableau 4. Comme nous l'avons prévu à la section 2, l'estimateur robuste de l'erreur-type pour le modèle à deux niveaux donne des résultats très semblables à ceux de l'estimateur par linéarisation qui ne tient pas compte de la mise en grappes. L'estimateur robuste de l'erreur-type pour le modèle à trois niveaux donne des résultats fort semblables à l'estimateur par linéarisation tenant compte de l'échantillonnage à deux degrés. Les légers écarts reflètent les différences entre les méthodes d'estimation de V.

La méthode de linéarisation en présence d'échantillonnage à deux degrés s'approche donc fort de la méthode d'estimation robuste de la variance utilisée dans la littérature sur la modélisation multiniveaux. La distinction entre les méthodes devient plus prononcée si nous tenons également compte de la stratification et de la pondération. Une autre différence est que, dans la modélisation multiniveaux, les écarts entre les erreurs-types fondées sur le modèle pourraient servir d'outil diagnostique pour déceler des divergences par rapport au modèle (Maas et Hox 2004). Ainsi, les écarts importants entre les erreurs-types à trois niveaux pour les coefficients des groupes d'âge du tableau 4 pourraient mener à envisager l'inclusion de coefficients aléatoires pour le groupe d'âge. Cette situation diffère de l'approche du plan de sondage, où la structure d'erreur incluse dans le modèle (5) est traitée uniquement comme un modèle de travail et où on ne s'attend pas nécessairement à ce que les erreurs-types basées sur ce modèle soient approximativement valides.

Tableau 4 Erreurs-types estimées des coefficients de régression

	Linéarisation		**Modélisation multiniveaux**			
	EAS	Complexe	Fondée sur le modèle à 2 niveaux	Robuste à 2 niveaux	Fondée sur le modèle à 3 niveaux	Robuste à 2 niveaux
Ordonnée à l'origine	0,287	0,296	0,253	0,288	0,259	0,293
Année, *t*	0,014	0,014	0,013	0,014	0,013	0,014
Groupe d'âge						
16 à 21 ans						
22 à 27 ans	0,191	0,245	0,155	0,192	0,155	0,243
28 à 33 ans	0,214	0,270	0,187	0,215	0,187	0,266
34 ans et plus	0,237	0,275	0,218	0,238	0,218	0,271
Activité économique						
Occupée temps plein						
Occupée temps partiel	0,103	0,098	0,098	0,103	0,098	0,096
Autre inactive	0,166	0,150	0,146	0,166	0,146	0,148
Étudiante temps plein	0,207	0,238	0,199	0,207	0,199	0,236
Soin de la famille	0,125	0,102	0,112	0,125	0,112	0,101
Qualification						
Diplôme						
Qualif.	0,228	0,210	0,207	0,228	0,208	0,211
Niveau A	0,238	0,239	0,209	0,240	0,210	0,237
Niveau O	0,234	0,199	0,217	0,235	0,218	0,199
Autre	0,247	0,224	0,229	0,249	0,230	0,223

5. Discussion

Nous avons présenté certains arguments théoriques et certaines données empiriques soutenant la thèse selon laquelle ne pas tenir compte de la mise en grappes dans l'estimation des erreurs-types pourrait avoir un effet plus important dans certaines analyses longitudinales que dans les analyses transversales correspondantes. Cela signifie, qu'en général, il est au moins aussi important de tenir compte de la mise en grappes dans l'estimation des erreurs-types pour les analyses longitudinales que pour les analyses transversales et que les constatations de Kish et Frankel (1974), par exemple, ne devraient pas être utilisées pour justifier d'ignorer l'échantillonnage complexe dans le premier cas.

Les analyses longitudinales examinées dans le présent article sont d'un certain type, et il convient de souligner que les tendances observées pour les meffs dans ce genre d'analyse pourraient fort bien ne pas s'étendre à toutes les formes d'analyse longitudinale. En guise de spéculation quant à la classe de modèles et d'estimateurs à laquelle les profils observés dans le présent article pourraient s'appliquer, nous conjecturons que l'accroissement du meff dans les analyses longitudinales survient quand le plan longitudinal permet d'extraire des différences entre individus la variation temporelle « aléatoire » des réponses individuelles, donc, de réduire la composante des erreurs-types due à ces différences, mais qu'il fournit moins d'« explication » sur les différences entre grappes, de sorte que l'importance relative de cette composante des erreurs-types s'accroît.

Les travaux empiriques présentés ici ont également été restreints par l'effet de la mise en grappes. Nous avons entrepris des analyses correspondantes en tenant compte de la pondération et de la stratification, et avons obtenu des résultats en général semblables. La stratification a tendance à avoir un effet plus faible que la mise en grappes. Dans le cas de la BHPS, les probabilités de sélection dans l'échantillon varient peu, et l'effet de la pondération par l'inverse de ces probabilités sur les estimations ponctuelles ainsi que sur les estimations de la variance a tendance à ne pas être important. Nous observons plutôt une variation plus importante des poids longitudinaux qui sont fournis avec les données de la BHPS pour les analyses des ensembles d'individus qui ont répondu à chaque vague jusqu'à une année T donnée inclusivement. L'effet de ces poids sur les estimations ponctuelles et sur les estimations de la variance est un peu plus grand. À mesure que T augmente et que l'érosion de l'échantillon se poursuit, les poids longitudinaux tendent à devenir plus variables et donnent lieu à une augmentation plus importante des variances, ce qui a tendance à accroître l'effet que nous avons décrit des meffs croissants avec T.

En laissant de côté la stratification et la pondération, nous avons comparé deux approches en vue de tenir compte de l'échantillonnage en grappes. Nous avons considéré l'approche du plan de sondage comme étant l'approche de référence. Nous avons également examiné une approche de modélisation multiniveaux pour tenir compte de la mise en grappes. Nous soutenons que l'utilisation d'un simple effet aléatoire additif pour représenter la mise en grappes peut donner lieu à une sous-estimation grave de l'effet de cette dernière et entraîner une sous-estimation des erreurs-types. Si la mise en grappes présente un intérêt scientifique, une solution consisterait à envisager d'inclure des coefficients aléatoires. Une autre consisterait à utiliser l'approche des équations d'estimation généralisées « GEE2 » (Liang, Zeger et Qaqish 1992) et à spécifier un modèle paramétrique supplémentaire pour $E(y_i\, y_i')$. Si la mise en grappes est traitée comme une perturbation reflétant simplement une mesure de convenance administrative lors de la collecte des données, nous sommes d'avis que l'approche du plan de sondage offre plusieurs avantages pratiques. Ceux-ci sont discutés plus en détail dans Lavange et coll. (1996, 2001) dans le contexte d'autres applications à des données sur des mesures répétées.

Annexe

Justification de (11)

Pour simplifier, nous posons que x et β sont des scalaires et que $\hat{\beta}$ est l'estimateur par les moindres carrés ordinaires et nous supposons que les tailles des échantillons dans les grappes sont toutes égales à m. Dans (11), le meff est défini comme étant $\mathrm{var}_3(\hat{\beta}) / E_3[v_2(\hat{\beta})]$, où E_3 et var_3 sont les moments sous le modèle à trois niveaux (5) et $v_2(\hat{\beta})$ est un estimateur de variance fondé sur le modèle à deux nivaux (2). Sous (5), nous obtenons

$$\mathrm{var}_3(\hat{\beta}) = \left(\sum_{cit} x_{cit}^2\right)^{-2}\left(\sigma_\eta^2 \sum_c x_{c++}^2 + \sigma_u^2 \sum_{ci} x_{ci+}^2 + \sigma_v^2 \sum_{cit} x_{cit}^2\right),$$

où $+$ dénote la sommation sur un suffixe, $\sigma_\eta^2, \sigma_u^2$ et σ_v^2 sont les variances respectives de η_a, u_{ai} et v_{ait}, et x_{cit} est centrée à 0. Nous supposons en outre que $v_2(\hat{\beta})$ est définie de telle façon que $E[v_2(\hat{\beta})] \approx (\sum_{cit} x_{cit}^2)^{-2}[(\sigma_\eta^2 + \sigma_u^2)\sum_{ci} x_{ci+}^2 + \sigma_v^2 \sum_{cit} x_{cit}^2]$.

Après certaines opérations algébriques, nous pouvons montrer que :

$$\mathrm{meff} = 1 + (m-1)\tilde{\tau}\tau_z\rho[1 + (T-1)\tau_x]/[1 + (T-1)\rho\tau_x], \quad (12)$$

où $\tilde{\tau} = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_u^2)$, $\rho = (\sigma_\eta^2 + \sigma_u^2)/(\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2)$, $\tau_x = \sigma_{xB}^2/\sigma_x^2$, $\sigma_x^2 = \sum_{cit} x_{cit}^2/(nT)$, $\sigma_{xB}^2 = [\sum_{ci}(x_{ci+}/T)^2/n - \sigma_x^2/T]/[1 - 1/T]$, $\tau_z = \sigma_{zB}^2/\sigma_z^2$, $\sigma_z^2 = \sum_{ci} z_{ci}^2/n$, $\sigma_{zB}^2 = [\sum_c (z_{c+}/m)^2/C - \sigma_z^2/m]/[1 - 1/m]$ et $n = Cm$ est la taille d'échantillon. Notons que $\tilde{\tau}\rho = \tau_1$ et, quand $T = 1$,

$\tau_z = \tau_x$ de sorte que (12) se réduit à (10). En général, $\rho \leq 1$ et (11) découle de (12). En fait, dans notre application, nous estimons que ρ est 0,59, de sorte que, dans (11), les bornes ne devraient pas être très rapprochées.

Remerciements

Les travaux du deuxième auteur ont été financés par la subvention 20.0286/01.3 du Conseil national du Brésil pour le développement scientifique et technologique (CNPq).

Bibliographie

Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*. 2ième Éd. Chichester : John Wiley & Sons, Inc.

Berrington, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study. Dans *Meaning and Choice*: *Value Orientations and Life Course Decisions*. (Éd., R. Lesthaeghe) Brussels : NIDI.

Chambers, R.L., et Skinner, C.J. Éds. (2003). *Analysis of Survey Data*. Chichester : John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-92.

Diggle, P.J., Heagerty, P., Liang, K. et Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2ième Éd. Oxford : Oxford University Press.

Fan, P.-L., et Marini, M.M. (2000). Influences on gender-role attitudes during the transition to adulthood. *Social Science Research*, 29, 258-283.

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

Fuller, W.A. (1984). Application de la méthode des moindres carrées et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-137.

Fuller, W.A., et Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 74, 430-431.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3ième Éd. London : Arnold.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kish, L., et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Séries B, 36, 1-37.

Lavange, L.M., Koch, G.G. et Schwartz, T.A. (2001). Applying sample survey methods to clinical trials data. *Statistics in Medicine*, 20, 2609-23.

Lavange, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G. et Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, 5, 311-329.

Liang, K.Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Liang, K.Y., Zeger, S.L. et Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society*, Séries B, 54, 3-40.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. et Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 50, 270-278.

Lynn, P., et Lievesley, D. (1991). *Drawing General Population Samples in Great Britain*. London : Social and Community Planning Research.

Maas, C.J.M., et Hox, J.J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46, 427-440.

Morgan, S.P., et Waite, L.J. (1987). Parenthood and the attitudes of young adults. *Am. Sociological Review*, 52, 541-547.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Séries B, 60, 23-56.

Renard, D., et Molenberghs, G. (2002). Multilevel modelling of complex survey data. Dans *Topics in Modelling Clustered Data* (Éds., M. Aerts, H. Geys, G. Molenberghs et L.M. Ryan). Boca Raton : Chapman and Hall/CRC. 263-272.

Scott, A.J., et Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

Shah, B.V., Barnwell, B.G. et Bieler, G.S. (1997). SUDAAN User's manual, release 7.5. Research triangle park, NC : Research Triangle Institute.

Skinner, C.J. (1986). Design effects of two stage sampling. *Journal of the Royal Statistical Society*, Séries B, 48, 89-99.

Skinner, C.J. (1989a). Introduction to Part A. In *Analysis of Complex Surveys*, (Éds., C.J. Skinner, D. Holt et T.M.F. Smith) Chichester : Wiley. 23-58.

Skinner, C.J. (1989b). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*. (Éds., C.J. Skinner, D. Holt et T.M.F. Smith) Chichester : Wiley. 59-87.

Skinner, C.J., Holt, D. et Smith, T.M.F. Éds. (1989). *Analysis of Complex Surveys*. Chichester : Wiley.

Taylor, M.F. éd, Brice, J., Buck, N. et Prentice-Lane, E. (2001). *British Household Panel Survey - User Manual - Volume A*: *Introduction*, *Technical Report and Appendices*. Colchester, University of Essex.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 15-25
Statistique Canada, N° 12-001 au catalogue

Modélisation des durées de périodes multiples à partir de données d'enquête longitudinale

Milorad S. Kovačević et Georgia Roberts [1]

Résumé

Nous étudions certaines modifications du modèle de Cox à période unique classique afin de traiter les périodes multiples chez une même personne lorsque les données sont recueillies dans le cadre d'une enquête longitudinale à plan d'échantillonnage complexe. L'une des modifications est l'utilisation d'une approche fondée sur le plan de sondage pour l'estimation des coefficients du modèle et de leurs variances; dans l'estimation de la variance, chaque individu est traité comme une grappe de périodes, ce qui ajoute un degré supplémentaire de mise en grappes dans le plan de sondage. D'autres modifications du modèle ont pour but de rendre souple la spécification du risque de base afin de tenir compte de la dépendance différentielle éventuelle du risque à l'égard de l'ordre et de la durée des périodes successives, et de tenir compte aussi des effets différentiels des covariables sur les périodes de différents ordres. Ces approches sont illustrées en utilisant des données provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada.

Mots clés : Régression de Cox; inférence fondée sur le plan de sondage; inférence fondée sur un modèle; ordre de la période; EDTR.

1. Introduction

Le problème de modélisation abordé dans le présent article est connu sous divers noms, tels que modélisation des temps de défaillance corrélés, modélisation multivariée de la survie, modélisation de périodes multiples ou problème d'événements récurrents. Il a été étudié dans la littérature biomédicale (par exemple, Lin 1994, Hougaard 1999), sociale (Blossfeld et Hamerle 1989, Hamerle 1989) et économique (Lancaster 1979, Heckman et Singer 1982). Généralement, ce type de modélisation est requis pour résoudre les problèmes qui se posent dans les études du temps écoulé jusqu'à l'événement, lorsque deux événements ou plus surviennent chez le même sujet et que le but de la recherche est d'évaluer l'effet de diverses covariables sur la durée d'une période dans un état particulier. Les temps de défaillance sont corrélés chez un sujet donné et, donc, l'hypothèse d'indépendance des temps de défaillance conditionnellement à des covariables mesurées que requièrent les modèles de survie standard est vraisemblablement violée. Dans les études de durée des périodes (de pauvreté, de chômage, *etc.*), la « défaillance » équivaut à la sortie de l'état d'intérêt. Une propriété supplémentaire d'un grand nombre de périodes multiples, souvent ignorée, est que les périodes sont des « événements » ordonnés; autrement dit, la deuxième période ne peut pas survenir avant la première, et ainsi de suite. Le présent article a été motivé par une étude des périodes de chômage dont il est discuté plus en détail à la section 5.

L'interdépendance des périodes survenant chez un même individu est due au fait qu'elles ont en commun certaines caractéristiques inobservées de l'individu. L'effet de ces caractéristiques inobservées peut être modélisé explicitement sous forme d'un effet aléatoire (par exemple, Clayton et Cuzick 1985). Le cas échéant, il est supposé que l'effet aléatoire suit une loi statistique connue. La loi gamma de moyenne 1 et de variance inconnue est la loi privilégiée dans de nombreuses applications. Ensuite, des estimations des effets aléatoires et fixes peuvent être obtenues par une méthode appropriée (par exemple, vraisemblance en deux étapes (Lancaster 1979), en utilisant un algorithme EM (Klein 1992), *etc.*). Cette approche n'est pas explorée dans le présent article.

Une autre approche, qui est celle que nous utiliserons, consiste à adopter une méthode semi-paramétrique dans laquelle nous ne modélisons pas explicitement l'interdépendance des périodes multiples. Nous modélisons les lois marginales des périodes individuelles, en utilisant éventuellement l'ordre des périodes dans la spécification du modèle. Dans un contexte autre que les sondages, Lin (1994) décrit comment il est suffisant de modifier simplement la matrice de covariance « naïve » des coefficients estimés du modèle obtenue sous l'hypothèse d'indépendance, puisque les durées corrélées doivent être prises en compte dans les estimations de la variance, mais non dans les estimations des coefficients proprement dit.

Dans les études socioéconomiques des durées des périodes, les sources de données sont fréquemment des enquêtes longitudinales à plan d'échantillonnage complexe comportant une stratification, un échantillonnage à plusieurs degrés, la sélection avec probabilités inégales, des corrections stochastiques de l'érosion et de la non-réponse, le

1. Milorad S. Kovačević, conseiller en recherche méthodologique, Statistique Canada, Ottawa, Canada, K1A 0T6. Courriel : kovamil@statcan.ca; Georgia Roberts, chef du Centre de ressources en analyse de données à la Division des méthodes d'enquête sociales, Statistique Canada, Ottawa, Canada, K1A 0T6. Courriel : robertg@statcan.ca.

calage sur des paramètres connus, *etc*. Par conséquent, il est nécessaire de tenir compte de l'effet du plan d'échantillonnage sur la distribution des données d'échantillon lors de l'estimation des paramètres du modèle et des variances de ces estimations. Notre approche, lors de l'analyse de données d'enquête complexes, consiste à modéliser les lois marginales des périodes multiples selon des méthodes pour période unique, en traitant la dépendance entre les périodes comme une perturbation - aussi bien la dépendance due à la corrélation des périodes chez la même personne que la dépendance entre individus due au plan de sondage - mais à tenir compte des probabilités de sélection inégales à l'aide des poids de sondage. En fonction du modèle choisi, les paramètres de population finie sont définis et estimés comme dans Binder (1992). Les erreurs-types sont estimées par une méthode appropriée de linéarisation convergente sous le plan en posant l'hypothèse que les unités primaires d'échantillonnage sont échantillonnées avec remise dans les strates. Cette hypothèse est valide lorsque les fractions d'échantillonnage de premier degré sont faibles, comme cela est généralement le cas dans les enquêtes socioéconomiques. En outre, pour ce genre d'échantillon, la différence entre les inférences en population finie et en superpopulation (c'est-à-dire les erreurs-types et les statistiques de test) s'avère assez négligeable (Lin 2000). Par conséquent, les résultats de l'inférence basée sur notre approche peuvent s'étendre au-delà de la population finie étudiée.

À la section suivante, nous passons en revue la modélisation de périodes uniques et certaines méthodes d'estimation robuste des variances lorsque le modèle est spécifié incorrectement, d'abord dans un cadre fondé sur un modèle, puis dans un cadre fondé sur le plan de sondage. À la section 3, nous discutons plus en détail de l'estimation robuste de la variance pour des périodes multiples. À la section 4, nous introduisons trois modèles pour périodes multiples et décrivons comment les ajuster en utilisant des méthodes d'estimation robuste sous le plan de sondage. À la section 5, nous ajustons ces modèles aux données de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada et discutons des résultats. Enfin, à la section 6, nous présentons certaines remarques générales.

2. Inférence pour le modèle de taux de risque à période unique

La durée d'une période (ou simplement, une période) vécue par un individu est une variable aléatoire dénotée par T. Nous nous intéressons particulièrement à la fonction de risque $h(t)$ de T au temps t, définie comme étant le risque instantané d'achèvement de la période au temps t sachant qu'elle n'a pas été achevée avant le temps t, exprimée formellement par

$$h(t) = \lim_{dt \to 0} \frac{\text{Prob}\{t \le T < t + dt \mid T \ge t\}}{dt}.$$

La valeur de la fonction de risque au temps t est appelée taux de sortie afin de mettre l'accent sur le fait que l'achèvement de la période équivaut à la sortie de l'état d'intérêt. Les modèles de durée et l'analyse de la durée en général sont formulés et discutés en termes de la fonction de risque et de ses propriétés.

Du point de vue d'un spécialiste du domaine, l'intérêt principal est souvent d'étudier l'effet de certaines covariables clés sur la distribution de T. Un modèle à risques proportionnels est fréquemment choisi pour ce genre d'étude. Sous le modèle à risques proportionnels, la fonction de risque de la période T, étant donné un vecteur de covariables variant éventuellement en fonction du temps $\mathbf{x}(t) = (x_1(t), ..., x_p(t))'$ est

$$h(t \mid \mathbf{x}(t)) = \lambda_0(t)\, e^{\mathbf{x}'(t)\boldsymbol{\beta}}. \qquad (1)$$

La fonction $\lambda_0(t)$ est une fonction de risque de base non spécifiée qui donne la forme de $h(t \mid \mathbf{x}(t))$. Le risque de base décrit la dépendance de la durée, comme préciser si le taux de risque dépend du temps déjà écoulé dans la période. Par exemple, une dépendance négative décrit la situation où la probabilité de sortie d'un état est d'autant plus faible que la période est longue. Si la valeur de toutes les variables $\mathbf{x}(t)$ d'un individu est fixée à 0, la valeur (niveau) de la fonction de risque est égale au risque de base.

2.1 Inférence fondée sur le modèle

Le vecteur $\boldsymbol{\beta}$ contient les paramètres de régression inconnus montrant la dépendance du risque par rapport au vecteur $\mathbf{x}(t)$ et peut être estimé en maximisant la fonction de vraisemblance partielle (Cox 1975) :

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i'(T_i)\boldsymbol{\beta}}}{\sum_{j=1}^{n} Y_j(T_i)\, e^{\mathbf{x}_j'(T_i)\boldsymbol{\beta}}} \right]^{\delta_i}. \qquad (2)$$

Ici, $T_1, ..., T_n$ représentent n durées éventuellement censurées à droite; $\delta_i = 1$ si T_i est une durée observée et $\delta_i = 0$ autrement; et $\mathbf{x}_i(t)$ est le vecteur de covariables correspondant observé sur $[0, T_i]$. Au dénominateur, la somme est calculée sur les périodes qui risquent d'être achevées au temps T_i, c'est-à-dire $Y_j = 1$ si $t \le T_j$, et égale à 0 autrement. L'estimation $\hat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ du modèle est obtenue en résolvant l'équation de score de vraisemblance partielle

$$U_0(\boldsymbol{\beta}) = \sum_{i=1}^{n} u_{i0}(T_i, \boldsymbol{\beta}) = 0, \qquad (3)$$

où

$$u_{i0}(T_i,\ \boldsymbol{\beta}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{S^{(1)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} \right\}, \qquad (4)$$

$$S^{(0)}(t,\ \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}, \qquad (5)$$

et

$$S^{(1)}(t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\, \mathbf{x}_i(t)\, e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}. \qquad (6)$$

Si le modèle (1) est vérifié et que les durées sont indépendantes, la matrice des variances fondées sur le modèle de la fonction de score $U_0(\boldsymbol{\beta})$ est

$$J(\boldsymbol{\beta}) = -\partial U_0(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$$
$$= \sum_{i=1}^{n} \delta_i \left\{ \frac{S^{(2)}(T_i,\ \boldsymbol{\beta})}{S^{(0)}(T_i,\ \boldsymbol{\beta})} - \frac{S^{(1)}(T_i,\ \boldsymbol{\beta})[S^{(1)}(T_i,\ \boldsymbol{\beta})]'}{[S^{(0)}(T_i,\ \boldsymbol{\beta})]^2} \right\},$$

où

$$S^{(2)}(t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\ \mathbf{x}_i(t)\ \mathbf{x}_i'(t)\ e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}.$$

La variance approximative de $\hat{\boldsymbol{\beta}}$, obtenue par linéarisation, est $J^{-1}(\hat{\boldsymbol{\beta}})$.

Si la forme de (1) est incorrecte, mais que les observations sont indépendantes, Lin et Wei (1989) donnent l'estimateur robuste de la variance pour $\hat{\boldsymbol{\beta}}$ sous la forme

$$J^{-1}(\hat{\boldsymbol{\beta}})\, G(\hat{\boldsymbol{\beta}})\, J^{-1}(\hat{\boldsymbol{\beta}}), \qquad (7)$$

où

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta})\, g_i'(\boldsymbol{\beta})$$

et

$$g_i(\boldsymbol{\beta}) = u_{i0}(T_i, \boldsymbol{\beta})$$
$$-\sum_{j=1}^{n} \delta_j \frac{Y_i(T_j) e^{\mathbf{x}_i'(T_j)\boldsymbol{\beta}}}{nS^{(0)}(T_j, \boldsymbol{\beta})} \left\{ \mathbf{x}_i(T_j) - \frac{S^{(1)}(T_j, \boldsymbol{\beta})}{S^{(0)}(T_j, \boldsymbol{\beta})} \right\}. \qquad (8)$$

2.2 Inférence fondée sur le plan de sondage

Pour les observations provenant d'une enquête à plan d'échantillonnage complexe, Binder (1992) a utilisé une méthode de pseudovraisemblance pour estimer les paramètres d'un modèle à risques proportionnels et leurs variances dans le cas d'une seule période par individu. En particulier, il a commencé par définir le paramètre d'intérêt en population finie comme étant la solution de l'équation de score de vraisemblance partielle (3) calculée d'après les périodes de la population finie visée par l'enquête :

$$U_0(\mathbf{B}) = \sum_{i=1}^{N} u_{i0}(T_i, \mathbf{B}) = 0,$$

où $u_{i0}(T_i, \mathbf{B})$ est le résidu de score défini de la même manière que $u_{i0}(T_i, \boldsymbol{\beta})$, excepté que les moyennes dans les définitions de $S^{(0)}(t, \mathbf{B})$ et de $S^{(1)}(t, \mathbf{B})$ portent sur N plutôt que n observations. Notons que, si les membres de la population finie visée par l'enquête ne connaissent pas tous des périodes de l'état étudié, N représente la taille de la sous-population qui vit de telles périodes et la sommation est faite sur ces N individus.

Une estimation $\hat{\mathbf{B}}$ du paramètre $\mathbf{B}$ est obtenue sous forme d'une solution de l'équation d'estimation du pseudo-score partiel

$$\hat{U}_0(\hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = 0,$$

où $w_i(s) = w_i$, le poids de sondage, si $i \in s$, et 0 autrement. La fonction $\hat{u}_{i0}(T_i, \hat{\mathbf{B}})$ prend la forme

$$\hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})} \right\},$$

où

$$\hat{S}^{(0)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}},$$

et

$$\hat{S}^{(1)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i(s)\, Y_i(t)\, \mathbf{x}_i(t)\ e^{\mathbf{x}_i'(t)\hat{\mathbf{B}}}.$$

En général, la variance sous le plan de sondage d'une estimation $\hat{\boldsymbol{\theta}}$ qui satisfait une équation d'estimation de la forme $\hat{U}(\hat{\boldsymbol{\theta}}) = \sum w_i u_i(\hat{\boldsymbol{\theta}}) = 0$ peut être estimée, par linéarisation, en tant que

$$\hat{J}^{-1}\, \hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))\, \hat{J}^{-1}, \qquad (9)$$

où $\hat{J} = \partial \hat{U}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ est évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, et $\hat{V}(\hat{U}(\hat{\boldsymbol{\theta}}))$ est la variance estimée du total estimé $\hat{U}(\boldsymbol{\theta})$ obtenue par une méthode standard d'estimation de la variance fondée sur le plan (voir, par exemple, Cochran (1977)) et évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) énonce qu'afin d'utiliser cette approche pour calculer une estimation convergente de la variance, $\hat{U}(\hat{\boldsymbol{\theta}})$ doit être exprimée sous la forme d'une somme de vecteurs aléatoires indépendants. Dans le cas du modèle à risques proportionnels susmentionné, $\hat{U}_0(\hat{\mathbf{B}})$ ne satisfait pas cette condition, puisque chaque $\hat{u}_{i0}$ est une fonction de $\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})$ et de $\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})$, qui englobent tous deux de nombreux individus outre le i^{e}. Donc, Binder (1992) a trouvé pour $\hat{U}_0(\hat{\mathbf{B}})$ une expression de rechange conforme à

ces conditions, ce qui permet d'obtenir une estimation convergente sous le plan $\hat{V}(\hat{U}_0(\hat{\mathbf{B}}))$ par application d'une méthode d'estimation de la variance fondée sur le plan à l'expression de rechange, puis l'évaluation de cette estimation de la variance à $\mathbf{B} = \hat{\mathbf{B}}$. Si la méthode d'estimation de la variance sous le plan choisie est la méthode de linéarisation, alors la première étape consiste à calculer le résidu suivant pour chacun des individus échantillonnés :

$$\hat{u}_i(T_i, \hat{\mathbf{B}}) = \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) - \sum_{j=1}^{N} w_j(s)\,\delta_j \frac{Y_i(T_j)\, e^{\mathbf{x}_i'(T_j)\hat{\mathbf{B}}}}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \left\{ \mathbf{x}_i(T_j) - \frac{\hat{S}^{(1)}(T_j, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \right\}. \quad (10)$$

Chaque individu compris dans l'échantillon appartient à une UPE particulière dans une strate donnée. Donc, au lieu d'identifier un individu à l'aide d'un indice inférieur unique i, nous utiliserons un indique inférieur triple hci où $h = 1, 2, ..., H$ identifie la strate, $c = 1, 2, ..., c_h$ identifie l'UPE dans la strate et $i = 1, 2, ..., n_{hc}$ identifie l'individu dans l'UPE. Alors,

$$\hat{V}(\hat{U}_0(\hat{\mathbf{B}})) = \sum_{h=1}^{H} \frac{1}{c_h(c_h - 1)} \sum_{c=1}^{c_h} (t_{hc} - \bar{t}_h)\,(t_{hc} - \bar{t}_h)',$$

où

$$t_{hc} = c_h \sum_{i=1}^{n_{hc}} w_{hci}\,\hat{u}_{hci} \quad \text{et} \quad \bar{t}_h = \sum_{c=1}^{c_h} t_{hc} \Big/ c_h.$$

3. Inférence pour les modèles du taux de risque à périodes multiples

3.1 Inférence fondée sur le modèle

Si plus d'une période est observée pour un individu, il est raisonnable de supposer que ces périodes ne sont pas indépendantes. Donc, la fonction de vraisemblance partielle (2) est spécifiée incorrectement pour des périodes multiples, puisqu'elle ne tient pas compte de la corrélation intra-individu des périodes observées chez le même individu. En s'inspirant de Lin et Wei (1989), il suffit de modifier seulement la matrice de covariance des paramètres du modèle estimé, puisque les durées corrélées affectent la variance, tandis que les paramètres du modèle peuvent être estimés de manière convergente sans tenir compte de cette corrélation. Lin (1994) démontre comment la matrice de covariance des paramètres du modèle estimé peut être estimée en cas de corrélation intra-individu des périodes, à condition que les périodes provenant d'individus différents soient indépendantes.

3.2 Inférence fondée sur le plan de sondage

Dans une enquête longitudinale à plan de sondage à plusieurs degrés, les événements multiples peuvent être corrélés à divers niveaux : les périodes sont regroupées dans un individu et les individus sont groupés dans les unités de degré d'échantillonnage élevé. La corrélation intra-grappe positive à tout niveau ajoute aux estimations calculées à partir de ce genre de données une variation supplémentaire, outre celle attendue sous des conditions d'indépendance. L'hypothèse d'indépendance des observations lorsque celles-ci sont corrélées dans les grappes donne lieu à une sous-estimation des erreurs-types réelles, ce qui exagère les valeurs des statistiques de test et, en dernière analyse, aboutit au rejet trop fréquent des hypothèses nulles. Donc, pour les périodes multiples chez un individu, où les données proviennent d'une enquête longitudinale, il ne suffit pas de tenir compte simplement de la corrélation intra-individu.

L'estimation de la variance sous le plan de sondage dans le cas de données hiérarchiques corrélées dans les grappes peut être grandement simplifiée quand il est raisonnable de supposer que les individus provenant d'unités primaires d'échantillonnage (UPE) différentes ne sont pas corrélés. Cela équivaut à supposer que les UPE sont échantillonnées avec remise. Cette hypothèse est aussi vérifiée approximativement quand les unités de premier degré sont obtenues par échantillonnage sans remise, à condition que la fraction d'échantillonnage au premier degré soit très faible. Dans ces conditions, une estimation de la variabilité inter-UPE reflète la variabilité entre les unités à tous les degrés d'échantillonnage subséquents, quelle que soit la structure de dépendance entre les observations dans chaque UPE. Pour un résumé récent de l'estimation robuste de la variance dans le cas de données corrélées dans les grappes, voir Williams (2000). Cela implique que l'approche d'estimation robuste de la variance du modèle à période unique dans le cas d'un plan de sondage comportant un échantillonnage avec remise au premier degré décrite par Binder (1992) peut être appliquée directement au cas des périodes multiples, puisqu'elle tient compte de l'effet de la corrélation intra-grappe à tous les niveaux dans chaque UPE.

4. Trois modèles pour les périodes multiples

Afin de tenir compte de covariables ayant des effets différents pour des périodes d'ordres différents, ainsi que de dépendances de durée différentes (risques de base), nous explorons trois modèles pour les périodes multiples. Ces modèles se distinguent par la définition du risque et les hypothèses au sujet du risque de base. Deux de ces modèles tiennent compte de l'ordre des périodes.

Il convient toutefois de souligner que, dans nos travaux, l'ordre des périodes fait uniquement référence à celles survenant durant la période d'observation pour laquelle les données sont recueillies et non pas à la biographie complète d'un individu (à moins que les deux périodes ne coïncident). Par exemple, par première période, nous entendons la première période de l'état étudié survenant durant la période d'observation, alors qu'il pourrait s'agit d'une période d'ordre absolu plus élevé au cours de la vie de la personne. Cette limite implique qu'il convient d'interpréter avec précaution toute incidence que l'ordre de la période peut avoir sur les effets des covariables ou sur la dépendance par rapport au temps.

Modèle 1 : Dans le premier modèle, l'ensemble de risques est défini avec précaution afin de tenir compte de l'ordre des périodes en ce sens qu'un individu ne peut pas être exposé au risque d'achèvement de la deuxième période avant que la première soit achevée, *etc*. Ce modèle, connu sous le nom de modèle d'ensemble de risques conditionnels, a été proposé par Prentice, Williams et Peterson (1981) et revu par Lin (1994). Il a également été discuté par Hamerle (1989) et par Blossfeld et Hamerle (1989) dans le contexte de la modélisation de processus à épisodes multiples. En général, l'ensemble de risques conditionnels au temps t pour l'achèvement d'une période d'ordre j comprend tous les individus qui sont dans leur j^{e} période. Ce modèle permet que l'ordre de la période influence à la fois l'effet des covariables et la forme de la fonction de risque de base.

La fonction de risque pour le i^{e} individu pour la période de j^{e} ordre est

$$h_j(t \mid \mathbf{x}_{ij}(t)) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}_j},$$

où, pour chaque ordre de période, une fonction de risque de base différente et un vecteur de coefficients différent sont permis. Pour ce modèle et pour d'autres que nous examinerons à la présente section, le temps t est mesuré à partir du début de la j^{e} période. Bien que les périodes chez un même individu ne soient pas nécessairement indépendantes, la vraisemblance partielle qui suit reste valide pour l'estimation des $\boldsymbol{\beta}_j$:

$$L(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) = \prod_{j=1}^{K} \prod_{i=1}^{N_j} \left[\frac{e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}}{\sum_{r=1}^{N_j} Y_{rj}(T_{ij})\, e^{\mathbf{x}'_{ij}(T_{ij})\boldsymbol{\beta}_j}} \right]^{\delta_{ij}}. \quad (11)$$

Ici, $T_{1j}, ..., T_{N_j j}$ sont N_j durées de période de j^{e} ordre éventuellement censurées à droite, $\delta_{ij} = 1$ si T_{ij} est une durée observée et $\delta_{ij} = 0$ autrement, et K est l'ordre le plus élevé des périodes qui doivent être incluses dans le modèle de Cox. Au dénominateur, la somme est calculée sur les j^{e} périodes risquant d'être achevées au temps T_{ij}, c'est-à-dire $Y_{rj}(t) = 1$ si $t \le T_{rj}$ et est égale à 0 autrement. Le vecteur de covariables correspondant observé sur $[0, T_{ij}]$ est $\mathbf{x}_{ij}(t)$. La vraisemblance partielle (11) peut être maximisée séparément pour chaque j si aucune contrainte supplémentaire n'est appliquée aux $\boldsymbol{\beta}_j$.

Les équations de score correspondantes qui définissent le paramètre de population finie $\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, ..., \mathbf{B}'_K)'$ sont :

$$U_0(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u_{ij0}(T_{ij}, \mathbf{B}_j) = 0, \quad (12)$$

avec

$$u_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B}_j)}{S^{(0)}(T_{ij}, \mathbf{B}_j)} \right\},$$

et avec $S^{(0)}(t, \mathbf{B}_j)$ et $S^{(1)}(t, \mathbf{B}_j)$ prenant la forme (5) et (6) respectivement, mais avec N_j remplaçant n et $\mathbf{B}_j$ remplaçant $\boldsymbol{\beta}$.

Les estimations sous le plan des paramètres $\mathbf{B}_j$ sont obtenues en résolvant les équations $\sum_{i=1}^{N_j} w_i(s)\hat{u}_{ij0}(T_{ij}, \hat{\mathbf{B}}_j) = 0$ séparément pour chaque j, où $\hat{u}_{ij0}$ a la forme de u_{ij0}, mais avec $S^{(0)}$ et $S^{(1)}$ remplacés par $\hat{S}^{(0)}$ et $\hat{S}^{(1)}$ respectivement. Notons que les poids d'échantillonnage correspondent aux individus et non aux périodes. De même, l'estimation de la matrice de covariance de chaque $\hat{\mathbf{B}}_j$ se fera séparément, en utilisant la méthode d'estimation robuste sous le plan décrite à la section 2.2. Techniquement, il s'agit d'un ensemble d'analyses distinctes selon l'ordre de la période.

Modèle 2 : Le deuxième modèle étudié est le modèle marginal (Wei, Lin et Weissfeld 1989) :

$$h_j(t \mid \mathbf{x}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}},$$

où, pour chaque ordre de période, nous permettons une fonction de risque de base différente tandis que les effets des covariables sont maintenus les mêmes pour différents ordres de période. La fonction de vraisemblance partielle correspondante, ainsi que l'ensemble de risques, sous l'hypothèse que les périodes chez le même individu sont indépendantes, sont les mêmes que pour le modèle 1, avec $\boldsymbol{\beta}$ remplaçant les $\boldsymbol{\beta}_j$. L'équation de score correspondante qui définit le paramètre de population finie est

$$U_0^*(\mathbf{B}) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} u_{ij0}^*(T_{ij}, \mathbf{B}) = 0,$$

avec

$$u_{ij0}^*(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B})}{S^{(0)}(T_{ij}, \mathbf{B})} \right\},$$

où $S^{(0)}(t, \mathbf{B})$ et $S^{(1)}(t, \mathbf{B})$ sont définis par (5) et (6) respectivement, mais avec N_j remplaçant n et $\mathbf{B}$ remplaçant $\boldsymbol{\beta}$.

L'estimation sous le plan du paramètre $\mathbf{B}$ s'obtient en résolvant les équations de score pondérées

$$\sum_{i=1}^{K} \sum_{j=1}^{N_i} w_i(s) \hat{u}^*_{ij0}(T_{ij}, \hat{\mathbf{B}}) = 0,$$

où $\hat{u}^*_{ij0}$ a la forme de u^*_{ij0}, mais avec $S^{(0)}(t, \mathbf{B})$ et $S^{(1)}(t, \mathbf{B})$ remplacés par $\hat{S}^{(0)}(t, \hat{\mathbf{B}})$ et $\hat{S}^{(1)}(t, \hat{\mathbf{B}})$ respectivement.

L'estimation de la matrice de covariance de $\hat{\mathbf{B}}$ sera faite en utilisant la méthode d'estimation robuste sous le plan expliquée à la section 3.2.

Modèle 3 : Le dernier modèle étudié est le suivant :

$$h_j(t \mid \mathbf{x}_{ij}) = \lambda_0(t)\, e^{\mathbf{x}'_{ij}(t)\boldsymbol{\beta}}.$$

Dans ce modèle, nous supposons que les fonctions de risque de base et les effets des covariables sont communs aux divers ordres de période. L'ensemble de risques au temps T_{ij} est défini autrement que dans les modèles 1 et 2, et contient toutes les périodes pour lesquelles $t \le T_{ij}$, en supposant effectivement que toutes les périodes proviennent d'individus différents. Techniquement, ce modèle est un modèle à période unique, de sorte que l'estimation des coefficients et des variances par une méthode robuste sous le plan est simple.

5. Exemple de modélisation de périodes de chômage multiples

5.1 Les données

L'ensemble de données que nous utilisons pour l'illustration provient du premier panel de six ans (1993 à 1998) de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada. Dans ce panel, environ 31 000 personnes sélectionnées dans environ 15 000 ménages ont été suivies pendant six ans par la voie d'interviews annuelles. Certaines personnes sont sorties de l'échantillon au cours du temps pour diverses raisons, tandis que quelques autres, après avoir manqué une ou plusieurs interviews, ont recommencé à participer au panel. Une pondération compliquée des personnes répondant à l'EDTR chaque année tient compte des divers types d'érosion du panel, de sorte que chaque répondant durant une année particulière soit pondéré en fonction de la population de référence pertinente de 1993. Cette approche produit une pondération longitudinale distincte pour chaque cycle (c'est-à-dire année) de collecte de données. Pour la présente analyse, nous avons utilisé les poids longitudinaux provenant de la dernière année du panel, c'est-à-dire 1998, ce qui signifie que seules les données provenant des personnes qui ont répondu au dernier cycle du panel ont été incluses dans les analyses. Un bon résumé des questions relatives au plan d'échantillonnage de l'EDTR est donné dans Lavigne et Michaud (1998). Un examen des questions en rapport avec les études des périodes de chômage basées sur l'EDTR est donné dans Roberts et Kovačević (2001).

L'état d'intérêt est « être en chômage », défini ici comme étant l'état entre une mise à pied permanente dans le contexte d'un emploi à temps plein et le commencement d'un autre emploi à temps plein. Un emploi est « à temps plein » s'il exige au moins 30 heures de travail par semaine. L'évènement d'intérêt est « la sortie du chômage ». Seules les périodes débutant après le 1er janvier 1993 ont été incluses, puisque le 31 janvier 1993 est la date de début des observations auprès du panel. Les périodes de chômage qui n'étaient pas achevées à la fin de la période d'observation (31 décembre 1998) ont été considérées comme étant censurées. Les dénombrements d'échantillon des individus connaissant des périodes admissibles et des périodes en fonction de leur ordre sont présentés au tableau 1. Brièvement, 17 880 périodes ont été dénombrées auprès de 8 401 membres du panel longitudinal. Environ la moitié des individus échantillonnés (4 260) devenus chômeurs durant cette période ont connu deux périodes de chômage ou plus. En tout, 3 809 périodes sont demeurées inachevées à cause de la cessation du panel.

Sur une longue liste de covariables disponibles, nous n'en n'avons choisies que dix. La variable de sexe [SEX] de l'individu longitudinal est la seule qui demeure constante au cours des diverses périodes. Quatre variables ont des valeurs enregistrées à la fin de l'année durant laquelle la période a commencé, à savoir le niveau de scolarité [EDUCLEV] avec quatre catégories (faible, moyen-faible, moyen, élevé), l'état matrimonial [MARST] avec trois catégories (célibataire, marié(e)/union de fait, autre), le revenu familial par personne (en dollars canadiens) avec quatre catégories (<10 000, de 10 000 à 20 000, de 20 000 à 30 000, 30 000 et plus) et l'âge [AGE] (en années). Trois variables ont les valeurs correspondant à l'emploi avec mise à pied qui a précédé la période, à savoir le type de fin d'emploi [TYPJBEND] avec deux catégories (congédié(e) et départ volontaire), la profession [OCCUPATION] avec six catégories (professionnel, administration, secteur primaire, fabrication, construction et autre) et la taille de l'entreprise [FIRMSIZE] avec cinq catégories (<20, de 20 à 99, de 100 à 499, de 500 à 999, et 1 000 et plus employés). Deux variables binaires représentent la situation durant la période, à savoir travailler à temps partiel [PARTTJB] et être aux études [ATSCH].

L'ensemble de données a été préparé selon le mode de « dénombrement » où chaque individu présentant des

périodes admissibles est représenté par un ensemble de lignes dont chacune correspond à une période. Bien qu'une ligne contienne le temps d'entrée dans la période t_1 et le temps de sortie de la période t_2 ou le temps de censure t_c, pour l'analyse, la durée est toujours considérée sous la forme $(0, t_2 - t_1)$ ou $(0, t_c - t_1)$. Les covariables examinées sont reliées à chaque ligne. Sont également reliés à chaque ligne le poids longitudinal de 1998 et les identificateurs de la strate et de l'UPE de la personne dont la période est décrite par l'enregistrement en question.

5.2 Analyse

Pour les besoins de l'illustration, nous avons limité l'analyse aux quatre premières périodes, si bien que tous les individus échantillonnés présentant des périodes admissibles sont inclus dans l'analyse, mais que les enregistrements de période survenue après la quatrième ne sont pas pris en considération à cause de leur faible nombre dans l'échantillon.

Nous avons estimé les coefficients et leurs variances pour les trois modèles par les méthodes fondées sur le plan de sondage décrites à la section 4 en utilisant la procédures « SURVIVAL » du logiciel SUDAAN version 8. Pour chacun des trois modèles, nous avons spécifié un plan de sondage stratifié avec tirage des UPE avec remise (c'est-à-dire DESIGN = WR). Les trois modèles ont été ajustés au même nombre de périodes (16 307). Pour chaque modèle, nous avons ensuite calculé les fonctions de risque de base cumulatif empiriques en utilisant une approche produit-limite (voir Kalbfleisch et Prentice (2002), pages 114-116) telle qu'elle est implémentée dans la procédure SURVIVAL de SUDAAN.

L'approche robuste sous le modèle pour les périodes multiples décrites à la section 3.1 comporte un ajustement des estimations de la variance en vue de tenir compte de l'interdépendance éventuelle des périodes chez un même individu, en supposant que les périodes provenant d'individus différents sont indépendantes; cependant, dans cette approche, aucune mesure n'est prise pour tenir compte des probabilités inégales de sélection des individus échantillonnés, que ce soit dans les estimations des coefficients ou les estimations de la variance. Afin de le faire, pour les modèles 1 et 2, nous avons également utilisé la procédure SURVIVAL de SUDAAN version 8 pour estimer les variances des estimations pondérées des coefficients, où nous avons émis l'hypothèse d'indépendance inter-individus des périodes, mais avons tenu compte d'une corrélation intra-individu éventuelle des périodes. Pour cela, nous avons spécifié un plan d'échantillonnage non stratifié avec sélection des grappes avec remise en précisant que chaque individu formait sa propre grappe. Les hypothèses de dépendance sont les mêmes que celles utilisées par Lin (1994), mais nous avons tenu compte de l'utilisation de pondérations dans l'estimation des coefficients et des variances. Nous appellerons ces estimations des variances « estimations robustes modifiées de la variance sous le modèle des estimations pondérées des coefficients ».

5.3 Certaines statistiques descriptives

La durée moyenne estimée d'une période achevée est de 33,3 semaines, tandis que la durée moyenne estimée de la partie observée d'une période censurée (inachevée) est de 48,5 semaines.

Tableau 1 Dénombrement des individus du panel de six ans de l'EDTR ayant des périodes de chômage débutant entre janvier 1993 et décembre 1998, selon le nombre total de périodes et l'ordre de la période (A-achevée, I-inachevée)

Individus selon le nombre de périodes		Périodes par ordre									
		Première		Deuxième		Troisième		Quatrième		5e +	
		A	I	A	I	A	I	A	I	A	I
1 période	4 141	2 221	1 920	-	-	-	-	-	-	-	-
2 périodes	1 915	1 915	-	1 154	761	-	-	-	-	-	-
3 périodes	1 044	1 044	-	1 044	-	612	432	-	-	-	-
4 périodes	629	629	-	629	-	629	-	348	281	-	-
5 périodes et plus	672	672	-	672	-	672	-	672	-	1 158	415
Total	8 401	6 481	1 920	3 499	761	1 913	432	1 020	281	1 158	415

L'examen visuel des fonctions de survie estimées de Kaplan-Meier (non présentées) pour les périodes de chaque ordre révélait que, à mesure que croissait l'ordre, la valeur de la fonction de survie à n'importe quel temps fixe t diminuait, indiquant que les premières périodes sont les plus longues parmi les périodes achevées et que la durée d'une période multiple est d'autant plus courte que son ordre est élevé. Cette constatation est vraisemblablement une conséquence de la durée de vie limitée du panel, en ce sens qu'un individu comptant un plus grand nombre de périodes durant l'intervalle de temps de six ans donné est susceptible de présenter des périodes plus courtes.

5.4 Ajustement des modèles selon une approche fondée sur le plan

Comme nous l'avons mentionné plus haut, notre exemple est simplement une illustration de l'approche fondée sur le plan de sondage d'ajustement des modèles à risques proportionnels à des données sur des événements multiples provenant d'une enquête à plan de sondage complexe. Donc, nous consacrons peu de temps ici à discuter de la façon d'évaluer l'adéquation de ce ces modèles, dont l'adéquation des hypothèses de proportionnalité dans chacun des modèles ou le fait de savoir si un type de modèle est aussi bien ajusté qu'un autre.

Les coefficients estimés d'après l'ajustement des trois modèles aux données de l'EDTR sont présentés au tableau 2. Les valeurs des coefficients significatives au seuil de 5 % sur la base de tests t individuels sont en caractères gras.

Le modèle 1 est conditionnel à l'ordre de la période et comprend l'ajustement de quatre modèles distincts aux données provenant des quatre ordres de période différents. Comme le montre le tableau 2, les coefficients des variables SEXE, AGE et au moins une catégorie de la variable de revenu familial sont significatifs pour les périodes de tous ordres, quoique leur grandeur estimée varie en fonction de l'ordre de la période. Les coefficients estimés pour AGE sont négatifs, mais diminuent de grandeur à mesure que l'ordre de la période augmente, tandis qu'aucune tendance n'est discernable pour les coefficients estimés pour les deux autres variables. Les variables EDUCLEV, PARTJB et ATSCH ont des coefficients significatifs pour les périodes d'ordre 1, 2 et 3, mais non pour les périodes d'ordre 4. Ce résultat peut être attribué, en partie, à la petite taille d'échantillon pour les quatrièmes périodes. Pour chacune des trois autres variables du modèle (MARST, OCCUPATION et FIRMSIZE), un coefficient n'était significatif que pour un seul ordre de période.

Pour le modèle 2, les coefficients sont contraints d'être les mêmes pour tous les ordres de période. Comme le montre le tableau 2, numériquement, un grand nombre de valeurs des coefficients estimés, mais pas toutes, sont comprises entre les estimations calculées pour la première et pour la deuxième période à l'aide du modèle 1, ce qui pourrait être dû au fait qu'une proportion élevée de l'échantillon correspond aux événements de ces ordres. Toutes les variables, sauf OCCUPATION, ont un coefficient significatif. Les erreurs-types des coefficients sont plus faibles dans le cas du modèle 2 que dans celui du modèle 1.

Le modèle 3 est un modèle à période unique avec un seul ensemble de coefficients et une seule fonction de risque de base. Les coefficients estimés du modèle sont semblables aux estimations obtenues à l'aide du modèle 2.

Les fonctions de risque de base cumulatif estimées pour les modèles 1 à 3 sont illustrées aux figures 1 à 3 respectivement. Dans tous les cas, pour les durées allant jusqu'à environ 50 semaines, les fonctions ont une forme concave, ce qui sous-entend qu'il existe une dépendance positive par rapport au temps du taux de sortie (autrement dit, la probabilité de sortie est d'autant plus élevée que la période est longue). Pour les durées de plus de 50 semaines, la forme devient convexe, ce qui suggère une dépendance négative par rapport au temps dans le cas des périodes plus longues. À la figure 1, la position de la fonction de risque de base cumulatif estimée varie selon l'ordre de la période, la courbe pour les périodes d'ordre 1 étant la plus élevée et celle pour les périodes d'ordre 4, la plus basse. À la figure 2, pour le modèle 2, les positions des diverses courbes ne suivent pas l'ordre des périodes. Cette différence observée entre les figures 1 et 2 pourrait servir de diagnostic visuel indiquant qu'une étude plus approfondie est nécessaire afin d'évaluer lequel des modèles 1 ou 2 est un meilleur descripteur des données, puisque les coefficients estimés ont une incidence sur les risques de base estimés.

5.5 Comparaison aux estimations robustes modifiées de la variance sous le modèle

Comme il est décrit à la section 5.2, les estimations robustes modifiées de la variance sous le modèle tiennent compte de la corrélation éventuelle entre les périodes chez un même individu, sous l'hypothèse d'indépendance entre les individus. La comparaison, pour les modèles 1 et 2, des estimations des erreurs-types obtenues par cette approche aux estimations des erreurs-types sous le plan de sondage n'a révélé que des écarts faibles. Il semble donc que les estimations sous le plan reflètent toute corrélation entre les périodes chez un même individu et qu'il n'y ait pas de dépendance supplémentaire au-delà du niveau individuel dans notre exemple.

Tableau 2 Coefficients β estimés pour les trois modèles

	Modèle 1				Modèle 2	Modèle 3
	Ordre 1	**Ordre 2**	**Ordre 3**	**Ordre 4**		
SEX (F)						
M	**0,4417**	**0,3781**	**0,3299**	**0,4435**	**0,4049**	**0,4090**
EDUCLEV (É)						
F	**-0,4561**	**-0,5234**	**-0,3748**	-0,1065	**-0,4128**	**-0,4331**
MF	**-0,2330**	**-0,2700**	**-0,3310**	-0,1653	**-0,2436**	**-0,2474**
M	-0,0744	-0,1060	-0,1156	0,0668	-0,0684	-0,0671
MARST (M)						
Célibataire	**-0,1142**	-0,1290	-0,0622	-0,1375	**-0,1357**	**-0,1330**
Autre	0,0985	-0,0894	0,1124	-0,1072	0,0328	0,0401
TYPJBEND (Congédié(e))						
Départ volontaire	0,0704	**0,2752**	**0,4207**	**0,3413**	**0,1579**	**0,1284**
OCCUPATION (Autre)						
Professionnels	0,1592	-0,1364	-0,1388	0,0903	0,0490	0,0485
Administration	-0,0265	**-0,2930**	-0,1769	0,0579	-0,0971	-0,0938
Secteur primaire	-0,0211	-0,2175	-0,1187	0,2032	-0,0410	-0,0201
Fabrication	-0,0003	-0,0994	-0,1295	0,2862	-0,0093	-0,0088
Construction	0,1290	-0,1862	-0,0879	0,2339	0,0490	0,0813
FIRMSIZE (1 000+)						
<20	-0,0027	-0,0097	0,1005	**0,4403**	0,0441	0,0408
20 à 99	0,0358	0,0881	0,0815	0,3999	**0,0928**	**0,0951**
100 à 499	0,0436	-0,0905	0,0328	0,0257	0,0214	0,0278
500 à 999	-0,0006	0,0153	-0,0623	-0,0067	-0,0005	0,0020
PARTTJB (Non)						
Oui	**-0,2903**	**-0,5414**	**-0,5109**	-0,1407	**-0,3693**	**-0,3743**
ATSCH (Non)						
Oui	**-1,0832**	**-1,1516**	**-1,2956**	-1,3541	**-1,1205**	**-1,1266**
Revenu familial par personne (10 000 $ et moins)						
10 000 $ à 20 000 $	**0,1294**	**0,1802**	0,0692	**0,1117**	**0,1345**	**0,1330**
20 000 $ à 30 000 $	**0,1644**	**0,3611**	0,1572	**0,4900**	**0,2241**	**0,2141**
30 000 $ et plus	**0,1712**	**0,3916**	**0,3005**	**0,4241**	**0,2280**	**0,2115**
AGE	**-0,0491**	**-0,0311**	**-0,0269**	**-0,0207**	**-0,0424**	**-0,0435**
Périodes dans l'ensemble de risques	8 386	4 255	2 345	1 300	16 286	16 286
Censurées	1 913	759	432	281	3 385	3 385
Achevées	6 473	3 496	1 913	1 019	12 901	12 901

Les valeurs significatives au seuil de signification de 5 % sont en caractères gras.

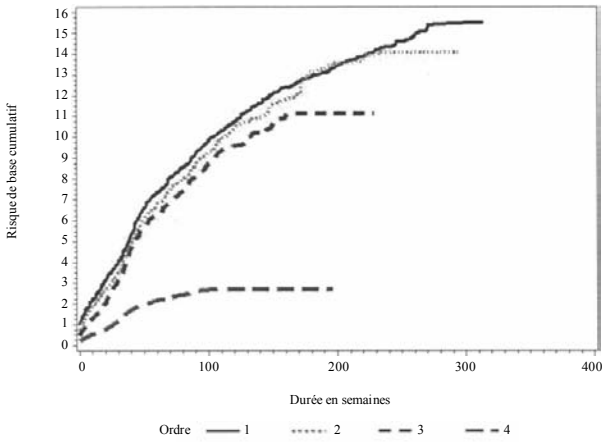


Figure 1 Risque de base cumulatif – Modèle 1

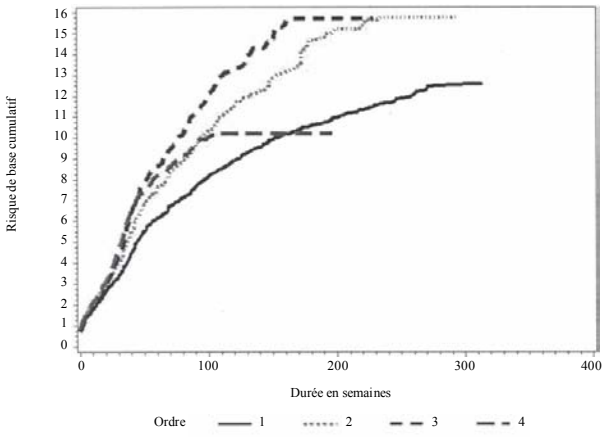


Figure 2 Risque de base cumulatif – Modèle 2

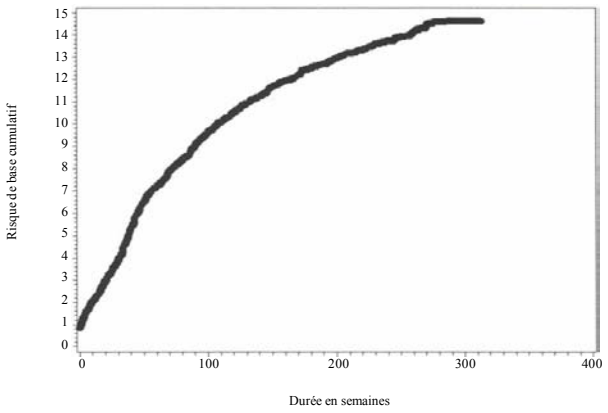


Figure 3 Risque de base cumulatif – Modèle 3

6. Conclusion

Nous avons étudié le problème de l'analyse de périodes multiples en considérant deux approches générales pour traiter le manque d'indépendance entre les temps de sortie, à savoir une approche robuste basée sur un modèle et une approche basée sur le plan de sondage. La première consiste à estimer les paramètres du modèle en supposant que les périodes sont indépendantes, puis à corriger la matrice de covariance naïve de façon à tenir compte des dépendances intra-individu postulées par le chercheur. Cette approche ne tient pas compte de la mise en grappes éventuelle entre individus (ou, en fait, de toute mise en grappes qui pourrait avoir lieu à un niveau d'agrégation plus élevé que l'individu) ni des probabilités inégales de sélection des individus (quoique, dans notre exemple, nous avons montré comment la méthode pourrait être étendue afin d'inclure les poids de sondage). La deuxième approche définit les coefficients du modèle comme des paramètres de population finie. Ces paramètres sont ensuite estimés en tenant compte des probabilités de sélection éventuellement inégales des individus. Une méthode d'estimation de la variance sous le plan de sondage qui tient compte des corrélations éventuelles entre individus dans la même UPE rend compte automatiquement des dépendances non spécifiées des périodes à des niveaux inférieurs à l'UPE, comme les dépendances intra-individu. Dans le cas des échantillons de grande taille, cette inférence sous le plan de sondage s'étend directement à la super-population à partir de laquelle la population finie a été hypothétiquement générée. Le défaut de la première approche est qu'elle ignore totalement la possibilité d'une mise en grappes entre individus. Un inconvénient éventuel de la deuxième approche, comme nous l'avons appliquée, est qu'elle repose sur l'hypothèse de l'échantillonnage avec remise des UPE comprenant les individus. Les deux approches coïncident dans le cas de l'échantillonnage aléatoire simple d'individus, où, dans le cas de l'approche robuste fondée sur un modèle, la dépendance entre les périodes chez un même individu est postulée explicitement et prise en compte dans la formule d'estimation de la variance et où, dans l'approche fondée sur le plan de sondage, les périodes chez un même individu sont traitées comme une grappe dans l'estimation de la variance sous le plan de sondage.

Nous avons appliqué l'approche fondée sur le plan de sondage à trois modèles de type à risques proportionnels. L'un tenait compte de risques de base non spécifiés différentiels et de coefficients différents pour chaque ordre de période. Le deuxième tenait encore compte de risques de base non spécifiés différentiels pour divers ordres de période, mais exigeait que les coefficients soient les mêmes pour tous les ordres. Le troisième était un simple modèle à période unique. Nous avons constaté que la façon dont l'information sur l'ordre de la période était utilisée avait une incidence sur les résultats de l'ajustement de nos modèles. Une comparaison visuelle des estimations des coefficients et des estimations des risques de base cumulatifs pour les modèles 1 et 2 indiquait des résultats différents. Comme l'a suggéré l'un des examinateurs, il serait bon d'élaborer un test formel en vue de déterminer si les coefficients diffèrent effectivement en fonction de l'ordre de la période (comme dans le modèle 1), étant donné des risques de base pouvant varier selon l'ordre de la période. Il est, en fait, facile de produire un tel test, de la façon suivante. Soit $\boldsymbol{\gamma} = (\mathbf{B}_1', \mathbf{B}_2', ..., \mathbf{B}_K')'$ le vecteur des K vecteurs de coefficients du modèle 1, où chacun est de longueur p, et soit $\mathbf{z}_{ij}(t) = (0', 0', ..., \mathbf{x}_{ij}(t)', 0', ..., 0')'$ le vecteur de longueur pK pour la j^{e} période du i^{e} individu où la j^{e} composante de ce vecteur contient le vecteur des covariables $\mathbf{x}_{ij}(t)$. Alors, le modèle 1 peut être exprimé par

$$h_j(t \mid \mathbf{z}_{ij}(\mathrm{t})) = \lambda_{0j}(t)\, e^{\mathbf{z}_{ij}'(t)\boldsymbol{\gamma}},$$

qui a la forme générale des risques de base variant avec l'ordre de la période, mais ayant un vecteur de coefficients fixe. Un test de la convergence des coefficients se rapportant à chaque ordre de période, c'est-à-dire $H_0 : \mathbf{B}_1 = \mathbf{B}_2 = ... \mathbf{B}_K$ équivaut à tester $H_0 : \mathbf{C}\boldsymbol{\gamma} = 0$ où C est la matrice $\mathbf{C} = I_p \otimes [1_{K-1} \; -I_{K-1}]$ de dimensions $(K-1)p \times Kp$. Étant donné une estimation $\hat{\boldsymbol{\gamma}}$ de $\boldsymbol{\gamma}$ et une estimation $\hat{V}(\hat{\boldsymbol{\gamma}})$ de la matrice de covariance de $\hat{\boldsymbol{\gamma}}$, obtenue comme il est décrit à la section 4 pour le modèle 2, une statistique de Wald peut être calculée pour tester l'hypothèse. Si l'hypothèse n'est pas rejetée, on peut conclure qu'un modèle à coefficients constants sur l'ordre de la période (mais à risques de base variables en fonction de l'ordre de la période) semble être aussi bien ajusté aux données qu'un modèle où les risques de base et les coefficients varient en fonction de l'ordre de la période. D'autres mesures de l'adéquation du modèle devraient également être faciles à élaborer dans le cadre d'estimations sous le plan de sondage.

Nous avons aussi comparé visuellement, dans le cas de notre exemple, les estimations des erreurs-types des coefficients obtenues sous le plan de sondage (en tenant compte de la mise en grappes au niveau de l'UPE et à des niveaux inférieurs), ainsi que sous une modification de l'approche robuste fondée sur un modèle (tenant compte de la mise en grappes au niveau de l'individu ou à un niveau inférieur) pour les modèles 1 et 2. Nous n'avons observé que des différences faibles, qui indiquaient l'absence d'effets de grappe à un niveau d'agrégation supérieur au niveau de l'individu pour les données en question. Nous avons également calculé les estimations des erreurrs-types en supposant que les périodes chez une même personne étaient indépendantes et de nouveau constaté des différences faibles seulement par rapport à celles obtenues suivant l'approche fondée sur le plan. Il semble donc que, pour l'exemple choisi, la dépendance inter-périodes soit faible. Cependant, en général, nous estimons qu'une approche fondée sur le plan de sondage protège contre l'omission de toute dépendance non postulée au niveau de l'UPE ou à un niveau inférieur dans les estimations de la variance.

Remerciements

Nous remercions Normand Laniel et Xuelin Zhang de leurs commentaires constructifs concernant une version antérieure du manuscrit. Nous remercions également le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions qui ont accru considérablement la lisibilité du manuscrit.

Bibliographie

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-291.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika,* 79, 139-147.

Blossfeld, H.-P., et Hamerle, A. (1989). Using Cox models to study multiepisode processes. *Sociological Methods and Research*, 17, 4, 432-448.

Clayton, D., et Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (avec discussion). *Journal of the Royal Statistical Society*, Séries A, 1985, 148, 82-117.

Cochran, W.G. (1977). *Sampling Techniques*. Deuxième édition. New York : John Wiley & Sons, Inc.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Hamerle, A. (1989). Multiple-spell regression models for duration data. *Applied Statistics*, 38, 1, 127-138.

Heckman, J., et Singer, B. (1982). Population heterogeneity in demographic models. Dans *Multidimensional Mathematical Demography*, (Éds., K. Land et A. Rogers), New York : Academic Press, 567-599.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55, 1, 13-22.

Kalbfleisch, J.D., et Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2ième Édition, New York : John Wiley & Sons, Inc.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939-956.

Lavigne, M., et Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Document de travail, Statistique Canada, 75F0002M No. 98-05.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 13, 2233-2247.

Lin, D.Y. (2000). On Fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.

Lin, D.Y., et Wei, L.J. (1989). The robust Inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Prentice, R.L., Williams, B.J. et Peterson, A.V. (1981). On the regression analysis of multivariate failure data. *Biometrika*, 68, 373-379.

Roberts, G., et Kovačević, M. (2001). New research problems in analysis of duration data arising from complexities of longitudinal surveys. *Proceedings of the Survey Methods Section of the Statistical Society of Canada,* 111-116.

Wei, L.J., Lin, D.Y. et Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 27-40
Statistique Canada, N° 12-001 au catalogue

Réduction bayésienne des poids pour les modèles de régression linéaire généralisée

Michael R. Elliott [1]

Résumé

Dans les sondages où les unités ont des probabilités inégales d'inclusion dans l'échantillon, les associations entre la probabilité d'inclusion et la statistique d'intérêt peuvent causer un biais. Des poids égaux à l'inverse de la probabilité d'inclusion sont souvent utilisés pour neutraliser ce biais. Les plans de sondage fortement disproportionnels comportent des poids de valeur élevée qui peuvent introduire une variabilité indésirable dans les statistiques telles que l'estimateur de la moyenne de population ou l'estimateur de la pente de la régression de population. La réduction des poids consiste à modifier ceux dont la valeur est élevée à une valeur seuil fixe et à ajuster ceux inférieurs à cette valeur de façon à ce que la somme de ces poids réduits demeure égale à la somme des poids non réduits, ce qui réduit la variabilité au prix de l'introduction d'un certain biais. La plupart des approches ordinaires sont ponctuelles en ce sens qu'elles n'utilisent pas les données en vue d'optimiser le compromis entre le biais et la variance. Les approches dictées par les données qui sont décrites dans la littérature sont un peu plus efficaces que les estimateurs entièrement pondérés. Dans le présent article, nous élaborons des méthodes bayésiennes de réduction des poids d'estimateurs par la régression linéaire et par la régression linéaire généralisée sous des plans de sondage avec probabilités d'inclusion inégales. Nous décrivons une application à l'estimation du risque de blessure chez les enfants installés sur le siège arrière dans les camionnettes compactes à cabine allongée à l'aide des données de la Partners for Child Passenger Safety surveillance survey.

Mots clés : Sondage; poids de sondage; winsorisation des poids; inférence bayésienne à la population; lissage des poids; modèles linéaires généralisés mixtes.

1. Introduction

Lors de l'analyse de données provenant d'échantillons sélectionnés avec probabilités différentielles, on utilise souvent comme poids des cas les inverses des probabilités d'inclusion afin de réduire ou d'éliminer le biais dans les estimateurs des quantités de population d'intérêt. Le remplacement des moyennes et des totaux implicites dans les statistiques par leurs équivalents pondérés par les poids des cas produit des estimateurs linéaires sans biais et des estimateurs non linéaires asymptotiquement sans biais des valeurs de population (Binder 1983). Les poids des cas peuvent aussi intégrer des ajustements pour la non-réponse, qui habituellement sont égaux à l'inverse de la probabilité estimée de réponse (Gelman et Carlin 2002, Oh et Scheuren 1983), ou des ajustements par calage, qui contraignent les poids des cas à être égaux à des totaux connus de pondération, soit conjointement, comme dans la poststratification ou l'estimation par la régression généralisée, soit aux marges, comme dans l'estimation par calage sur marges (raking) généralisé (Deville et Särndal 1992, Isaki et Fuller 1982).

L'utilisation des poids de sondage pour la production de statistiques descriptives, comme les moyennes et les totaux, d'après des plans avec probabilités d'inclusion inégales ne suscite guère de débat. Cependant, lorsqu'il s'agit d'estimer des quantités « analytiques » (Cochran 1977, page 4) axées sur les associations entre, par exemple, les facteurs de risque et les résultats en matière de santé au moyen de modèles linéaires ou linéaires généralisés, la décision d'utiliser les poids de sondage est moins catégorique (voir Korn et Graubard 1999, pages 180-182). Dans des conditions de régression, des différences entre les estimateurs pondéré et non pondéré de la pente de la droite de régression peuvent survenir parce que le modèle de données est spécifié incorrectement ou qu'il existe une association entre les erreurs résiduelles et (ou) la probabilité d'inclusion (l'échantillon est informatif). Si le modèle de données est spécifié incorrectement, une option consiste à améliorer la spécification du modèle. Cependant, il est parfois difficile de déterminer la forme fonctionnelle exacte; ou bien, il se peut que l'erreur de spécification soit très modeste, mais que le plan de sondage l'amplifie, ou bien, une approximation du modèle réel pourrait être souhaitée pour simplifier l'explication (approximation linéaire d'une tendance quadratique). Dans le cas de l'échantillonnage informatif ou non ignorable, les poids de sondage peuvent être nécessaires pour obtenir des estimateurs convergents des paramètres de régression (Korn et Graubard 1995). De manière plus formelle, les estimateurs entièrement pondérés des paramètres de régression sont des estimateurs du « pseudo-maximum de vraisemblance » (EPMV) (Binder 1983, Pfeffermann 1993) en ce sens qu'ils sont « convergents sous le plan » pour les EMV qui résoudraient les équations de

1. Michael R. Elliott est professeur adjoint au département de biostatistique, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI. Courriel : mrelliot@umich.edu.

score pour les paramètres de régression sous le modèle de régression de superpopulation hypothétique si nous disposions de données observées pour l'ensemble de la population. La convergence par rapport au plan implique que la différence entre la quantité cible de population et l'estimation calculée d'après l'échantillon tend vers zéro quand la taille de l'échantillon et la taille de la population augmentent concomitamment, ou que les différences tendent, en moyenne, vers zéro par échantillonnage répété de la population, où les échantillons sont sélectionnés de manière identique à partir de $t \to \infty$ répliques de la population : voir Särndal (1980) ou Isaki et Fuller (1982). Si les observations sont en grappes, plus de soins doivent être mis à élaborer des estimateurs EPMV convergents sous le plan, quoique les plans hiérarchiques à plusieurs degrés permettent d'approximer les estimations de la log-vraisemblance dans des conditions de recensement à l'aide d'équations de score pondérées si l'on veille à tenir compte du fait que les tailles d'échantillons intra-grappe sont habituellement faibles et le demeurent même si le nombre de grappes augmente (Pfeffermann, Skinner, Holmes, Goldstein et Rabash 1998, Korn et Graubard 2003).

Bien que les EPMV soient populaires, à cause de leur convergence sous le plan, cette propriété est obtenue au prix d'un accroissement de la variance. Cet accroissement peut engloutir la réduction du biais, de sorte qu'effectivement, l'EQM augmente dans une analyse pondérée. Cela risque surtout de se produire si a) la taille d'échantillon est faible, b) les écarts entre les probabilités d'inclusion sont grands ou c) le modèle est approximativement correctement spécifié et l'échantillonnage est approximativement non informatif. L'approche qui est peut-être la plus courante pour traiter ce problème est la réduction des poids (Potter 1990, Kish 1992, Alexander, Dahl et Weidman 1997), qui consiste à fixer la valeur des poids plus grands qu'une certaine valeur w_0 à cette valeur w_0. Habituellement, w_0 est choisie de manière ponctuelle - disons égale à trois ou à six fois le poids moyen - sans se soucier de savoir si le seuil de réduction choisi est optimal en ce qui concerne l'EQM. Donc, le biais est introduit pour réduire la variance, avec l'objectif d'une réduction globale de l'EQM.

D'autres méthodes fondées sur le plan de sondage ont été envisagées dans la littérature. Potter (1990) discute de méthodes systématiques en vue de choisir w_0, y compris des méthodes de distribution des poids et de réduction de l'EQM. La technique de distribution des poids consiste à supposer que les poids suivent une loi bêta inverse et à l'échelle; les paramètres de la loi bêta inverse sont estimés à l'aide d'estimateurs par la méthode des moments, et les poids provenant de la queue supérieure de la distribution, disons où $1 - F(w_i) < 0,01$, sont réduits à la valeur w_0 telle que $1 - F(w_0) = 0,01$. La méthode de réduction de l'EQM consiste à déterminer l'EQM empirique au niveau de réduction w_t, où le poids réduit est $w_i^* = w_t\, I(w_i \geq w_t) + w_i\, I(w_i < w_t)$, $i = 1, ..., n$ sous l'hypothèse que l'estimation entièrement pondérée est sans biais pour la moyenne réelle. En pratique, on envisage une série de niveaux de réduction $t = 1, ..., T$, où $t = 1$ correspond aux données non pondérées $(w_1 = \min_i(w_i))$ et $t = T$, aux données entièrement pondérées $(w_T = \max_i(w_i))$, et $\hat{\theta}_t$ est la valeur de la statistique en utilisant les poids réduits au niveau *t*. Le niveau de réduction choisi est alors donné par $w_0 = w_{t^*}$, où $t^* = \text{argmin}_t(\text{EQM}_t)$ pour $\text{EQM}_t = (\hat{\theta}_t - \hat{\theta}_T)^2 + \hat{V}(\theta_t)$.

La littérature sur le calage décrit des techniques qui ont été mises au point en vue de permettre que des ajustements par poststratification généralisée ou calage sur marges (raking) généralisé soient bornés pour éviter la construction de poids extrêmes (Deville et Särndal 1992, Folsom et Singh 2000). Beaumont et Alavi (2004) étendent cette notion à l'élaboration d'estimateurs axés sur la réduction des poids de valeur élevée d'observations fortement influentes ou aberrantes. Bien que l'imposition de ces bornes réduit les poids extrêmes pour les ramener à une valeur seuil fixe, le choix de ce seuil demeure arbitraire.

Une autre approche des méthodes de réduction direct des poids a été élaborée dans la littérature sur l'inférence bayésienne en population finie (Elliott et Little 2000, Holt et Smith 1979, Ghosh et Meeden 1986, Little 1991, 1993, Lazzeroni et Little 1998, Rizzo 1992). Ces approches tiennent compte des probabilités d'inclusion inégales en considérant les poids des cas comme des variables de stratification à l'intérieur des strates définies par la probabilité d'inclusion. Ces « strates d'inclusion » peuvent correspondre aux strates formelles définies par un plan de sondage disproportionnel stratifié, ou peuvent être des « pseudo-strates » fondées sur des poids regroupés dérivés de la sélection, de la poststratification et (ou) des redressements pour la non-réponse. Des estimations pondérées standard sont alors obtenues quand les moyennes de strate de poids des résultats de l'enquête sont traitées comme des effets fixes et que la réduction des poids est réalisé en considérant les moyennes de strate de poids sous-jacentes comme des effets aléatoires. Ces méthodes tiennent compte de la présence éventuelle de données « partiellement pondérées » qui utilisent les données proprement dites pour moduler comme il convient le compromis entre le biais et la variance, et permettent aussi que l'estimation et l'inférence à partir de données recueillies sous des plans de sondage avec probabilités d'inclusion inégales soit fondées sur des modèles utilisés dans d'autres domaines d'estimation et d'inférence statistiques.

Le présent article étend ces modèles à effets aléatoires, que nous appelons modèles « de lissage des poids », afin

d'inclure l'estimation des paramètres de population des modèles linéaires ainsi que linéaires généralisés. À la section 2, nous passons brièvement en revue l'inférence bayésienne en population finie, formalisons les concepts de mécanismes d'échantillonnage ignorable et non ignorable, et développons les modèles de lissage des poids pour les modèles de régression linéaire et de régression linéaire généralisée dans un cadre entièrement bayésien. À la section 3, nous présentons les résultats de simulations effectuées pour examiner les propriétés sous rééchantillonnage des estimateurs par lissage des poids des paramètres de régression linéaire et logistique sous un plan de sondage disproportionnel stratifié, et nous les comparons à celles des estimateurs fondés sur le plan de sondage standard. À la section 4, nous illustrons l'utilisation des estimateurs par lissage des poids dans une analyse du risque de blessure chez les enfants dans les collisions impliquant des véhicules transportant des passagers. À la section 5, nous résumons les résultats des simulations et considérons des extensions à des plans de sondage plus complexes.

2. Inférence bayésienne en population finie

Représentons les données de population pour une population comptant $i = 1, ..., N$ unités par $Y = (y_1, ..., y_N)$, avec les vecteurs de covariables connexes $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ et la variable indicatrice d'échantillonnage $I = (I_1, ..., I_N)$, où $I_i = 1$ si le i^{e} élément est échantillonné et 0 autrement. Comme l'inférence fondée sur le plan de sondage, l'inférence bayésienne se concentre sur les quantités étudiées de population $Q(Y)$, comme les moyennes de population $Q(Y) = \overline{Y}$ ou les paramètres de régression par les moindres carrés de population $Q(Y, X) = \min_{B_0, B_1} \sum_{i=1}^{N} (y_i - B_0 - B_1 x_i)^2$. Contrairement à l'inférence fondée sur le plan, mais en accordance avec la plupart des autres domaines de la statistique, on postule pour les données de population Y un modèle ayant la forme d'une fonction des paramètres θ : $Y \sim f(Y|\theta)$. L'inférence au sujet de $Q(Y)$ est faite en se fondant sur la loi prédictive a posteriori de $p(Y_{\text{nob}} | Y_{\text{obs}}, I)$, où Y_{nob} consiste en les éléments de Y_i pour lesquels $I_i = 0$:

$$p(Y_{\text{nob}} | Y_{\text{obs}}, I) = \frac{\int\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi)\, p(I | Y, \theta, \phi)\, p(Y_{\text{obs}} | \theta)\, p(\theta, \phi)\, d\theta\, d\phi}{\int\int\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi)\, p(I | Y, \theta, \phi)\, p(Y_{\text{obs}} | \theta)\, p(\theta, \phi)\, d\theta\, d\phi\, dY_{\text{nob}}} \quad (1)$$

où $p(I | Y, \theta, \phi)$ modélise l'indicateur d'inclusion.

Si nous supposons que ϕ et θ sont indépendants a priori et si la loi de l'indicateur d'échantillonnage I est indépendante de Y, le plan d'échantillonnage est dit « non confondu » ou « non informatif »; si la loi de I dépend uniquement de Y_{obs}, alors le mécanisme d'échantillonnage est dit « ignorable » (Rubin 1987), ce qui équivaut à la terminologie standard des données manquantes (les éléments non observés de la population peuvent être considérés comme manquants par conception). Sous des plans d'échantillonnage ignorables, $p(\theta, \phi) = p(\theta)p(\phi)$ et $p(I | Y, \theta, \phi) = p(I | Y_{\text{obs}}, \phi)$, et donc (1) se réduit à

$$\frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta)\, p(Y_{\text{obs}} | \theta)\, p(\theta)\, d\theta}{\int\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta)\, p(Y_{\text{obs}} | \theta)\, p(\theta)\, d\theta\, dY_{\text{nob}}} = p(Y_{\text{nob}} | Y_{\text{obs}}), \quad (2)$$

ce qui permet de faire l'inférence au sujet de $Q(Y)$ sans modéliser explicitement le paramètre I d'inclusion dans l'échantillonnage (Ericson 1969, Holt et Smith 1979, Little 1993, Rubin 1987, Skinner, Holt et Smith 1989). Les plans d'échantillonnage non informatifs représentent un cas particulier des plans d'échantillonnage ignorables équivalant aux mécanismes de création de données manquant entièrement au hasard qui sont un cas particulier des mécanismes de création de données manquant au hasard.

Dans les conditions de régression, où on souhaite faire des inférences au sujet des paramètres qui régissent la loi de Y sachant les covariables fixes et connues X, (1) devient

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X, I) = \frac{\int\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi)\, p(Y_{\text{obs}} | X, \theta)\, p(\theta, \phi)\, d\theta\, d\phi}{\int\int\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi)\, p(Y_{\text{obs}} | X, \theta)\, p(\theta, \phi)\, d\theta\, d\phi\, dY_{\text{nob}}}$$

qui se réduit à

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X) = \frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi)\, p(Y_{\text{obs}} | X, \theta)\, p(\theta, \phi)\, d\theta}{\int\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi)\, p(Y_{\text{obs}} | X, \theta)\, p(\theta, \phi)\, d\theta\, dY_{\text{nob}}}$$

si, et uniquement si, I dépend seulement de (Y_{obs}, X), dont la dépendance à l'égard de X uniquement est un cas particulier. Donc, si l'on souhaite faire une inférence au sujet d'un paramètre de régression $Q(Y, X)$, alors un plan d'échantillonnage non informatif ou, plus généralement, ignorable peut permettre que les probabilités d'inclusion soient une fonction des covariables fixes.

2.1 Adaptation à des probabilités d'inclusion inégales

Maintenir l'hypothèse d'ignorabilité du mécanisme d'échantillonnage oblige souvent à tenir compte du plan d'échantillonnage dans la structure du modèle de vraisemblance et du modèle a priori. Dans le cas des plans d'échantillonnage avec probabilités d'inclusion inégales,

cela peut s'accomplir en créant un indice $h = 1, ..., H$ de la probabilité d'inclusion (Little 1983, 1991); il pourrait s'agir d'une application bijective des statistiques d'ordre des poids des cas sur leurs classements, ou d'un « regroupement » préliminaire des poids des cas en utilisant, par exemple, les centiles $100/H$ des poids des cas. Les données sont alors modélisées par

$$y_{hi} \mid \theta_h \sim f(y_{hi}; \theta_h),\ i = 1, ..., N_h$$

pour tous les éléments figurant dans la h^{e} strate d'inclusion, où θ_h tient compte d'une interaction entre le ou les paramètres du modèle θ et la strate d'inclusion h. Appliquer une loi a priori non informative à θ_h reproduit alors une analyse entièrement pondérée en ce qui concerne l'espérance de la loi prédictive a posteriori de $Q(Y)$.

Pour concrétiser ce qui précède, supposons que nous cherchions à estimer une moyenne de population $Q(Y) = \bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ d'après un échantillon à probabilités d'inclusion inégales avec un échantillonnage aléatoire simple dans les strates d'inclusion. En réécrivant l'équation sous la forme $Q(Y) = \sum_h P_h \bar{Y}_h$ où $\bar{Y}_h = N_h^{-1} y_{hi}$ est la moyenne de population de la strate d'inclusion et $P_h = N_h/N$, nous avons

$$E(\bar{Y} \mid Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h \mid Y_{\text{obs}}) = N^{-1}\sum_h \{n_h \bar{y}_{h,\text{obs}} + (N_h - n_h) E(\bar{Y}_{h,\text{nob}} \mid Y_{\text{obs}})\}$$

où $\bar{Y}_h$ est décomposé en la moyenne de strate d'inclusion observée $\bar{y}_{h,\text{obs}} = n_h^{-1}\sum_{i=1}^{N_h} I_{hi} y_{hi}$ et la moyenne de strate d'inclusion non observée $\bar{Y}_{h,\text{nob}} = (N_h - n_h)^{-1}\sum_{i=1}^{N_h} (1 - I_{hi}) y_{hi}$. Si nous supposons que

$$y_{hi} \mid \mu_h, \sigma_h^2 \overset{\text{ind}}{\sim} N(\mu_h, \sigma_h^2)$$

$$p(\mu_h, \sigma_h^2) \propto 1$$

alors

$$E(\bar{Y}_{h,\text{nob}} \mid Y_{\text{obs}}) = E(E(\bar{Y}_{h,\text{nob}} \mid Y_{\text{obs}}) \mid Y_{\text{obs}}, \mu_h, \sigma_h^2) = E(\mu_h \mid Y_{\text{obs}}) = \bar{y}_{h,\text{obs}}.$$

et la moyenne prédictive a posteriori de la moyenne de population est donnée par la moyenne d'échantillon pondérée :

$$E(\bar{Y} \mid Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h \mid Y_{\text{obs}}) = N^{-1}\sum_h N_h \bar{y}_{h,\text{obs}} = N^{-1}\sum_h \sum_{i=1}^{N_h} I_{hi} w_h y_{hi}$$

où $w_{hi} \equiv w_h = N_n/n_h$ pour tous les éléments observés dans la strate d'inclusion h. En outre, la moyenne pondérée sera l'espérance prédictive a posteriori de la moyenne de population pour toute loi hypothétique de Y à condition que $E(y_{hi} \mid \mu_h) = \mu_h$. Par contre, un simple modèle d'échangeabilité pour les données

$$y_i \mid \mu, \sigma^2 \overset{\text{ind}}{\sim} N(\mu, \sigma^2)$$

$$p(\mu, \sigma^2) \propto 1$$

donne $E(\bar{Y} \mid Y_{\text{obs}}) = n^{-1}\sum_{i=1}^{N} I_i y_i$, l'estimateur non pondéré de la moyenne, qui peut être gravement biaisé si l'échangeabilité n'est pas vérifiée, comme cela serait le cas s'il existait une association entre la probabilité d'inclusion et Y.

2.2 Modèles de lissage des poids

Sous sa forme générale, la « méthode de lissage des poids » que nous proposons consiste à stratifier les données en fonction de la probabilité d'inclusion, puis d'utiliser un modèle hiérarchique pour effectuer la réduction par la voie d'un rétrécissement (shrinkage). Une description d'un tel modèle est donnée par

$$y_{hi} \mid \theta_h \sim f(y_{hi}; \theta_h) \tag{3}$$

$$\theta_h \mid M_h, \mu, R \sim N(\hat{y}_h, R),\ \hat{y}_h = g(M_h, \mu)$$

$$\mu, R \mid M_h \sim \Pi.$$

où $h = 1, ..., H$ indice les probabilités d'inclusion de la plus élevée à la plus faible, $g(M_h, \mu)$ est une fonction liant l'information M_h provenant de la strate de probabilités d'inclusion et un paramètre de lissage μ au paramètre de distribution des données θ_h indicé par la strate d'inclusion, et Π est une loi uniforme ou faiblement informative d'un hyperparamètre (Little 2004).

Les détails de spécification de la fonction de la vraisemblance et de la distribution a priori dépendent du paramètre de population étudié, du plan d'échantillonnage, des hypothèses concernant la loi de y, et des compromis entre l'efficacité et la robustesse. Postuler un modèle d'échangeabilité sur les moyennes de strate d'inclusion provenant de l'exemple précédent donne (Lazzeroni et Little 1998, Elliott et Little 2000)

$$y_{hi} \mid \theta_h \overset{\text{ind}}{\sim} N(\theta_h, \sigma^2)$$

$$\theta_h \overset{\text{ind}}{\sim} N(\mu, \tau^2).$$

Si nous supposons pour le moment que σ^2 et τ^2 sont connus, nous avons

$$E(\bar{Y} \mid Y_{\text{obs}}) = N^{-1}\sum_h \{n_h \bar{y}_{h,\text{obs}} + (N_h - n_h) E(\mu_h \mid Y_{\text{obs}})\}$$

où $E(\mu_h \mid Y_{\text{obs}}) = w_h \bar{y}_h + (1 - w_h)\tilde{y}$ pour $w_h = \tau^2 n_h/(\tau^2 n_h + \sigma^2)$ et $\tilde{y} = (\sum_h n_h/(n_h\tau^2 + \sigma^2))^{-1}\sum_h n_h/(n_h\tau^2 + \sigma^2)\bar{y}_h$. À mesure que $\tau^2 \to \infty$, $w_h \to 1$ de sorte que $E(\bar{Y} \mid Y_{\text{obs}}) = \sum_h P_h \bar{y}_h$; donc, une loi a priori uniforme

rétablit l'estimateur entièrement pondéré, comme nous l'avons montré plus haut. Par ailleurs, à mesure que $\tau^2 \to 0$, $w_h \to 0$ de sorte que $E(\mu_h \mid Y_{\text{obs}}) \to \tilde{y}\mid_{\tau^2=0} = \bar{y}$, la moyenne non pondérée; donc, les unités exclues de l'échantillon sont estimées à la moyenne regroupée, puisque le modèle suppose que tous les y_{hi} sont tirés à partir d'une moyenne commune. Par conséquent, ce modèle de lissage des poids permet un compromis entre l'estimateur convergeant sous le plan qui pourrait être très inefficace et l'estimateur non pondéré qui est entièrement efficace sous l'hypothèse forte que les probabilités d'inclusion et la moyenne de Y sont indépendantes. En supposant que τ^2 suit une hyperloi a priori faible, le degré de compromis entre les moyennes pondérée et non pondéré sera « dicté par les données », quoique sous les hypothèses de modélisation.

2.3 Lissage des poids pour les modèles de régression linéaire et de régression linéaire généralisée

Les modèles de régression linéaire généralisée (McCullagh et Nelder 1989) postulent pour y_i une vraisemblance de la forme

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right] \qquad (4)$$

où $a_i(\phi)$ fait intervenir une constante connue et un paramètre d'échelle (de perturbation) ϕ, et la moyenne de y_i est reliée à une combinaison linéaire de covariables fixes $\boldsymbol{x}_i$ par une fonction lien $g(\cdot)$: $E(y_i \mid \theta_i) = \mu_i$, où $g(\mu_i) = g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Nous avons aussi $\text{Var}(y_i \mid \theta_i) = a_i(\phi)V(\mu_i)$, où $V(\mu_i) = b''(\theta_i)$. Le lien est canonique si $\theta_i = \eta_i$, auquel cas $g'(\mu_i) = V^{-1}(\mu_i)$. Des exemples bien connus sont la loi normale, où $a_i(\phi) = \sigma^2$ et le lien canonique est $g(\mu_i) = \mu_i$; la loi binomiale, où $a_i(\phi) = n_i^{-1}$ et le lien canonique est $g(\mu_i) = \log(\mu_i/(1-\mu_i))$; et la loi de Poisson, où $a_i(\phi) = 1$ et le lien canonique est $g(\mu_i) = \log(\mu_i)$.

En indiçant la strate d'inclusion par h, nous avons $g(E[y_{hi} \mid \boldsymbol{\beta}_h]) = \boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h$. Soit un modèle hiérarchique hypothétique de la forme

$$(\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_H^T)^T \mid \boldsymbol{\beta}^*, G \sim N_{Hp}(\boldsymbol{\beta}^*, G). \qquad (5)$$

où $\boldsymbol{\beta}^*$ est un vecteur inconnu des valeurs moyennes des coefficients de régression et G est une matrice de covariance inconnue.

Nous considérons que la quantité de population cible d'intérêt $\boldsymbol{B} = (B_1, ..., B_p)^T$ est la pente qui résout l'équation de score de population $U_N(\boldsymbol{B}) = 0$, où

$$U_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial}{\partial\boldsymbol{\beta}} \log f(y_i;\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{(y_{hi} - g^{-1}(\mu_i(\boldsymbol{\beta})))x_{hi}}{V(\mu_{hi}(\boldsymbol{\beta}))g'(\mu_{hi}(\boldsymbol{\beta}))}. \qquad (6)$$

Notons que la quantité $\boldsymbol{B}$ telle que $U(\boldsymbol{B}) = 0$ est toujours une quantité de population d'intérêt significative, même si le modèle est spécifié incorrectement (c'est-à-dire que η_i n'est pas exactement linéaire en ce qui concerne les covariables), puisqu'il s'agit de l'approximation linéaire de $\boldsymbol{x}_i$ à $\eta_i = g(\mu_i)$. Sous le modèle donné par (4) et (5), une approximation d'ordre un (en supposant que la fraction d'échantillonnage est négligeable) à $E(\boldsymbol{B} \mid y, X)$ est donnée par $\hat{B}$ où

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \frac{(\hat{y}_{hi} - g^{-1}(\mu_i(\hat{\boldsymbol{B}})))\boldsymbol{x}_{hi}}{V(\mu_{hi}(\hat{\boldsymbol{B}}))g'(\mu_{hi}(\hat{\boldsymbol{B}}))} = 0 \qquad (7)$$

où $W_h = N_h/n_h$, $\hat{y}_{hi} = g^{-1}(\boldsymbol{x}_{hi}^T\hat{\boldsymbol{\beta}}_h)$ et $\hat{\boldsymbol{\beta}}_h = E(\boldsymbol{\beta}_h \mid y, X)$, tel qu'il est déterminé par la forme de (5). (Si N_h est inconnu, il peut être remplacé par $\hat{N}_h = \sum_{i\in h} w_{hi}$, et les $\hat{N}_1, ..., \hat{N}_H$ peuvent être traités comme une loi multinomiale de taille N paramétrisée par les probabilités inconnues de la strate d'inclusion $q_1, ..., q_H$ avec, par exemple, une loi a priori de Dirichlet.) Donc, dans l'exemple d'une régression linéaire, où $V(\mu_i) = \sigma^2$ et $g'(\mu_i) = 1$, (7) se résout en

$$\hat{\boldsymbol{B}} = E(\boldsymbol{B} \mid y, X) = \left[\sum_h W_h \sum_{i=1}^{n_h} \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right]^{-1}\left[\sum_h W_h\left(\sum_{i=1}^{n_h}\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right)\hat{\boldsymbol{\beta}}_h\right]. \qquad (8)$$

Dans l'exemple de la régression logistique, où $V(\mu_i) = \mu_i(1-\mu_i)$ et $g'(\mu_i) = \mu_i^{-1}(1-\mu_i)^{-1}$, $E(\boldsymbol{B} \mid y, X)$ s'obtient en résolvant, pour trouver les paramètres de régression de population B_j, $j = 1, ..., p$,

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij}\frac{\exp(x_{hij}B_j)}{1+\exp(x_{hij}B_j)} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij}\frac{\exp(x_{hij}\hat{\beta}_{hj})}{1+\exp(x_{hij}\hat{\beta}_{hj})}. \qquad (9)$$

On peut, pour cela, appliquer de simples méthodes numériques de calcul de racine comme la méthode de Newton.

Dans le présent article, nous considérons quatre formes de $\boldsymbol{\beta}^*$ et G dans (5) :

1. Pente aléatoire échangeable (PAE) :

$$\boldsymbol{\beta}_h^* = (\beta_0^*, ..., \beta_p^*) \text{ pour tous } h, G = I_H \otimes \sum. \qquad (10)$$

2. Pente aléatoire autorégressive (PAA) :

$$\boldsymbol{\beta}_h^* = (\beta_0^*, ..., \beta_p^*) \text{ pour tous } h,$$
$$G = A \otimes \sum, A_{jk} = \rho^{|j-k|}, j, k = 1, ..., H.$$

3. Pente aléatoire linéaire (PAL) :

$$\boldsymbol{\beta}_h^* = (\beta_{00}^* + \beta_{01}^* h, ..., \beta_{p0}^* + \beta_{p1}^* h),$$
$$G = I_H \otimes \Lambda.$$

4. Pente aléatoire non paramétrique (PANP) :

$$\boldsymbol{\beta}_h^* = (f_0(h), ..., f_p(h)), G = 0.$$

$$\left\{ \begin{array}{l} f_j : f_j^v \text{ absolument continue, } v = 0,1, \\ \int (f_j^{(2)}(u))^2 du < \infty, \\ \min\limits_{f_j} \sum\limits_h (\beta_{hj}^* - f_j(h))^2 + \lambda_j \int (f_j^{(2)}(u))^2 du \end{array} \right\}$$

où h indice de nouveau la probabilité d'inclusion, I_H est une matrice d'identité $H \times H$, ρ est un paramètre d'autocorrélation qui contrôle le degré de rétrécissement sur l'ensemble de strates de poids, Σ est une matrice de covariance $p \times p$ non contrainte, Λ est une matrice diagonale $p \times p$ et $f_j(h)$ est une fonction lisse doublement dérivable de h qui minimise la somme des carrés des résidus plus une pénalité d'irrégularité paramétrisée par λ_j (Wahba 1978, Hastie et Tibshirani 1990). En reformulant le modèle PANP comme dans Wang (1998) nous obtenons

$$y_{hi} \mid \boldsymbol{\beta}_h \overset{\text{ind}}{\sim} N(\boldsymbol{x}_{hi}^T \boldsymbol{\beta}_h, \sigma^2)$$

$$\beta_{hj} = \beta_{j0}^* + \beta_{j1}^* h + \boldsymbol{\omega}_h \boldsymbol{u}_j$$

$$\boldsymbol{u}_j \overset{\text{ind}}{\sim} N_{H-1}(0, I\tau_j^2), \tau_j^2 = \sigma^2/(H\lambda_j)\ j = 0, ..., p$$

où $\boldsymbol{\omega}_h$ est la i^{e} ligne de la décomposition de Choleski de la matrice de base des splines cubiques Ω, où $\Omega_{hk} = \int_0^1 ((h-1)/(H-1) - t)_+ ((k-1)/(H-1) - t)_+ dt$, $(x)_+ = x$ si $x \geq 0$ et $(x)_+ = 0$ si $x < 0$, $h, k = 1, ..., H$. Le modèle PANP peut être étendu à la forme du modèle linéaire généralisé comme dans Lin et Zhang (1999), où l'hypothèse de normalité au premier degré est remplacée par une fonction lien qui est linéaire en les covariables : $g(E(y_{hi} \mid \boldsymbol{\beta}_h)) = \boldsymbol{x}_{hi}^T \boldsymbol{\beta}_h$, pour $g(\cdot)$ comme dans (4).

En supposant pour le moment que les paramètres de deuxième degré sont connus, nous voyons que, dans le cas du modèle PAE avec des données normales, à mesure que $|G| \to \infty$, le partage d'information entre les strates d'inclussion cesse et $\hat{\boldsymbol{\beta}}_h \approx (\boldsymbol{x}_h^T \boldsymbol{x}_h)^{-1} \boldsymbol{x}_h^T \boldsymbol{y}_h$, l'estimateur par la régression dans la strate d'inclusion. En introduisant cette expression par substitution dans (8) nous obtenons $\hat{\boldsymbol{B}} \approx \hat{\boldsymbol{B}}^w$, l'estimateur entièrement pondéré de la pente de population. De même, à mesure que $|G| \to 0$, les pentes dans les strates d'inclusion $\hat{\boldsymbol{\beta}}_h \approx \boldsymbol{\beta}^*$, qui est la pente a priori commune, ce qui donne $\hat{\boldsymbol{B}} \approx \boldsymbol{\beta}^*$ après introduction par substitution dans (8), ou $\hat{\boldsymbol{B}}^u$ si une hyperloi a priori non informative est appliquée à $\boldsymbol{\beta}^*$ et que sa moyenne a posteriori est obtenue par $(\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y}$. Les méthodes empiriques ou entièrement bayésienne qui permettent que les données estiment les paramètres de deuxième degré admettent donc le « lissage des poids » dicté par les données, qui établit un compromis entre les estimateurs non pondéré et entièrement pondéré.

En pratique, naturellement, la moyenne et les composantes de la variance de deuxième degré sont en général inconnues; donc, nous complétons les spécifications du modèle en postulant une hyperloi a priori pour les paramètres de deuxième degré :

$$p(\phi, \boldsymbol{\beta}^*, G) \propto p(\boldsymbol{\zeta}).$$

Habituellement, l'hyperloi a priori $p(\boldsymbol{\zeta})$ est faiblement informative ou non informative. Nous pouvons alors utiliser l'échantillonnage de Gibbs (Gelfand et Smith 1990; Gelman et Rubin 1992) pour obtenir des tirages à partir de la loi a posteriori conjointe complète de $(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \phi, G)^T \mid \boldsymbol{y}, X$. Dans le modèle PAE, nous considérons $p(\sigma, \boldsymbol{\beta}^*, \Sigma) \propto \sigma^{-2} \mid \Sigma \mid^{-(p+1/2)} \exp(-1/2 \text{ tr } \{r\Sigma^{-1}\})$, c'est-à-dire des lois a priori non informatives pour les paramètres d'échelle et de moyenne a priori, ainsi qu'une hyperloi a priori Wishart-inverse de la variance a priori G centrée à la matrice d'identité à facteur d'échelle r avec p degrés de liberté. La même loi a priori est utilisée pour le modèle PAA, avec l'hypothèse supplémentaire que $\rho \overset{\text{ind}}{\sim} U(0,1)$ (autocorrélation non négative entre les strates d'inclusion). Dans les modèles PAL et PANP, $p(\sigma, \boldsymbol{\beta}^*, \Lambda) \propto \sigma^{-2}$ et $p(\sigma, \boldsymbol{\beta}^*, \boldsymbol{\tau}) \propto \sigma^{-2}$ (loi a priori non informative du paramètre d'échelle et hyperloi a priori standard). La description des tirages conditionnels de l'échantillonneur de Gibbs peut être consultée à http://www.sph.umich.edu/mrelliot/trim/meth2.pdf.

Le degré de compromis est une fonction de la structure de la moyenne et de la variance du modèle choisi. Les modèles PAE et PAA reposent sur l'hypothèse que les moyennes de pentes sont échangeables; le modèle PAA est plus souple en ce sens que la structure de sa variance permet que les unités ayant des probabilités d'inclusion presque égales soient lissées plus fortement que celles dont les probabilités d'inclusion sont très inégales. Le modèle PAL suppose une tendance linéaire sous-jacente dans les pentes, tandis que le modèle PANP suppose uniquement une tendance sous-jacente lisse jusqu'à la deuxième dérivée. Notons que, dans les modèles PAL et PANP, nous supposons l'indépendance a priori des paramètres de régression associés à une covariable donnée, c'est-à-dire $(\beta_{1j}, ... \beta_{Hj}) \perp (\beta_{1j'}, ... \beta_{Hj'})$, $j \neq j'$. Il en est ainsi parce que nous modélisons les tendances dans ces paramètres sur l'ensemble de la strate d'inclusion et que nous ne souhaitons pas « lier » ces tendances sur l'ensemble des covariables.

Le rétrécissement le plus important, correspondant à la réduction le plus rigoureux des poids, s'obtient quand la variabilité des pentes des strates de poids est faible, ou quand les strates des probabilités d'inclusion les plus faibles sont mal estimées. Le rétrécissement devrait être faible si les pentes de strate de poids sont estimées avec précision et qu'elles sont systématiquement associées à leurs probabilités d'inclusion. En nous basant sur Elliott et Little (2000), nous nous attendrions à ce que le modèle PAE soit le plus efficace lorsqu'une réduction importante des poids est nécessaire pour minimiser l'EQM, mais soit le plus vulnérable au « sur-rétrécissement » quand la correction du biais est l'aspect le plus important. Accroître la structure, particulièrement dans la partie moyenne du modèle, comme dans les modèles PAL et PANP, donnera une estimation plus robuste en ce sens que le sur-rétrécissement aura lieu uniquement dans les situations quasi pathologiques (par exemple, lorsque les tendances des moyennes sont non monotones et très discontinues) et, même dans de telles situations, pourrait donner lieu à une correction du biais un peu plus faible que ne le justifient les données. Le prix à payer pour cette robustesse sera toutefois une réduction de l'efficacité relative des modèles d'échangeabilité.

3. Résultats des simulations

Parce que nous souhaitons des modèles simultanément plus efficaces que les estimateurs fondés sur le plan de sondage, mais néanmoins raisonnablement robustes à l'erreur de spécification du modèle - et en général nous estimons que même les modèles bayésiens devraient avoir de bonnes propriétés fréquentistes - nous évaluons nos modèles proposés dans un contexte de rééchantillonnage. Nous considérons les régressions linéaire et logistique, sous un modèle spécifié incorrectement avec un plan d'échantillonnage non informatif.

3.1 Régression linéaire

Dans le cas du modèle de régression linéaire en présence d'erreur de spécification, nous avons généré des données de population de la façon suivante :

$$Y_i \mid X_i, \sigma^2 \sim N(\alpha X_i + \beta X_i^2, \sigma^2), \qquad (11)$$

$$X_i \sim U(0, 10), i = 1, ..., N = 20\,000.$$

Nous avons utilisé un schéma d'échantillonnage disproportionnellement spécifié, non informatif, pour échantillonner les éléments en fonction de X_i (I_i égale 1 si l'élément est échantillonné et 0 autrement) :

$$h_i = \lceil X_i \rceil$$

$$P(I_i = 1 \mid h_i) = \pi_i \propto (1 + h_i/2{,}5)\, h_i$$

Nous avons créé ainsi 10 strates, définies par les parties entières des valeurs de X_i. Les éléments (Y_i, X_i) avaient $\approx 1/36^{e}$ de la probabilité de sélection quand $0 < X_i \leq 1$ que lorsque $9 < X_i < 10$. Nous avons échantillonné $n = 500$ éléments sans remise pour chaque simulation. L'objet de l'analyse était d'obtenir la pente de population $B_1 = \sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X}) / \sum_{i=1}^{N}(X_i - \bar{X})^2$. Nous avons fixé $\alpha = \beta = 1$, ce qui a introduit un biais positif dans l'estimation de B_1, et avons fait varier σ^2. L'effet de l'erreur de spécification du modèle augmente à mesure que $\sigma^2 \to 0$, car le biais des estimateurs devient grand comparativement à la variance, et inversement diminue quand $\sigma^2 \to \infty$. Nous avons considéré les valeurs de $\sigma^2 = 10^l$, $l = 1, ..., 5$; 200 simulations ont été générées pour chaque valeur de σ^2.

Ici et plus loin, nous avons utilisé une hyperloi a priori Wishart inverse sur la variance a priori G, centrée à la matrice d'identité avec deux degrés de liberté.

En plus des modèles à pente aléatoire échangeable (PAE), à pente aléatoire autorégressive (PAA), à pente aléatoire linéaire (PAL) et à pente aléatoire non paramétrique (PANP) discutés à la section 2.3, nous avons considéré l'estimateur fondé sur le plan (entièrement pondéré) standard, ainsi que les estimateurs à poids réduits et non pondéré. Pour l'estimateur entièrement pondéré (EPD), nous avons utilisé l'EPMV $\boldsymbol{B}_w = (X'WX)^{-1}X'W\boldsymbol{y}$ où, en dénotant par une minuscule les éléments échantillonnés $(I_i = 1)$, $w_{hi} \equiv w_h$ pour $h = 1, ..., H$, $i = 1, ..., n_h$, $W = \text{diag}(w_{hi})$, $\boldsymbol{x}_{hi} = (1\, x_{hi})'$, X_h contient les lignes empilées de $\boldsymbol{x}'_{hi}$ et X contient les matrices empilées X_h. Nous avons obtenu l'inférence au sujet de $\hat{\boldsymbol{B}}_w$ par l'approximation standard par développement en série de Taylor (Binder 1983) :

$$\text{Var}(\hat{\boldsymbol{B}}_w) = \hat{S}_{XX}^{-1} \hat{\sum}(\hat{\boldsymbol{B}}_w) \hat{S}_{XX}^{-1}$$

où $\hat{S}$ est un estimateur convergeant d'un total de population $\sum_{i=1}^{N} \boldsymbol{x}'_i \boldsymbol{x}_i$ donné par $X'WX$ et $\hat{\sum}(\hat{\boldsymbol{B}}_w)$ est une estimation convergente sous le plan de la variance du total $\sum_{i=1}^{N} \epsilon_i \boldsymbol{x}_i$ où $\epsilon_i = y_i - \boldsymbol{x}_i \boldsymbol{B}$ est la différence entre la valeur de y_i et sa valeur estimée sous la pente de population réelle $\boldsymbol{B}$: $\hat{\sum}(\hat{\boldsymbol{B}}_w) = \sum_h n_h/(n_h - 1) \sum_{i=1}^{n_h} (\tilde{\boldsymbol{x}}_{hi} - \bar{\tilde{\boldsymbol{x}}}_h)'(\tilde{\boldsymbol{x}}_{hi} - \bar{\tilde{\boldsymbol{x}}}_h)$, où $\tilde{\boldsymbol{x}}_{hi} = w_{hi} e_{hi} \boldsymbol{x}_{hi}$ pour $e_{hi} = y_{hi} - \boldsymbol{x}_{hi} \hat{\boldsymbol{B}}_w$. Nous considérons aussi l'estimateur à poids réduits (PEL) obtenu en remplaçant les poids w_{hi} par les valeurs réduites w_{hi}^t qui fixent la valeur maximale normalisée à 3 : $w_{hi}^t = N \tilde{w}_{hi}^t / \sum_{h=1}^{H} n_h \tilde{w}_h^t$, où $\tilde{w}_{hi}^t = \min(w_{hi}, 3N/n)$, et l'estimateur non pondéré (NPD) obtenu en fixant $w_{hi} = N/n$ pour tous h, i.

Le tableau 1 donne le biais relatif, la racine de l'erreur quadratique moyenne (REQM) et la couverture réelle de l'intervalle de confiance à taux nominal de 95 % pour les

trois estimateurs fondés sur le plan de sondage et les quatre estimateurs fondés sur un modèle de la pente de population (deuxième composante de $\hat{\boldsymbol{B}}$) étudiés, en fonction de la variance σ^2.

L'estimateur entièrement pondéré de la pente de population est essentiellement sans biais par rapport au plan sous erreur de spécification du modèle; les estimateurs non pondérés et à poids réduits sont biaisés. Les biais des modèles d'échangeabilité et autorégressif augmentent à mesure que la variance augmente, car ces modèles échangent l'absence de biais de l'estimateur entièrement pondéré pour la variance réduite de l'estimateur non pondéré. Les modèles linéaire et non paramétrique sont approximativement sans biais.

Les estimateurs non pondéré et à poids réduits donnent de mauvais résultats en ce qui concerne l'EQM pour les petites valeurs de σ^2, où le biais dû à l'erreur de spécification du modèle est critique, et de bons résultats pour les valeurs plus grandes de σ^2, où l'instabilité de l'estimateur entièrement pondéré est plus importante que la réduction du biais. L'estimateur basé sur un modèle d'échangeabilité a de bonnes propriétés en ce qui concerne la REQM pour les petites et les grandes valeurs de σ^2, les réductions de l'EQM étant de plus de 30 %, mais produit un lissage excessif pour les degrés intermédiaires de spécification du modèle. Les propriétés du modèle autorégressif sont égales à celles du modèle d'échangeabilité pour les petites et les grandes valeurs de σ^2, mais il est protégé en grande partie contre le surlissage observé pour le modèle d'échangeabilité aux niveaux intermédiaires. Les modèles linéaire et non paramétrique dominent essentiellement les estimateurs entièrement pondérés en ce qui concerne l'EQM sous toutes les simulations envisagées, quoique les réductions de l'EQM soient seulement de l'ordre de 10 %.

Les estimateurs non pondéré et à poids réduits ont une mauvaise couverture de l'intervalle de confiance, sauf quand l'erreur de spécification du modèle est quasi absente. L'échec du compromis entre le biais et la variance dans le cas de l'estimateur d'échangeabilité en présence d'erreur de spécification du modèle est mis en évidence par la mauvaise couverture de l'estimateur pour les valeurs intermédiaires de σ^2; cette situation est améliorée, mais n'est pas entièrement éliminée, dans le cas de l'estimateur autorégressif. Les estimateurs linéaire et non paramétrique ont une bonne couverture lorsque l'erreur de spécification du modèle est moins importante, mais produisent une certaine sous-couverture lorsque l'erreur de spécification du modèle est plus importante.

3.2 Régression logistique

Dans le modèle de régression logistique, nous avons généré des données de population de la façon suivante :

$$P(Y_i = 1 \mid X_i) \sim B(\text{expit}(3,25 - 0,75X_i + \gamma X_i^2)), \quad (12)$$

$$X_i \sim U(0,10), i = 1, ..., N = 20\,000$$

où $B(p)$ est une loi de Bernoulli avec probabilité de « succès » p, $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. L'objet de l'analyse est d'obtenir la pente de population de la régression logistique définie comme étant la valeur B_1 dans l'équation $\sum_i^N (y_i - \text{expit}(B_0 + B_1 x_i))\begin{pmatrix}1\\ x_i\end{pmatrix} = 0$. Un plan d'échantillonnage avec probabilités de sélection inégales a été appliqué comme il est décrit dans les simulations par la régression linéaire. Nous considérons les valeurs de $\gamma = 0, 0,0158, 0,0273, 0,0368, 0,0454,$ qui correspondent à des mesures de courbures de $K = 0, 0,02, 0,04, 0,06, 0,08$ au point médian 5 du support de X, où $K(X;\gamma) = |2\gamma/[1 + (2\gamma X - 0,75)^2]^{3/2}|$; 200 simulations ont été exécutées pour chaque valeur de γ. Comme dans les simulations par régression linéaire, les éléments ont été échantillonnés sans remise avec probabilité proportionnelle à $(1 + h_i/2,5)h_i$; en tout, 1 000 éléments ont été échantillonnés pour chaque simulation. Nous avons de nouveau considéré les estimateurs basés sur l'EPMV entièrement pondéré (EPD), non pondéré (NPD) et à poids réduits (PEL), ainsi que les estimateurs à pente aléatoire échangeable (PAE), à pente aléatoire autorégressive (PAA), à pente aléatoire linéaire (PAL) et à pente aléatoire non paramétrique (PANP). L'inférence au sujet des estimateurs EPMV est obtenue au moyen d'approximations par développement en série de Taylor (Binder 1983), comme il est discuté à la section précédente.

Le tableau 2 donne le biais relatif, la REQM relative à la REQM de l'estimateur entièrement pondéré et la couverture réelle des IC ou des intervalles prédictifs a posteriori à taux nominal de 95 % associés à chacun des sept estimateurs de la pente de population (B) pour diverses valeurs de la courbure K, correspondant à des degrés croissants d'erreur de spécification.

Le sous-échantillonnage des petites valeurs de X signifie que l'estimateur du maximum de vraisemblance de B dans les conditions d'erreurs de spécification du modèle est sans biais pour $K = 0$ et présente un biais par défaut pour $K = 0,02, 0,04, 0,06,$ et 0,08 à moins qu'il soit tenu compte du plan d'échantillonnage. Le biais de l'estimateur à poids réduits est compris entre ceux de l'estimateur non pondéré et de l'estimateur entièrement pondéré. Le biais de l'estimateur à échangeabilité est compris entre ceux de l'estimateur à poids réduits et de l'estimateur entièrement pondéré; le biais de l'estimateur autorégressif est compris entre ceux de l'estimateur à échangeabilité et de l'estimateur entièrement pondéré, tandis que les estimateurs linéaire et non paramétrique sont essentiellement sans biais.

L'estimateur non pondéré possède une EQM sensiblement améliorée (40 % de réduction) lorsque le modèle à

pente linéaire est approximativement correctement spécifié, mais donne de mauvais résultats pour un degré modéré à important d'erreur de spécification. Les estimateurs à poids réduits, autorégressif et non paramétrique sont tous supérieurs à l'estimateur standard entièrement pondéré, et l'estimateur à échangeabilité et l'estimateur linéaire le sont presque, sur l'étendue des simulations considérées. L'estimateur à poids réduits brut a produit une réduction de 30 % de l'EQM, les estimateurs non paramétrique, à échangeabilité et autorégressif, des réductions allant jusqu'à 20 à 25 %, et l'estimateur linéaire, des réductions de seulement 10 % ou moins.

L'estimateur non pondéré donne une mauvaise couverture de l'IC nominal, sauf quand le modèle à pente linéaire est spécifié correctement ou qu'il l'est presque. Les estimateurs fondés sur un modèle ont généralement de bonnes propriétés de couverture lorsque le modèle linéaire a été spécifié correctement, une légère réduction de la couverture étant observée lorsque la courbure est importante.

Tableau 1
Biais relatif (%), racine carrée de l'erreur quadratique moyenne (REQM) relativement à la REQM de l'estimateur entièrement pondéré, et couverture réelle de l'intervalle de confiance ou de l'intervalle prédictif a posteriori à 95 % de l'estimateur de la pente de la régression linéaire de population sous erreur de spécification du modèle. La pente et l'ordonnée à l'origine de population sont estimées au moyen des estimateurs fondés sur le plan de sondage non pondéré (NPD), entièrement pondéré (EPD) et à poids réduits (PEL), et en tant que moyenne a posteriori dans (8) sous une loi a priori d'échangeabilité (PAE), autorégressif (PAA), linéaire (PAL) et non paramétrique (PANP) pour les paramètres de régression. Les valeurs de l'EQM relatives à l'estimateur entièrement pondéré inférieures à 1 sont en caractères gras

	Biais relatif (%) $\log_{10}$ de la variance					REQM relative à EPD $\log_{10}$ de la variance					Couverture réelle $\log_{10}$ de la variance				
Estimateur	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
NPD	21,5	21,8	22,2	20,8	22,3	12,1	4,57	1,76	**0,75**	**0,67**	0	0	6	78	92
EPD	0,0	0,1	1,4	1,6	-0,2	1	1	1	1	1	94	95	96	94	96
PEL	8,3	8,4	9,6	8,8	7,8	4,74	1,88	1,02	**0,71**	**0,75**	0	13	78	95	96
PAE	0,2	2,2	11,4	15,1	18,3	1,00	1,17	1,18	**0,73**	**0,68**	87	86	64	91	96
PAA	0,1	1,4	9,6	14,5	17,4	1,00	1,03	1,11	**0,74**	**0,69**	87	89	78	90	96
PAL	-0,2	-0,4	1,1	1,6	-0,3	**0,99**	**0,91**	**0,91**	**0,91**	**0,93**	85	91	96	95	94
PANP	-0,1	-0,3	0,9	1,5	-0,4	**0,89**	**0,90**	**0,95**	**0,90**	**0,95**	86	92	96	94	94

Tableau 2
Biais relatif (%), racine carrée de l'erreur quadratique moyenne (REQM) relativement à la REQM de l'estimateur entièrement pondéré, et couverture réelle de l'intervalle de confiance ou de l'intervalle prédictif a posteriori à 95 % de l'estimateur de la pente de la régression logistique de population sous erreur de spécification du modèle. La pente et l'ordonnée à l'origine de population sont estimées au moyen des estimateurs fondés sur le plan de sondage non pondéré (NPD), entièrement pondéré (EPD) et à poids réduits (PEL), et en tant que moyenne a posteriori dans (8) sous une loi a priori d'échangeabilité (PAE), autorégressif (PAA), linéaire (PAL) et non paramétrique (PANP) pour les paramètres de régression. Les valeurs de l'EQM relatives à l'estimateur entièrement pondéré inférieures à 1 sont en caractères gras

	Biais relatif (%) Courbure K					REQM relative à EPD Courbure K					Couverture réelle Courbure K				
Estimateur	0	0,02	0,04	0,06	0,08	0	0,02	0,04	0,06	0,08	0	0,02	0,04	0,06	0,08
NPD	1,0	-4,9	-11,9	-21,6	-34,6	**0,57**	**0,73**	**0,88**	1,19	1,61	96	89	66	32	17
EPD	1,1	2,2	1,3	-0,3	1,6	1	1	1	1	1	95	94	90	94	94
PEL	0,5	-1,0	-3,5	-7,2	-12,1	**0,70**	**0,77**	**0,77**	**0,78**	**0,95**	98	97	94	84	92
PAE	1,3	-0,8	-1,9	-5,6	-8,7	**0,75**	**0,82**	**0,85**	**0,88**	1,02	97	94	92	91	90
PAA	1,3	-0,5	-2,2	-4,8	-7,5	**0,78**	**0,85**	**0,84**	**0,84**	**0,95**	94	92	90	92	90
PAL	0,8	1,7	1,5	-0,4	1,1	**0,89**	**0,97**	**0,94**	**0,91**	1,02	95	91	88	92	89
PANP	0,3	1,5	1,1	0,9	0,5	**0,87**	**0,88**	**0,87**	**0,80**	**0,90**	95	92	88	94	96

4. Application : estimation de la prévalence des blessures chez les enfants passagers dans les camionnettes compactes à cabine allongée

L'ensemble de données de Partners for Child Passenger Safety correspond à l'échantillon disproportionné, à probabilités connues, provenant de toutes les demandes d'indemnisation reçues par State Farm depuis décembre 1998 comportant au moins un enfant de 15 ans ou moins passager dans un véhicule de modèle 1990 ou plus récent assuré par State Farm (Durbin, Bhatia, Holmes, Shaw, Werner, Sorenson et Winston 2001). Comme les blessures, et particulièrement les blessures « graves » définies comme étant des lacérations faciales ou d'autres blessures recevant une cote de 2 ou plus sur l'échelle AIS (Abbreviated Injury Scale) (Association for the Advancement of Automotive Medicine 1990), sont relativement rares, même chez les enfants dans la population de demandes de remboursement pour dommage à un véhicule accidenté, un échantillon en grappes disproportionnel stratifié est utilisé pour sélectionner les véhicules (l'unité d'échantillonnage) en vue de réaliser un sondage téléphonique auprès des conducteurs. Les véhicules contenant des enfants ayant reçu un traitement médical après l'accident ont été sur-échantillonnés afin que la majorité des enfants blessés soient sélectionnés, tout en veillant à ce que l'échantillon demeure représentatif de l'ensemble de la population. (Par traitement médical, on entend un traitement prodigué par des ambulanciers paramédiques, un traitement reçu au cabinet d'un médecin ou dans un service d'urgence, ou l'hospitalisation.) Pour tout véhicule échantillonné, tous les enfants passagers de ce véhicule ont été inclus dans l'enquête. Les conducteurs des véhicules échantillonnés ont été contactés par téléphone et, si un traitement médical avait été reçu par un passager, soumis à une présélection au moyen d'un questionnaire abrégé afin de confirmer la présence d'au moins un enfant passager ayant subi une blessure. Tous les véhicules contenant au moins un enfant pour lesquels la présélection était positive pour une blessure, ainsi qu'un échantillon aléatoire de 10 % des véhicules pour lesquels il avait été déclaré que les enfants passagers avaient reçu un traitement médical, mais avait donné un résultat de présélection négatif pour les blessures ont été sélectionnés pour un interview complète; un échantillon à 2 % (par après 2,5 %) d'accidents pour lesquels aucun traitement médical n'avait été reçu a également été sélectionné. Parce que la stratification du traitement est imparfaitement associée au risque de blessure (plus de 15 % de la population ayant subi des blessures graves se situe, selon les estimations, dans la catégorie des probabilités de sélection les plus faibles et près de 20 % des personnes n'ayant pas subi de blessure grave sont comprises dans la catégorie des probabilités de sélection les plus élevées), le plan d'échantillonnage est informatif, avec des rapports de cotes non pondérées biaisés vers zéro (Korn et Graubard 1995). En outre, les poids qui s'appliquent à cet ensemble de données sont plutôt variables : $1 \le w_i \le 50$, où 9 % des poids ont une valeur normalisée supérieure à 3.

Winston, Kallan, Elliott, Menon et Durbin (2002) ont déterminé que les enfants assis sur le siège arrière dans les camionnettes compactes à cabine allongée courent un plus grand risque de blessure grave que ceux assis sur le siège arrière d'autres véhicules. Cependant, la quantification du degré de risque excédentaire, donc de l'importance du problème de santé publique, posait des difficultés. Le rapport de cotes (RC) non pondérées exprimant le risque d'une blessure grave chez les enfants voyageant dans une camionnette compacte à cabine allongée comparativement à d'autres véhicules était de 3,54 (IC à 95 % : 2,01, 6,23), comparativement à 11,32 (IC à 95 % : 2,67, 48,03) pour l'estimateur entièrement pondéré. Puisque le risque de blessure ainsi que l'utilisation d'une camionnette compacte à cabine allongée étaient tous deux associés à l'âge de l'enfant, à la gravité de l'accident (intrusion et le fait que le véhicule soit utilisable ou non), à la direction de l'impact et au poids du véhicule, nous avons également considéré un modèle de régression logistique multivariée tenant compte de ces facteurs. Les rapports de cotes corrigées non pondérées et entièrement pondérées pour le risque de blessure chez les enfants assis sur le siège arrière d'une camionnette compacte à cabine allongée comparativement à d'autres véhicules sont de 3,50 (IC à 95 % : 1,88, 6,53) et de 14,56 (IC à 95 % : 3,45, 61,40) respectivement. L'utilisation de l'estimateur non pondéré posait des problèmes, à cause du biais vers zéro induit par le plan d'échantillonnage informatif; toutefois, l'estimateur entièrement pondéré semblait hautement instable, en partie à cause de la présence d'un enfant ayant subi une blessure grave dans une camionnette compacte à cabine allongée pour lequel la probabilité de sélection était très faible (0,025). Dans Winston et coll. (2002), cet enfant a été éliminé avant d'effectuer l'analyse.

Le tableau 3 donne les résultats pour les rapports des cotes non corrigées et corrigées du risque de blessure grave obtenues en utilisant les estimateurs fondés sur le plan de sondage non pondéré, entièrement pondéré et à poids réduits, ainsi que les estimateurs fondés sur les modèles de la pente de la régression à échangeabilité, autorégressif et linéaire. (Résultats pour les estimateurs basés sur un modèle d'après 250 000 tirages d'une seule chaîne après un tirage de rodage de 50 000; la convergence a été évaluée par la méthode de Geweke (1992).) Pour les modèles PAE et PAA, $p(\sum) \sim \text{INVERSE-WISHART}(p, 0,1I)$, où $p = 2$ pour le modèle non corrigé et $p = 13$ pour le modèle

corrigé. Dans les résultats non corrigés, les estimateurs PAE et PAA sont compris entre l'estimateur non pondéré et l'estimateur entièrement pondéré, tandis que les estimateurs linéaires et non paramétriques tendent à suivre l'estimateur entièrement pondéré. Dans l'analyse corrigée, les trois estimateurs fondés sur un modèle sont compris entre l'estimateur non pondéré et l'estimateur entièrement pondéré, l'estimateur PAE étant le plus proche de l'estimateur non pondéré et l'estimateur PAL, le plus proche de l'estimateur entièrement pondéré. Si l'on s'en tient aux résultats des simulations, il semble que l'estimateur PAA, qui suggère des risques relatifs de blessure de l'ordre de 7 pour les enfants passagers dans une camionnette compacte avec cabine allongée comparativement à d'autres véhicules, pourrait être un meilleur estimateur du risque relatif que l'estimateur non pondéré ou entièrement pondéré. (À titre d'« évaluation », nous notons que des données portant sur deux années supplémentaires, qui n'étaient pas disponibles au moment de Winston et coll. (2002), englobant 4 091 enfants supplémentaires assis sur le siège arrière de véhicules de transport de passagers [44 dans des camionnettes compactes à cabine allongée] ont donné un rapport de cotes non corrigées entièrement pondérées pour les blessures chez les enfants passagers dans une camionnette compacte à cabine allongée de 6,3, et un rapport de cotes corrigées de 7,0.)

5. Discussion

Les modèles dont il est discuté dans le présent article généralisent les travaux de Lazzeroni et Little (1998), ainsi que d'Elliott et Little (2000) dans lesquels l'inférence à la population a été limitée aux moyennes de population sous des hypothèses de loi gaussienne. Considérer la pondération comme une interaction entre les probabilités d'inclusion et les paramètres du modèle offre un autre paradigme pour la réduction des poids sous forme d'un modèle à effets aléatoires qui lisse les paramètres d'intérêt du modèle sur l'ensemble des classes d'inclusion. Les modèles avec structures de moyenne échangeables offrent le degré le plus important de rétrécissement ou de réduction, mais sont les plus sensibles à l'erreur de spécification du modèle; les modèles dont les moyennes sont fortement structurées pourraient être moins efficaces, mais sont plus robustes à l'erreur de spécification du modèle. Cette propriété de robustesse peut être particulièrement importante, sachant que les éléments des grandes strates d'inclusion donnent le degré le plus important de réduction possible de la variance dans l'estimation fondée sur un modèle, mais sont aussi sujets au degré le plus important de biais et de variance dans le modèle à cause de l'extrapolation.

Nous considérons des simulations sous divers degrés d'erreur de spécification du modèle et d'échantillonnage informatif pour les modèles de régression linéaire et logistique. Les modèles de lissage linéaire et non paramétrique surpassent presque les estimateurs entièrement pondérés en ce qui concerne la réduction de l'erreur quadratique dans les simulations considérées. Le modèle d'échangeabilité manifeste une certaine tendance à surlisser, favorisant la réduction de la variance au détriment de la correction du biais, spécialement dans les conditions de régression linéaire. Tous les estimateurs à lissage des poids ont tendance à avoir une couverture des intervalles de confiance inférieure au taux nominal lorsque l'erreur de spécification du modèle est très forte, quoique dans aucun cas, la couverture n'était catastrophiquement faible. Le modèle de lissage autorégressif, qui permet d'utiliser divers degrés de lissage local dans les strates de poids semble produire un accroissement non négligeable de l'efficacité tout en posant un risque limité de surlissage ou de sous-dénombrement grave.

Tableau 3
Rapport de cotes estimées d'une blessure chez les enfants assis sur le siège arrière d'une camionnette compacte à cabine allongée ($n = 60$) comparativement à ceux assis sur le siège arrière d'autres véhicules ($n = 8\,060$), en utilisant les estimateurs non pondéré (NPD), entièrement pondéré (EPD), à poids réduits à une valeur normalisée de 3 (PEL), de la pente aléatoire à modèle d'échangeabilité (PAE), de la pente aléatoire à modèle autorégressif (PAA), de la pente aléatoire à modèle linéaire (PAL) et de la pente aléatoire à modèle non paramétrique (PANP); non corrigés et corrigés pour l'âge de l'enfant, la gravité de l'accident, la direction de l'impact et le poids du véhicule. Estimations ponctuelles des modèles PAE, PAA et PAL à partir de la médiane a posteriori. Intervalle de confiance ou intervalle prédictif a posteriori à 95 % en indice inférieur. Données provenant du Partners for Child Passenger Safety

	NPD	EPD	PEL	
Non corr.	3,54 (2,01 à 6,23)	11,32 (2,67 à 48,02)	9,15 (2,65 à 31,57)	
Corr.	3,50 (1,88 à 6,53)	14,56 (3,45 à 61,40)	10,99 (2,97 à 34,64)	
	PAE	**PAA**	**PAL**	**PANP**
Non corr.	6,70 (2,51 à 20,92)	6,69 (2,64 à 21,05)	11,17 (3,21 à 24,94)	10,34 (3,27 à 24,62)
Corr.	4,45 (2,39 à 8,67)	6,67 (3,56 à 11,94)	11,87 (3,33 à 36,93)	10,23 (3,02 à 37,93)

L'application des méthodes aux données du Partners for Child Passenger Safety en vue de déterminer le risque excédentaire de blessure lors d'un accident chez les enfants installés sur le siège arrière d'une camionnette compacte à cabine allongée comparativement à ceux installés sur le siège arrière d'autres véhicules pour le transport de passagers, il semble que la décision dans Winston et coll. (2002) d'éliminer de l'analyse un enfant dont la probabilité de sélection était faible en vue de stabiliser les estimations était effectivement prudente. En effet, l'estimateur PAA, favorisé par les mesures de l'EQM dans les simulations suggère un risque excédentaire corrigé de 6,7 avec un IPP à 95 % de (3,6, 11,9), comparativement à celui de 14,6 avec un IC à 95 % de (3.4, 61,4) de l'estimateur entièrement pondéré.

Quoique nous appliquions, dans le présent article, une approche entièrement bayésienne de l'inférence au sujet de la loi prédictive a posteriori de la pente de régression de population, des estimations empiriques bayésiennes (EB) peuvent aussi être obtenues par estimation du maximum de vraisemblance ou du maximum de vraisemblance restreint en utilisant des méthodes standard à modèle mixte linéaire ou linéaire généralisé. Dans les conditions gaussiennes, les estimations EB de G et de σ^2 peuvent être « insérées » dans les expressions explicites de $E(\boldsymbol{B} \mid y, X)$ et $\mathrm{Var}(\boldsymbol{B} \mid y, X)$. Les conditions exponentielles générales posent plus de problèmes. Les estimations insérées peuvent être utilisées pour déterminer $E(\boldsymbol{B} \mid y, X)$ par des méthodes de recherche de racine. L'absence de formules explicites pour $E(\boldsymbol{B} \mid y, X)$ rend difficile l'obtention d'estimateurs empiriques bayésiens basés sur un modèle pour $\mathrm{Var}(\boldsymbol{B} \mid y, X)$. En outre, les estimateurs empiriques bayésiens standard ne tiennent pas compte de l'incertitude dans l'estimation de G.

Nous notons aussi que, bien que le calcul des valeurs réelles de réduction des poids des cas ne soit pas nécessaire dans cette approche, il est possible de déterminer les poids de sondage révisés qu'implique le rétrécissement. Dans les conditions du modèle linéaire, ceux-ci peuvent être obtenus par une application itérative d'un schéma de pondération par calage, tel que les estimateurs par la régression généralisée ou GREG (Deville et Särndal 1992). Les conditions exponentielles générales requièrent l'intégration de l'algorithme de calage des poids dans l'algorithme des moindres carrés itératifs repondérés utilisé pour ajuster un modèle linéaire généralisé.

Lorsque les poids d'échantillonnage sont utilisés pour tenir compte de l'erreur de spécification de la moyenne dans des conditions de régression, on pourrait soutenir que l'approche correcte consiste à spécifier correctement la moyenne afin d'éliminer les divergences entre les estimations entièrement pondérées et non pondérées des paramètres de régression. Cependant, la spécification parfaite est un objectif impossible à atteindre et, même de bonnes approximations pourraient être fortement biaisées si l'on ne tient pas compte des poids des cas quand les probabilités d'échantillonnage sont très variables. Dans les conditions d'échantillonnage informatif, il pourrait être impossible de déterminer si les divergences entre les estimations pondérées et non pondérées sont dues à une erreur de spécification du modèle ou au plan d'échantillonnage proprement dit. Enfin, même les modèles de régression spécifiés incorrectement ont la caractéristique séduisante, dans les conditions de population finie, de produire une seule grandeur de population cible. Par conséquent, nous continuons de recommander de tenir compte de la probabilité d'inclusion sous des modèles linéaires ou des modèle linéaires généralisés, et les méthodes établissant un compromis entre une analyse entièrement pondérée à faible biais et variance élevée et une analyse non pondérée à biais élevé et faible variance demeurent utiles.

Les méthodes dont il est discuté dans le présent article offrent la promesse d'adapter les méthodes fondées sur un modèle pour s'attaquer au problème de l'analyse des données d'enquête. Notre objectif n'est pas d'élaborer un modèle hiérarchique bayésien unique, finement ajusté à un ensemble de données ou à une question particulière, mais plutôt d'élaborer des méthodes robustes, mais néanmoins efficaces, qui peuvent être appliquées dans les conditions « automatisées » à rythme rapide dans lesquelles de nombreux analystes spécialisés dans la recherche par sondage appliquée doivent parfois travailler. Quoiqu'elles demandent de nombreux calculs, les méthodes considérées sont des applications ou des extensions de la « boîte à outils » des modèles à effets aléatoires existants et peuvent être implémentées dans les progiciels statistiques existants ou exécutées à l'aide de méthodes MCMC relativement simples. Notre approche conserve un côté fondé sur le plan en ce sens que nous essayons d'élaborer des méthodes d'estimation bayésiennes « automatisées » fondées sur un modèle qui produisent des inférences robustes dans des conditions d'échantillonnage répété lorsque le modèle proprement dit est spécifié incorrectement. Cependant, parce que ces modèles s'appuient sur la stratification des données par la probabilité de sélection en prélude à l'utilisation des techniques de regroupement ou de rétrécissement en vue d'induire une réduction des poids dicté par les données, il existe une correspondance naturelle entre cette méthodologie et les plans d'échantillonnage (post) stratifiés dans lesquels les strates correspondent aux probabilités inégales d'inclusion. L'élaboration de méthodes adaptées à une classe plus générale de plans d'échantillonnage complexes comprenant des échantillons en grappes à un ou à plusieurs degrés et (ou) des strates qui « recoupent »

les strates d'inclusion demeure un important domaine dans lequel poursuivre de futurs travaux.

Remerciements

L'auteur remercie Roderick J.A. Little, ainsi que le rédacteur en chef, le rédacteur adjoint et deux examinateurs anonymes, de leurs révisions et commentaires. Il remercie aussi les docteurs Dennis Durbin et Flaura Winston du projet de la Partners for Child Passenger Safety de leur aide, ainsi que les compagnies d'assurance State Farm de leur appui pour le projet Partners for Child Passenger Safety. La présente étude a été financée par la subvention R01-HL-068987-01 du National Institute of Heart, Lung and Blood.

Bibliographie

Alexander, C.H., Dahl, S. et Weidman, L. (1997). Making estimates from the American Community Survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 2000, 88-97.

Association for the Advancement of Automotive Medicine (1990). *The Abbreviated Injury Scale*, *1990 Revision*. Association for the Advancement of Automotive Medicine, Des Plaines, Illinois.

Beaumont, J.-F., et Alavi, A. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 217-231.

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

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Durbin, D.R., Bhatia, E., Holmes, J.H., Shaw, K.N., Werner, J.V., Sorenson, W. et Winston, F.K. (2001). Partners for child passenger safety: A unique child-specific crash surveillance system. *Accident Analysis and Prevention*, 33, 407-412.

Elliott, M.R., et Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Ericson, W.A. (1969). Subjective bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society*, Séries B, 31, 195-234.

Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2000, 598-603.

Gelfand, A.E., et Smith, A.M.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 389-409.

Gelman, A., et Carlin, J.B. (2002). Poststratification and weighting adjustments. *Survey Nonresponse*, (Éds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), 289-302.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4*, *Proceedings of the Fourth Valencia International Meeting*, (Éds., J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith), 89-193.

Ghosh, M., et Meeden, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.

Hastie, T.J., et Tibshirani, R.J. (1990). *Generalized Additive Models*, London: Chapman and Hall.

Holt, D., et Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, Séries A, 142, 33-46.

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

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 183-200.

Korn, E.L., et Graubard, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.

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

Korn, E.L., et Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society*, Séries B, 65, 175-190.

Lazzeroni, L.C., et Little, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

Lin, X., et Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society*, Séries B, 61, 381-400.

Little, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R.J.A. (1993). Poststratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J.A. (2004). To model or not model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*, 2ième édition. CRC Press : Boca Raton, Floride.

Oh, H.L., et Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse. *Incomplete Data in Sample Surveys*, (Éds., W.G. Madow, I. Olkin et D.B. Rubin), 2, 143-184.

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

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rabash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Séries B, 60, 23-40.

Potter, F. (1990). A study of procedures to identify and trim extreme sample weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1990, 225-230.

Rizzo, L. (1992). Conditionally consistent estimators using only probabilities of selection in complex sample surveys. *Journal of the American Statistical Association*, 87, 1166-1173.

Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York : John Wiley & Sons, Inc.

Särndal, C.-E. (1980). On π -inverse weighting verses best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

Skinner, C.J., Holt, D. et Smith, T.M.F (1989). *Analysis of Complex Surveys*. New York : John Wiley & Sons, Inc.

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, Séries B, 40, 364-372.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93, 341-348.

Winston, F.K., Kallan, M.K., Elliott, M.R., Menon, R.A. et Durbin, D.R. (2002). Risk of injury to child passengers in compact extended pick-up trucks. *Journal of the American Medical Association*, 287, 1147-1152.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 41-51
Statistique Canada, N° 12-001 au catalogue

Estimation assistée par un modèle semi-paramétrique pour les enquêtes sur les ressources naturelles

F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson et M. Giovanna Ranalli [1]

Résumé

De l'information auxiliaire est souvent utilisée pour améliorer la précision des estimateurs des moyennes et des totaux de population finie grâce à des techniques d'estimation par le ratio ou par la régression linéaire. Les estimateurs résultants ont de bonnes propriétés théoriques et pratiques, dont l'invariance, le calage et la convergence par rapport au plan de sondage. Cependant, il n'est pas toujours certain que les modèles de ratio et les modèles linéaires sont de bonnes approximations de la relation réelle entre les variables auxiliaires et la variable d'intérêt, ce qui cause une perte d'efficacité si le modèle n'est pas approprié. Dans le présent article, nous expliquons comment on peut étendre l'estimation par la régression afin d'intégrer des modèles de régression semi-paramétriques dans le cas de plans de sondage simples ainsi que plus complexes. Tout en retenant les bonnes propriétés théoriques et pratiques des modèles linéaires, les modèles semi-paramétriques reflètent mieux les relations complexes entre les variables, ce qui se traduit souvent par des gains importants d'efficacité. Nous illustrerons l'applicabilité de l'approche à des plans de sondage complexes comportant de nombreux types de variables auxiliaires en estimant plusieurs caractéristiques liées à l'acidification dans le cas d'une enquête sur les lacs du Nord-Est des États-Unis.

Mots clés : Estimation par la régression; lissage; régression à noyau; chimie des lacs.

1. Introduction

La post-stratification, le calage et l'estimation par la régression sont différentes approches qui permettent d'améliorer la précision des estimateurs lorsqu'on dispose d'information auxiliaire à l'étape de l'estimation. L'*estimation assistée par un modèle* (Särndal, Swensson et Wretman 1992) offre un cadre commode dans lequel élaborer ces estimateurs et des estimateurs connexes. Dans ce cadre, un modèle de superpopulation décrit la relation entre la variable d'intérêt et les variables auxiliaires. Ce modèle est alors utilisé pour construire des estimateurs sur échantillon dont la précision est meilleure si le modèle est spécifié correctement, mais qui retiennent d'importantes propriétés relatives au plan de sondage, comme la convergence et la possibilité d'estimer la variance lorsque le modèle est incorrect.

Jusqu'à récemment, les modèles de superpopulation utilisés dans ce contexte étaient formulés comme des modèles paramétriques, le plus souvent des modèles de ratio ou des modèles linéaires. Bien que ce genre de modèles relativement simples soient raisonnables dans beaucoup d'applications pratiques, il existe de nombreuses situations où ils ne donnent pas une bonne représentation de la relation entre la variable d'intérêt et les variables auxiliaires. Breidt et Opsomer (2000) ont proposé un estimateur assisté par un modèle non paramétrique basé sur la régression polynomiale locale qui généralise les estimateurs par la régression paramétrique bien établis. Grâce à cet estimateur, la superpopulation ne doit plus nécessairement posséder une forme paramétrique spécifiée a priori. Au lieu de cela, la relation entre la ou les variables d'intérêt de l'enquête et la variable auxiliaire doit être lisse (continue), mais est par ailleurs entièrement non spécifiée.

Dans le présent article, nous étendons en termes formels la théorie de Breidt et Opsomer (2000) au contexte de la régression semi-paramétrique, dans lequel certaines variables sont intégrées linéairement et d'autres le sont à l'aide de termes additifs lisses. Cette extension rend les résultats de ces auteurs plus utiles en pratique, puisque l'information auxiliaire est très fréquemment de nature multidimensionnelle et qu'elle contient presque toujours des variables catégoriques qui doivent entrer paramétriquement dans le modèle de régression (grâce à l'utilisation de variables indicatrices). En guise d'illustration, nous utilisons une enquête sur les lacs des États du Nord-Est des États-Unis réalisées par l'Environmental Monitoring and Assessment Program de l'Environmental Protection Agency des États-Unis. Cette enquête porte sur 334 lacs échantillonnés parmi une population de 21 026 lacs entre 1991 et 1996. Nous appliquons l'estimateur assisté par un modèle semi-paramétrique pour produire des estimations de la moyenne et de la fonction de répartition de la *capacité de neutralisation des acides* et d'autres variables de composition chimique d'intérêt. Dans cette application, nous introduisons linéairement dans le modèle des variables catégoriques ainsi que continues, de même qu'une variable continue à titre de terme additif lisse.

1. F. Jay Breidt, Department of Statistics, Colorado State University, Fort Collins CO 80523, É.-U.; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames, IA 50011, É.-U. Courriel : jopsomer@iastate.edu; Alicia A. Johnson, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis MN 55455, É.-U.; M. Giovanna Ranalli, Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli, 06123 Perugia, Italie.

Dans Opsomer, Breidt, Moisen et Kauermann (2007), le principe de l'estimation assistée par un modèle non paramétrique a été étendu à des modèles additifs généralisés (MAG) et appliqué dans un modèle d'interaction pour estimer les variables provenant des enquêtes de Forest Inventory and Analysis. Bien que les MAG contiennent aussi un mélange de termes catégoriques (paramétriques) et non paramétriques, ils ne permettent pas d'arriver à un développement théorique complet, si bien que ce dernier n'y a pas été présenté. Le modèle semi-paramétrique faisant l'objet du présent article peut être considéré comme un cas particulier d'un MAG comportant une fonction de lien identité. Contrairement au MAG « général », le modèle semi-paramétrique permet la dérivation formelle des propriétés statistiques de l'estimateur assisté par un modèle.

La présentation de la suite de l'article est la suivante. À la section 2, nous définissons l'estimateur assisté par un modèle semi-paramétrique. À la section 3, nous énonçons et prouvons les propriétés de l'estimateur par rapport au plan de sondage. À la section 4, nous décrivons l'application de l'estimation assistée par un modèle semi-paramétrique aux données sur les lacs du Nord-Est . Enfin, à la section 5, nous présentons nos conclusions.

2. Estimation assistée par un modèle semi-paramétrique

Pour commencer, nous examinons le modèle de superpopulation contenant un seul terme non paramétrique univarié et une composante paramétrique; l'extension à plusieurs termes non paramétriques est abordée à la section 3.2. La composante paramétrique peut être constituée d'un nombre arbitraire de termes linéaires. Il s'agit du modèle semi-paramétrique étudié, entre autre, par Speckman (1988). Ce modèle de superpopulation, que nous dénotons par ξ, peut s'écrire sous la forme

$$\mathrm{E}_\xi(y_k) = g(x_k,\ \boldsymbol{z}_k) = m(x_k) + \boldsymbol{z}_k\boldsymbol{\beta}$$

$$\mathrm{Var}_\xi(y_k) = v(x_k,\ \boldsymbol{z}_k) \qquad (1)$$

où x_k est une variable auxiliaire continue devant être modélisée non paramétriquement et $\boldsymbol{z}_k = (z_{1k}, ..., z_{Dk})$ est un vecteur de D variables auxiliaires catégoriques ou continues qui sont spécifiées paramétriquement. Les fonctions $m(\cdot)$ et $v(\cdot, \cdot)$, ainsi que le vecteur de paramètres $\boldsymbol{\beta}$ sont inconnus. Pour les besoins d'identifiabilité, nous supposerons que le vecteur $\boldsymbol{z}_k$ contient un terme d'ordonnée à l'origine et que la fonction $m(\cdot)$ est centrée autour de 0 en ce qui concerne la distribution de x_k. Nous dériverons l'estimateur assisté par un modèle utilisant le modèle (1) en commençant par définir des estimateurs de population pour les fonctions et paramètres inconnus, puis en construisant des estimateurs sur échantillon. Cette approche est la même que celle utilisée pour le cas paramétrique dans Särndal et coll. (1992, chapitre 6).

Soit $U = \{1, 2, ..., N\}$ les étiquettes ordonnées pour une population finie d'intérêt. À titre d'estimateur de population de $g(x_k, \boldsymbol{z}_k)$, nous utiliserons l'estimateur par rétro-justement (backfitting estimator) décrit dans Opsomer et Ruppert (1999). Nous commençons par présenter la notation requise. Soit $K(\cdot)$ une fonction noyau utilisée pour définir les voisinages dans lesquels les polynômes locaux seront ajustés (les hypothèses concernant les K sont spécifiées en annexe). Le vecteur de lisseurs de population pour la régression par polynômes locaux de degré p au point d'observation x_k est défini comme étant

$$\boldsymbol{s}_{Uk}^T = \boldsymbol{e}_1^T (\boldsymbol{X}_{Uk}^T \boldsymbol{W}_{Uk} \boldsymbol{X}_{Uk})^{-1} \boldsymbol{X}_{Uk}^T \boldsymbol{W}_{Uk}$$

où $\boldsymbol{e}_1$ est un vecteur de longueur $p+1$ avec une valeur 1 à la première position et des valeurs 0 ailleurs, $\boldsymbol{W}_{Uk} = \mathrm{diag}\{h^{-1}K((x_1 - x_k)/h), ..., h^{-1}K((x_N - x_k)/h)\}$ et

$$\boldsymbol{X}_{Uk} = \begin{bmatrix} 1 & x_1 - x_k & \dots & (x_1 - x_k)^p \\ \vdots & & \ddots & \vdots \\ 1 & x_N - x_k & \dots & (x_N - x_k)^p \end{bmatrix}.$$

Le lisseur $\boldsymbol{s}_{Uk}$ peut être appliqué au vecteur $\boldsymbol{Y}_U = (y_1, ..., y_N)^T$ pour produire l'ajustement de la régression non paramétrique en fonction de la variable x au point d'observation x_k. Il peut aussi être appliqué à n'importe quelle colonne de $\boldsymbol{Z}_U = (\boldsymbol{z}_1^T, ..., \boldsymbol{z}_N^T)^T$ pour lisser celle-ci par rapport à x. Nous le ferons lors de la dérivation des propriétés de l'estimateur semi-paramétrique (section 3).

En plus du vecteur de lisseurs au point d'observation x_k, $\boldsymbol{s}_{Uk}^T$, nous devons définir la matrice de lisseurs à tous les points d'observation $x_1, ..., x_N$,

$$\boldsymbol{S}_U = \begin{bmatrix} \boldsymbol{s}_{U1}^T \\ \vdots \\ \boldsymbol{s}_{UN}^T \end{bmatrix},$$

et la matrice de lisseurs centrée $\boldsymbol{S}_U^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T / N)\boldsymbol{S}_U$. Lorsqu'elle est appliquée à $\boldsymbol{Y}_U$, la matrice de lisseurs produit le vecteur des ajustements de la régression non paramétrique à tous les points d'observation. La matrice de lisseurs centrée $\boldsymbol{S}_U^*$ produit les ajustements centrés, ce qui signifie que la moyenne globale des valeurs ajustées est soustraite de chaque valeur ajustée. Le centrage est utilisé pour préserver l'identifiabilité des estimateurs, comme l'ont expliqué Opsomer et Ruppert (1999).

Pour toute observation x_k, un estimateur possible de $m(x_k)$ pourrait être défini comme étant $\boldsymbol{s}_{Uk}^T Y_U$, avec ou sans ajustement de centrage. Cet estimateur serait généralement médiocre, puisqu'il ne tient pas compte du fait que les y_k contiennent une composante paramétrique qui

dépend des z_k. Un estimateur plus efficace est obtenu en estimant conjointement $m(\cdot)$ et β, comme le fait l'ensemble suivant d'estimateurs

$$\boldsymbol{B} = (\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Z}_U)^{-1}\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Y}_U$$

$$m_k = \boldsymbol{s}_{Uk}^T(\boldsymbol{Y}_U - \boldsymbol{Z}_U\boldsymbol{B}) \quad k = 1, \ldots, N. \tag{2}$$

Dans ces estimateurs, $\boldsymbol{B}$ est calculé pour commencer, puis le « vecteur résiduel » $\boldsymbol{Y}_U - \boldsymbol{Z}_U\boldsymbol{B}$ est lissé en fonction de x. Les estimateurs dans (2) sont identiques aux *estimateurs par rétro-ajustement* pour les modèles additifs décrits dans Hastie et Tibshirani (1990) et implémentés dans `gam` dans S-Plus, R ou SAS. À titre d'estimateur de population de $\mathrm{E}_\xi(y_k) = g(x \mid k, z_k)$, nous utilisons

$$g_k = m_k + z_k\boldsymbol{B}.$$

Nous expliquons maintenant comment construire un estimateur assisté par modèle fondé sur l'approche de régression semi-paramétrique. Soit $A \subset U$ un échantillon de taille n tiré à partir de U conformément au plan d'échantillonnage $p(A)$ avec les probabilités d'inclusion uni et bidimensionnelles $\pi_k = \sum_{A \ni k} p(A)$, $\pi_{kl} = \sum_{A \ni k,l} p(A)$, respectivement. Si les g_k, $k = 1, \ldots, N$ étaient disponibles, il serait possible de construire un estimateur par différence pour la moyenne de population de y_k, $\bar{y}_N = \sum_U y_k / N$ de la forme

$$\hat{y}_{\mathrm{dif}} = \frac{1}{N}\sum_U g_k + \frac{1}{N}\sum_A \frac{y_k - g_k}{\pi_k}, \tag{3}$$

qui est sans biais par rapport au plan et possède une variance due au plan

$$\mathrm{Var}_p(\hat{y}_{\mathrm{dif}}) = \frac{1}{N^2}\sum\sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l}$$

(Särndal et coll. 1992, page 221). La variance due au plan est faible si les écarts entre y_k et g_k sont petits. Cet estimateur n'est pas faisable, car son calcul nécessite la connaissance de toutes les valeurs de x_k, z_k et y_k pour la population. Nous construirons plutôt un estimateur faisable en remplaçant les g_k par des estimateurs fondés sur un échantillon. Les estimateurs fondés sur un échantillon correspondant aux estimateurs de population donnés en (2) sont construits de la façon suivante. Le vecteur pondéré selon le plan de lisseurs polynômiaux locaux est

$$\boldsymbol{s}_{Ak}^{0T} = \boldsymbol{e}_1^T(\boldsymbol{X}_{Ak}^T\boldsymbol{W}_{Ak}\boldsymbol{X}_{Ak})^{-1}\boldsymbol{X}_{Ak}^T\boldsymbol{W}_{Ak}, \tag{4}$$

avec $\boldsymbol{X}_{Ak}$ contenant les lignes de $\boldsymbol{X}_{Uk}$ qui correspondent à $k \in A$ et

$$\boldsymbol{W}_{Ak} = \mathrm{diag}\left\{\frac{1}{\pi_j h}K\left(\frac{x_j - x_k}{h}\right) : j \in A\right\}.$$

La matrice $\boldsymbol{X}_{Ak}^T\boldsymbol{W}_{Ak}\boldsymbol{X}_{Ak}$ dans (4) sera singulière si, pour un échantillon A, il existe moins que $p+1$ observations pour appuyer le noyau à un point d'observation x_k. En pratique, ce problème peut être évité en sélectionnant une fenêtre suffisamment large pour rendre la matrice inversible. Cependant, cette situation ne peut pas être écartée en général et nous avons besoin d'un estimateur qui existe pour chaque échantillon A pour les calculs théoriques de la section 3. Donc, nous considérerons le vecteur ajusté de lisseurs d'échantillon

$$\boldsymbol{s}_{Ak}^T = \boldsymbol{e}_1^T(\boldsymbol{X}_{Ak}^T\boldsymbol{W}_{Ak}\boldsymbol{X}_{Ak} + \mathrm{diag}(\delta N^{-2}))^{-1}\boldsymbol{X}_{Ak}^T\boldsymbol{W}_{Ak}, \tag{5}$$

pour une distance $\delta > 0$ faible, comme l'ont fait Breidt et Opsomer (2000). La matrice de lisseurs d'échantillon et sa version centrée sont données par

$$\boldsymbol{S}_A = [\boldsymbol{s}_{Ak}^T : k \in A] \quad \boldsymbol{S}_A^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T\boldsymbol{\Pi}_A^{-1}/N)\boldsymbol{S}_A$$

avec $\boldsymbol{\Pi}_A = \mathrm{diag}\{\pi_k : k \in A\}$. Les estimateurs pondérés selon le plan de $\boldsymbol{B}$ et des m_k sont

$$\hat{\boldsymbol{B}} = (\boldsymbol{Z}_A^T\boldsymbol{\Pi}_A^{-1}(\boldsymbol{I} - \boldsymbol{S}_A^*)\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}_A^T\boldsymbol{\Pi}_A^{-1}(\boldsymbol{I} - \boldsymbol{S}_A^*)\boldsymbol{Y}_A \tag{6}$$

$$\hat{m}_k = \boldsymbol{s}_{Ak}^T(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T\hat{\boldsymbol{B}}), \tag{7}$$

où $\boldsymbol{Z}_A$ et $\boldsymbol{Y}_A$ dénotent les versions d'échantillon de $\boldsymbol{Z}_U$ et $\boldsymbol{Y}_U$, respectivement. Notons que l'estimateur $\hat{m}_k$ est défini pour tout point d'observation x_k dans la population, et non pas uniquement pour ceux figurant dans l'échantillon. Comme pour les estimateurs de population, ces estimateurs peuvent s'écrire comme étant la solution d'équations de rétro-ajustement, de sorte qu'ils peuvent être calculés grâce à des versions convenablement pondérées des algorithmes existants. L'estimateur de g_k est

$$\hat{g}_k = \hat{m}_k + z_k\hat{\boldsymbol{B}}.$$

Nous construisons alors l'estimateur assisté par un modèle semi-paramétrique en remplaçant g_k par $\hat{g}_k$ dans (3) :

$$\hat{y}_{\mathrm{reg}} = \frac{1}{N}\sum_U \hat{g}_k + \frac{1}{N}\sum_A \frac{y_k - \hat{g}_k}{\pi_k}. \tag{8}$$

En définissant $\bar{y}_\pi = \sum_A y_k/\pi_k$ et en faisant de même pour $\bar{z}_\pi$, nous obtenons pour $\hat{y}_{\mathrm{reg}}$ une expression équivalente donnée par

$$\hat{y}_{\mathrm{reg}} = \bar{y}_\pi + (\bar{z}_N - \bar{z}_\pi)\hat{\boldsymbol{B}} + \frac{1}{N}\sum_U \hat{m}_k - \frac{1}{N}\sum_A \frac{\hat{m}_k}{\pi_k}, \tag{9}$$

qui montre que l'estimateur semi-paramétrique peut être interprété comme étant un estimateur par la régression linéaire « classique » utilisant la composante paramétrique du modèle $z\beta$, avec un terme de correction supplémentaire pour la composante non paramétrique du modèle. Cet

estimateur a également certaines propriétés souhaitables en commun avec les estimateurs par la régression entièrement paramétrique. Il est invariant par rapport à l'échelle et à la localisation, et il est calé pour les composantes de modélisation paramétriques ainsi que non paramétriques, en ce sens que $\hat{x}_{\text{reg}} = \overline{x}_N$ et $\hat{\boldsymbol{z}}_{\text{reg}} = \overline{\boldsymbol{z}}_N$. Le calage pour les variables figurant dans le terme paramétrique peut être vérifié directement en utilisant les expressions (6) et (7), tandis que le calage pour la variable spécifiée non paramétriquement x_k découle du fait que $\boldsymbol{s}_{Ak}^T \boldsymbol{X}_A = x_k$, où $\boldsymbol{X}_A = (x_k : k \in A)^T$ (nous ignorons l'effet de l'ajustement $\text{diag}(\delta N^{-2})$ dans (5), parce que ce dernier peut être rendu arbitrairement petit). En outre, l'estimateur peut s'écrire comme une somme pondérée des y_k, $k \in A$, de sorte qu'il est possible d'obtenir et d'appliquer un ensemble de poids w_k à n'importe quelle variable d'enquête d'intérêt.

3. Propriétés et extensions

3.1 Propriétés relatives au plan de sondage

Dans la présente section, nous explorons les propriétés relatives au plan de sondage de l'estimateur semi-paramétrique (8). Plus précisément, nous prouvons que $\hat{y}_{\text{reg}}$ est convergeant par rapport au plan au taux $\sqrt{n}$ et nous dérivons sa loi asymptotique, y compris une variance estimée. Cet exercice est fait dans le contexte de convergence asymptotique par rapport au plan utilisé dans Isaki et Fuller (1982) et dans Breidt et Opsomer (2000), dans lequel la taille de la population ainsi que celle des échantillons augmente quand $N \to \infty$. Toutes les preuves et les hypothèses nécessaires sont données en annexe.

Dans le théorème qui suit, nous prouvons la convergence par rapport au plan de l'estimateur semi-paramétrique. Nous montrons aussi que le taux de convergence est $\sqrt{n}$, c'est-à-dire le taux habituel pour les estimateurs fondés sur le plan

Théorème 3.1 *Sous les hypothèses A1 à A8, l'estimateur* $\hat{y}_{\text{reg}}$ *donné par (8) est convergent par rapport au plan au taux* $\sqrt{n}$, *en ce sens que*

$$\hat{y}_{\text{reg}} = \overline{y}_N + O_p\left(\frac{1}{\sqrt{n}}\right).$$

Le théorème qui suit prouve qu'il existe un théorème de la limite centrale pour $\hat{y}_{\text{reg}}$ s'il existe pour l'estimateur par extension $\overline{y}_\pi$.

Théorème 3.2 *Sous les hypothèses A1 à A8, si*

$$\frac{\overline{y}_\pi - \overline{y}_N}{\sqrt{\hat{V}(\overline{y}_\pi)}} \to N(0, 1),$$

avec

$$\hat{V}(\overline{y}_\pi) = \frac{1}{N^2} \sum\sum_A \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

pour un plan d'échantillonnage donné, alors nous avons aussi

$$\frac{\hat{y}_{\text{reg}} - \overline{y}_N}{\sqrt{\hat{V}(\hat{y}_{\text{reg}})}} \to N(0, 1),$$

avec

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum\sum_A \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - \hat{g}_k}{\pi_k} \frac{y_l - \hat{g}_l}{\pi_l}. \quad (10)$$

3.2 Modèle additif semi-paramétrique

Les résultats des théorèmes 3.1 et 3.2 sont obtenus en utilisant le modèle semi-paramétrique (1), qui contient un seul terme non paramétrique univarié $m(\cdot)$. Dans de nombreuses applications pratiques, plusieurs variables auxiliaires susceptibles d'être incluses dans la partie non paramétrique du modèle sont disponibles, mais la malédiction de la dimensionnalité fait qu'il est souvent difficile de combiner plusieurs variables en un seul terme non paramétrique multidimensionnel. Au lieu de cela, nous traitons les variables qui doivent être incluses non paramétriquement comme des composantes univariées. Nous obtenons ainsi le *modèle additif semi-paramétrique*, qui s'écrit

$$\begin{aligned} \text{E}_\xi(y_k) &= g(\boldsymbol{x}_k, \boldsymbol{z}_k) = m_1(x_{1k}) + \ldots + m_Q(x_{Qk}) + \boldsymbol{z}_k \boldsymbol{\beta} \\ \text{Var}_\xi(y_k) &= v(\boldsymbol{x}_k, \boldsymbol{z}_k) \end{aligned}$$

où les $m_q(\cdot)$, $q = 1, \ldots, Q$ et $v(\cdot, \cdot)$ sont des fonctions lisses inconnues.

Quand $Q = 2$, des expressions semblables à (6) et (7) peuvent être développées en utilisant les décompositions du modèle additif de Opsomer et Ruppert (1997) et, quand $Q > 2$, des expressions récursives peuvent être dérivées en utilisant l'approche d'Opsomer (2000). L'estimateur s'écrirait alors comme dans les équations (6) et (7), mais en remplaçant les vecteurs de lisseurs $\boldsymbol{s}_{Ak}$ et la matrice de lisseurs $\boldsymbol{S}_A$ par des lisseurs compliqués à modèle additif de plus grande dimensionnalité (voir Opsomer (2000) pour des détails). Par conséquent, prouver les propriétés de l'estimateur assisté par modèle dans le cas où la valeur de Q est arbitraire serait une tâche difficile dépassant le cadre du présent article.

En pratique, la formulation de l'algorithme de rétro-ajustement offre un moyen nettement plus efficace et plus simple de calculer l'estimateur semi-paramétrique. Soit $\boldsymbol{s}_{Aqk}$ le vecteur de lisseurs d'échantillon, tel qu'il est défini en (5), pour la variable x_q au point d'observation x_{qk} et

$\boldsymbol{S}_{Aq}$, la matrice de lisseurs correspondante pour la variable x_q. En outre, $\hat{m}_{qk}$ dénote l'estimateur par rétro-ajustement pondéré selon l'échantillon pour $m_q(x_{qk})$ et $\hat{\boldsymbol{m}}_{Aq} = (\hat{m}_{qk},\ k \in A)$. L'algorithme de rétro-ajustement pour un modèle contenant des termes non paramétriques Q est constitué de l'ensemble d'équations suivant, itéré jusqu'à convergence :

$$\hat{\boldsymbol{B}} = (\boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1} \boldsymbol{Z}_A)^{-1} \boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1} \left(\boldsymbol{Y}_A - \sum_{q=1}^{Q} \hat{\boldsymbol{m}}_{Aq} \right)$$

$$\hat{\boldsymbol{m}}_{A1} = \boldsymbol{S}_{A1} \left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}} - \sum_{q \neq 1} \hat{\boldsymbol{m}}_{Aq} \right)$$

$$\vdots \quad \vdots$$

$$\hat{\boldsymbol{m}}_{AQ} = \boldsymbol{S}_{AQ} \left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}} - \sum_{q \neq Q} \hat{\boldsymbol{m}}_{Aq} \right).$$

Ces équations fournissent des ajustements pondérés uniquement aux points d'observation (localisations) compris dans l'échantillon $k \in A$. Pour les autres points d'observation $k \in U$ non compris dans A, une étape de lissage supplémentaire est nécessaire après l'obtention de $\hat{\boldsymbol{m}}_{Aq}$, $q = 1, ..., Q$:

$$\hat{m}_{kq} = \boldsymbol{s}_{Aqk}^T \left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}} - \sum_{q' \neq q} \hat{\boldsymbol{m}}_{Aq'} \right).$$

Les estimateurs fondés sur l'échantillon de la fonction moyenne à tous les points d'observation $k \in U$ sont alors définis comme étant $\hat{g}_k = \hat{m}_{k1} + ... + \hat{m}_{kQ} + \boldsymbol{z}_k \hat{\boldsymbol{B}}$, qui sont utilisés dans l'expression (8) pour construire l'estimateur assisté par modèle.

4. Application à l'enquête sur les lacs du Nord-Est

À la présente section, nous démontrons l'applicabilité de l'estimateur par la régression semi-paramétrique à un ensemble de données sur la composition chimique d'échantillons d'eau. Comme nous l'illustrerons, une fois que l'on a choisi un ensemble de variables auxiliaires et un modèle, il est aussi facile de calculer des estimateurs pour le modèle semi-paramétrique que pour les modèles linéaires, ce qui peut donc aboutir à une amélioration de la précision à relativement peu de frais.

La National Surface Water Survey (NSWS) parrainée par l'Environmental Protection Agency (EPA) des États-Unis entre 1984 et 1996 a permis d'estimer que 4,2 % des lacs de la région du Nord-Est des États-Unis étaient acides (Stoddard, Kahl, Deviney, DeWalle, Driscoll, Herlihy, Kellogg, Murdoch, Webb et Webster 2003). Les lacs du Nord-Est sensibles aux acides faisaient partie des préoccupations auxquelles visait à répondre le Clean Air Act Amendment (CAAA) de 1990, grâce à la restriction des émissions industrielles de souffre et d'azote en vue de réduire l'acidité de ces eaux. Une mesure courante de l'acidité est la capacité de neutralisation des acides (CNA), qui est définie comme étant le pouvoir tampon de l'eau, c'est-à-dire la capacité de l'eau de résister aux variations de l'acidité. Une valeur de la CNA inférieure à zéro $\mu éq / L$ indique que l'eau a perdu son pouvoir tampon. Les eaux de surface dont la valeur de la CNA est inférieure à 200 $\mu éq / L$ sont considérées comme courant un risque d'acidification, et les valeurs inférieures à 50 $\mu éq / L$ sont considérées comme posant un risque élevé (National Acid Precipitation Assessment Program (1991), page 15).

Entre 1991 et 1996, l'Environmental Monitoring and Assessment Program (EMAP) de l'Environmental Protection Agency des États-Unis a réalisé une enquête sur lacs des États du Nord-Est des États-Unis. Les données ont été recueillies afin de déterminer l'effet que des restrictions imposées par le CAAA ont eu sur les conditions écologiques de ces eaux. Parmi une population de 21 026 lacs, on en a sélectionné pour l'enquête 334, dont certains ont été visités plusieurs fois durant la période d'étude. On a calculé la moyenne des mesures multiples faites sur un même lac afin d'obtenir une seule mesure par lac échantillonné. Les lacs étudiés ont été sélectionnés selon un plan d'échantillonnage complexe fondé sur une base de sondage à grille hexagonale utilisée fréquemment par l'EMAP (voir Larsen, Thornton, Urquhart et Paulsen (1993) pour une description du plan d'échantillonnage).

Soit y_k la valeur (éventuellement moyenne) de la CNA du k^{e} lac échantillonné. Une estimation très simple de la moyenne des CNA des lacs est représentée par l'estimateur à facteur d'extension $\bar{y}_\pi$. Ici, comme dans le cas de nombreuses enquêtes, un meilleur choix est l'estimateur de Hájek,

$$\hat{y}_H = \frac{1}{\hat{N}} \sum_{k \in A} \frac{y_k}{\pi_k}, \qquad (11)$$

qui applique un ajustement de type ratio pour l'estimation de la taille de population à l'aide de $\hat{N} = \sum_{k \in A} 1/\pi_k$. Cependant, des variables auxiliaires sont disponibles pour chaque lac dans cette population, si bien qu'il devrait être possible d'améliorer davantage l'efficacité de l'estimateur de Hájek. Les variables qui suivent sont disponibles pour chaque lac $k \in U$:

- x_k = UTMX, coordonnée géographique x du centroïde de chaque lac dans le système de coordonnées UTM,
- $z_{j,k}$ = variable indicatrice pour l'écorégion $j = 1, ..., 6$;
- $z_{7,k}$ = UTMY, coordonnée géographique y ;
- $z_{8,k}$ = élévation.

Comme sept écorégions différentes sont incluses dans la population, nous construisons des variables muettes $z_{j,k}$ pour $j = 1, ..., 6$. Nous construirons un estimateur par la régression semi-paramétrique pour la variable y en traitant la variable UTMX x comme un terme non paramétrique et les autres variables $z_1 - z_8$ comme une composante paramétrique. Une approche de sélection de modèle nous a permis de déterminer que le fait de traiter les deux autres variables continues comme étant non paramétriques n'amélioraient pas l'ajustement du modèle. Aux fins de comparaison, nous avons également calculé un estimateur par la régression qui traite tous les termes comme étant paramétriques. Cet estimateur est par conséquent identique à l'estimateur semi-paramétrique, excepté que la coordonnée géographique x est modélisée linéairement. Nous dénoterons cet estimateur par la régression entièrement paramétrique par $\hat{y}_{\text{par}}$.

Afin de déterminer l'efficacité estimée des estimateurs d'après les données d'enquête, nous devons calculer les estimations de la variance. Cependant, comme les probabilités d'inclusion de deuxième ordre n'étaient pas disponibles, nous ne pouvons pas évaluer $\hat{V}(\hat{y}_{\text{reg}})$ comme dans (10). Afin de produire des estimations appropriées de la variance, nous traitons le plan d'échantillonnage complexe comme un plan d'échantillonnage stratifié avec remise. Les 14 strates que nous avons sélectionnées correspondent à des groupes de grappes spatiales de lacs qui figuraient dans le plan d'échantillonnage original et qui ont été utilisées pour assurer la répartition spatiale des lacs échantillonnés sur la région d'intérêt. Larsen et coll. (1993) donnent des précisions sur la construction des grappes spatiales.

Soit H le nombre de strates, n_h, le nombre d'observations dans la strate h, et A_h l'ensemble d'éléments échantillonnés qui sont compris dans la strate h. Définissons $p_k = n_h^{-1}\pi_k$. En utilisant cette notation et l'hypothèse d'un échantillon stratifié avec remise, nous réécrivons l'estimateur semi-paramétrique sous la forme

$$\hat{y}_{\text{reg}} = \frac{1}{N}\sum_{k\in U}\hat{g}(x_k, z_k) + \frac{1}{N}\sum_{h\in H}\frac{1}{n_h}\sum_{k\in A_h}\frac{y_k - \hat{g}(x_k, z_k)}{p_k} \qquad (12)$$

et l'estimateur de la variance sous la forme

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2}\sum_{h\in H} S_h^2,$$

où S_h^2 est la variance résiduelle pondérée intrastrate estimée pour la strate h. En supposant que les strates sont échantillonnées avec remise, Särndal et coll. (1992, pages 421-422) proposent de calculer S_h^2 comme il suit :

$$S_h^2 = \frac{1}{n_h(n_h-1)}\sum_{k\in A_h}\left(\frac{y_k - \hat{g}(x_k, z_k)}{p_k} - \sum_{l\in A_h}\frac{y_l - \hat{g}(x_l, z_l)}{\pi_l}\right)^2. \qquad (13)$$

De même, nous estimons $\hat{V}(\hat{y}_H)$ par

$$\hat{V}(\hat{y}_H) = \frac{1}{\hat{N}^2}\sum_{h\in H}\frac{1}{n_h(n_h-1)}\sum_{k\in A_h}\left(\frac{y_k - \hat{y}_H}{p_k} - \sum_{l\in A_h}\frac{y_l - \hat{y}_H}{\pi_l}\right)^2, \qquad (14)$$

et nous obtenons l'expression pour $\hat{V}(\hat{y}_{\text{par}})$ de façon entièrement analogue à celle utilisée pour $\hat{V}(\hat{y}_{\text{reg}})$, excepté que $\hat{g}(x_k, z_k)$ est calculé par régression linéaire.

Ces conditions nous permettent d'obtenir les estimations suivantes de la CNA moyenne pour les lacs du Nord-Est, ainsi que les estimations de la variance et les intervalles de confiance (IC) à 95 % approximatif. Nous avons recouru à un ajustement linéaire local pour le terme non paramétrique avec la largeur de la fenêtre fixée à un dixième de l'étendue de l'UTMX.

$\hat{y}_{\text{reg}} = 558{,}0\ \mu\text{éq/L}$ $\quad \hat{V}(\hat{y}_{\text{reg}}) = 2534{,}6$ $\quad \text{IC} = (459{,}3; 656{,}6)$

$\hat{y}_{\text{par}} = 577{,}3\ \mu\text{éq/L}$ $\quad \hat{V}(\hat{y}_{\text{par}}) = 3239{,}6$ $\quad \text{IC} = (465{,}8; 688{,}9)$

$\hat{y}_H = 555{,}9\ \mu\text{éq/L}$ $\quad \hat{V}(\hat{y}_H) = 4313{,}3$ $\quad \text{IC} = (427{,}2; 684{,}7)$

L'intervalle de confiance construit en utilisant l'estimateur de Hájek est environ 31 % plus large que celui construit en utilisant l'estimateur semi-paramétrique, tandis que l'intervalle pour l'estimateur par la régression entièrement paramétrique est 13 % plus large. Ces résultats donnent la preuve d'une amélioration de l'efficacité due au fait de tenir compte de l'information auxiliaire de façon paramétrique ainsi que non paramétrique dans la procédure d'estimation de la moyenne, l'estimateur non paramétrique étant capable de fournir une efficacité supplémentaire en sus de celle de l'estimateur paramétrique.

Comme nous l'avons mentionné plus haut, un objectif important de cette application est de déterminer combien de lacs risques d'être acidifiés ou le sont déjà. Autrement dit, nous cherchons à estimer la proportion de lacs du Nord-Est dont la valeur de la CNA est inférieure à une valeur seuil particulière. Nous pouvons déterminer ce genre de proportion en estimant la fonction de répartition en population finie,

$$F_N(t) = \frac{1}{N}\sum_{k\in U} I_{\{y_k \le t\}}$$

à des valeurs seuil particulières t, où $I_{\{y_k \le t\}}$ dénote la fonction indicatrice prenant une valeur de 1 si $y_k \le t$ et 0 autrement. Puisque les trois estimateurs peuvent être exprimés sous forme de sommes pondérées d'observations d'échantillon, les poids obtenus pour chacun peuvent être appliqués directement à la fonction indicatrice $I_{\{y_k \le t\}}$ pour l'échantillon afin d'estimer $F_N(t)$ pour toute valeur t souhaitée. Soit $\hat{F}_H(t)$, $\hat{F}_{\text{reg}}(t)$ et $\hat{F}_{\text{par}}(t)$ les estimateurs par la régression de Hájek, semi-paramétrique et paramétrique de la fonction de répartition, respectivement. Les estimations de leur variance due au plan sont calculés en introduisant les variables indicatrices dans les équations (13) et (14).

La figure 1 montre les estimations de la fonction de répartition de la CNA produites par $\hat{F}_H(t)$, $\hat{F}_{\text{par}}(t)$ et $\hat{F}_{\text{reg}}(t)$ évaluées sur une grille de 1 000 valeurs uniformément espacées pour t. Sont inclus leurs intervalles de confiance à 95 % ponctuels respectifs calculés en chaque point de la grille. Les trois estimateurs sont comparables, mais les bandes de confiance pour les estimateurs par la régression paramétrique et semi-paramétrique ont tendance à être plus étroites. Si l'on calcule la moyenne sur l'ensemble des 1 000 points de la grille, les largeurs des bandes de confiance sont égales à 0,093 pour $\hat{F}_H(t)$, 0,084 pour $\hat{F}_{\text{par}}(t)$ et 0,075 pour $\hat{F}_{\text{reg}}(t)$, respectivement.

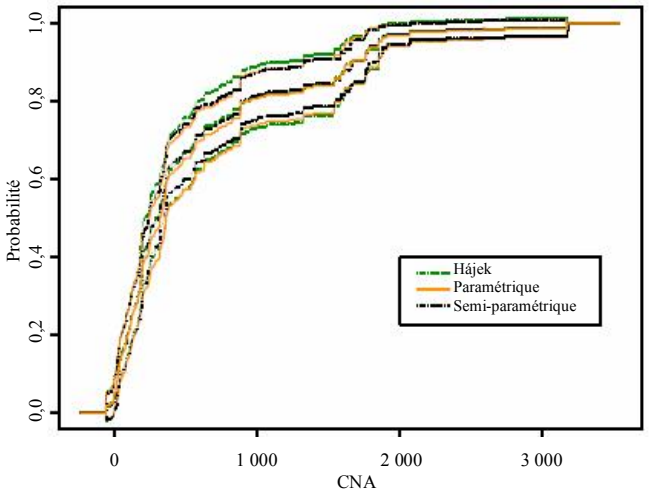


Figure 1
Estimation de la fonction de répartition de population de la CNA et limites de confiance produites par les estimateurs par la régression de Hájek, paramétrique et semi-paramétrique

En plus de la CNA, l'enquête du EMAP sur les lacs du Nord-Est visait à mesurer la concentration de plusieurs substances chimiques, dont les sulfates, le magnésium et les chlorures, de sorte que les poids de sondage obtenus pour la CNA peuvent également être appliqués à ces concentrations, ainsi qu'à leurs fonctions de répartition respectives. En guise d'illustration supplémentaire de l'approche par estimation semi-paramétrique, il est possible d'« inverser » $\hat{F}_{\text{reg}}(t)$ pour obtenir les estimateurs des quantiles $\hat{\theta}_{\text{reg}}(\alpha) = \min\{t : \hat{F}_{\text{reg}}(t) \ge \alpha\}$ pour ces variables de composition chimique supplémentaires. Le tableau 1 donne les estimations semi-paramétriques des premier, deuxième et troisième quartiles pour les concentrations de sulfates, de magnésium et de chlorures exprimées en $(\mu\text{éq}/L)$. L'estimation de la variance pour ces quantiles pourrait être traitée en utilisant les résultats asymptotiques de Francisco et Fuller (1991), mais nous ne nous pencherons pas sur ce problème ici.

Tableau 1 Estimation des quartiles des variables chimiques

α	Sulfates	Magnésium	Chlorures
0,25	73,3	63,8	27,4
0,50	104,3	127,0	162,2
0,75	201,4	221,9	462,2

5. Conclusion

Dans le présent article, nous avons décrit un estimateur assisté par un modèle qui s'appuie sur la régression semi-paramétrique pour refléter la relation entre de multiples variables auxiliaires au niveau de la population et les variables de sondage. Nous avons élaboré une théorie asymptotique qui montre que l'estimateur résultant est convergent par rapport au plan et asymptotiquement normal sous des hypothèses faibles concernant le plan de sondage et la population. Cette théorie généralise les résultats de Breidt et Opsomer (2000), qui ont prouvé des résultats similaires pour un estimateur assisté par un modèle non paramétrique univarié. L'estimateur semi-paramétrique a été appliqué à des données provenant d'une enquête sur les lacs du Nord-Est des États-Unis, où il s'est avéré plus efficace qu'un estimateur ne tirant pas parti des variables auxiliaires et qu'un estimateur par la régression entièrement paramétrique.

En plus de ses propriétés théoriques, l'estimateur assisté par un modèle semi-paramétrique présente des propriétés pratiques séduisantes. Comme nous l'avons mentionné plus haut, il est entièrement calé pour les variables auxiliaires, qu'elles soient utilisées dans la composante de modélisation paramétrique ou non paramétrique, et il est invariant par rapport à la localisation et à l'échelle. L'estimateur peut être exprimé comme une somme pondérée des observations sur échantillon, de sorte qu'il est conforme aux paradigmes d'estimation par sondage classique et qu'un ensemble unique de poids peut être appliqué à toutes les variables étudiées, ce qui permet de préserver les relations entre les variables.

L'un des problèmes qui n'a pas été abordé dans le présent article est celui de la sélection du paramètre de

lissage pour la composante non paramétrique du modèle de régression. Il s'agit d'un sujet difficile dans le contexte de l'estimation assistée par un modèle, que complique encore davantage le fait qui vient d'être mentionné qu'un seul ensemble de poids de régression sur données d'enquête est appliqué à toutes les variables étudiées : parce que le choix de la largeur de fenêtre optimale dépend de la variable qui est lissée, aucune largeur de fenêtre (et par conséquent aucun ensemble de poids) unique ne sera optimale pour toutes les variables de l'enquête. Ce sujet est étudié à l'heure actuelle par les auteurs.

Remerciements

Les travaux de recherche sur lesquels s'appuie cet article ont été financés par les subventions DMS-0204531 et DMS-0204642 de la National Science Foundation et par les STAR Research Assistance Agreements CR-829095 et CR-829096 octroyés par l'Environmental Protection Agency (EPA) des États-Unis à l'Université de l'État du Colorado et à l'Université de l'État de l'Oregon. Le présent manuscrit n'a pas été révisé officiellement par l'EPA. Les opinions exprimées ici n'engagent que les auteurs. L'EPA ne sanctionne aucun produit ni service commercial mentionné dans le présent rapport.

Annexe

Hypothèses techniques et calculs

Nous commençons par énoncer les hypothèses nécessaires, qui étendent celles utilisées dans Breidt et Opsomer (2000) au modèle semi-paramétrique.

Hypothèses :

- **A1** *Distribution des erreurs sous* ξ : *les erreurs* ε_k *sont indépendantes et de moyenne nulle, de variance* $v(x_k, z_k)$, *et à support compact, uniformément pour tout* N.
- **A2** *Distribution des covariables* : *les* x_k *et* z_k *sont considérées fixes par rapport au modèle de superpopulation* ξ. *Il est supposé que les* z_k *ont un support borné et que les* x_k *sont indépendantes et de même loi* $F(x) = \int_{-\infty}^{x} f(t)\,dt$, *où* $f(\cdot)$ *est une densité avec support compact* $[a_x, b_x]$ *et* $f(x) > 0$ *pour tout* $x \in [a_x, b_x]$.
- **A3** *Fonctions de moyenne et de variance non paramétriques* : *la fonction de moyenne* $m(\cdot)$ *est continue et la fonction de variance* $v(\cdot, \cdot)$ *est bornée et strictement supérieure à* 0.
- **A4** *Noyau* K : *le noyau* $K(\cdot)$ *a un support compact* $[-1, 1]$, *est symétrique et continue, et satisfait* $\int_{-1}^{1} K(u)\,du = 1$.
- **A5** *Taux d'échantillonnage* nN^{-1} *et largeur de fenêtre* h_N : *quand* $N \to \infty$, $nN^{-1} \to \pi \in (0, 1)$, $h_N \to 0$ *et* $Nh_N^2/(\log\log N) \to \infty$.
- **A6** *Probabilités d'inclusion* π_k *et* π_{kl} : *pour tout* N, $\min_{k \in U_N} \pi_k \geq \lambda > 0$, $\min_{k,\, l \in U_N} \pi_{kl} \geq \lambda^* > 0$ *et*

$$\limsup_{N \to \infty} n \max_{k,\, l \in U_{N:\, i \neq j}} |\pi_{kl} - \pi_k \pi_l| < \infty.$$

- **A7** *Hypothèses supplémentaires faisant intervenir des probabilités d'inclusion d'ordre supérieur* :

$$\lim_{N \to \infty} n^2 \max_{(k_1, k_2, k_3, k_4) \in D_{4,\, N}} |E_p(I_{k_1} - \pi_{k_1})(I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})(I_{k_4} - \pi_{k_4})| < \infty,$$

où $D_{t,\, N}$ *dénote l'ensemble de tous les tuples* t *distincts* $(k_1, k_2, \ldots, k_t)$ *provenant de* U_N,

$$\lim_{N \to \infty} \max_{(k_1, k_2, k_3, k_4) \in D_{4,\, N}} |E_p(I_{k_1} I_{k_2} - \pi_{k_1 k_2})(I_{k_3} I_{k_4} - \pi_{k_3 k_4})| = 0,$$

et

$$\limsup_{N \to \infty} n \max_{(k_1, k_2, k_3) \in D_{3,\, N}} |E_p(I_{k_1} - \pi_{k_1})^2 (I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})| < \infty.$$

- **A8** *La matrice* $N^{-1} \boldsymbol{Z}_U^T (\boldsymbol{I} - \boldsymbol{S}_U^*) \boldsymbol{Z}_U$ *est inversible pour tout* N *avec une probabilité de modèle de* 1.

L'hypothèse A8 est requise pour que l'estimateur de population $\boldsymbol{B}$ soit bien défini. L'inversibilité de la matrice en A8 dépend de l'effet combiné de la largeur de fenêtre h et de la distribution conjointe des x_k et z_k. Bien qu'il soit possible, en principe, d'écrire suffisamment de conditions pour cela, nous avons opté pour cette approche plus simple et plus explicite.

Avant de donner les preuves des théorèmes 3.1 et 3.2, nous énonçons et prouvons un certain nombre de lemmes.

Lemme 1 *Sous les hypothèses* $A1$ à $A7$,

a) *pour tout* $k \in U$ *et* $d = 1, \ldots, D$,

$$\frac{1}{N} \sum_U E_p (\boldsymbol{s}_{Ak}^T \boldsymbol{Y}_A - \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U)^2 = O\left(\frac{1}{nh}\right)$$

et

$$\frac{1}{N} \sum_U E_p (\boldsymbol{s}_{Ak}^T \boldsymbol{Z}_{dA} - \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_{dU})^2 = O\left(\frac{1}{nh}\right);$$

b) *les* $\boldsymbol{s}_{UK}^T \boldsymbol{Y}_U$ *et* $\boldsymbol{s}_{UK}^T \boldsymbol{Z}_U$ *sont bornées uniformément sur tout* $k \in U$.

Preuve du lemme 1 : Puisque les y_k et z_{dk} sont les unes et les autres bornées par hypothèse, la partie (a) peut être démontrée en utilisant un raisonnement identique à celui du lemme 4 de Breidt et Opsomer (2000). Bien que ce lemme n'inclue pas de taux de convergence, ce taux est calculé facilement en notant que

$$\frac{1}{N} \sum_{i,\, k \in U_N} z_{ik}^2 = O\left(\frac{1}{nh}\right)$$

dans la notation de Breidt et Opsomer (2000), puis en poursuivant comme dans cette preuve.
La partie b) a été prouvée directement dans le lemme 2 (iv) de Breidt et Opsomer (2000).

Lemme 2 *Sous les hypothèses A*1 *à A*8,

$$\hat{\boldsymbol{B}} = \boldsymbol{B} + O_p(1/\sqrt{nh}),$$

avec le taux vérifié en ce qui concerne les composantes, et $\boldsymbol{B}$ *est borné pour tout* N.

Preuve du lemme 2 : Écrivons $\tilde{y}_k^{[s_U]} = \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U$ et $\tilde{y}_k^{[s_A]} = \boldsymbol{s}_{Ak} \boldsymbol{Y}_A$ pour les versions lissées en population et en échantillon de y_k, et, similairement, $\tilde{\boldsymbol{z}}_k^{[s_U]} = \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_U$ et $\tilde{\boldsymbol{z}}_k^{[s_A]} = \boldsymbol{s}_{Ak} \boldsymbol{Z}_A$. Nous réécrivons l'expression (6) sous forme d'une fonction des termes pondérés selon l'échantillon $\hat{\boldsymbol{t}}_l$, $l = 1, \ldots, 6$:

$$\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{t}_1 & \hat{\boldsymbol{t}}_2 \\ \hat{\boldsymbol{t}}_3 & \hat{\boldsymbol{t}}_4 \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_5 \\ \hat{\boldsymbol{t}}_6 \end{bmatrix},$$

où

$$\hat{t}_1 = \left(\frac{\hat{N}}{N}\right)^2$$

$$\hat{\boldsymbol{t}}_2 = \overline{\boldsymbol{z}}_\pi - \frac{1}{N} \sum_A \frac{\tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k} \left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_3 = \overline{\boldsymbol{z}}_\pi^T \left(\frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_4 = \frac{1}{N} \sum_A \frac{\boldsymbol{z}_k^T \boldsymbol{z}_k}{\pi_k} - \frac{1}{N} \sum_A \frac{\boldsymbol{z}_k^T \tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k} + \overline{\boldsymbol{z}}_\pi^T \frac{1}{N} \sum_A \frac{\tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k}$$

$$\hat{t}_5 = \overline{y}_\pi - \frac{1}{N} \sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k} \left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_6 = \frac{1}{N} \sum_A \frac{\boldsymbol{z}_k^T y_k}{\pi_k} - \frac{1}{N} \sum_A \frac{\boldsymbol{z}_k^T \tilde{y}_k^{[s_A]}}{\pi_k} + \overline{\boldsymbol{z}}_\pi^T \frac{1}{N} \sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k}.$$

L'estimateur pondéré selon l'échantillon $\hat{\boldsymbol{B}}$ sera étendu autour de

$$\boldsymbol{B} = \begin{bmatrix} 1 & \overline{\boldsymbol{z}}_N \\ \overline{\boldsymbol{z}}_N^T & \boldsymbol{t}_4 \end{bmatrix}^{-1} \begin{bmatrix} \overline{y}_N \\ \boldsymbol{t}_6 \end{bmatrix}, \qquad (15)$$

où

$$\boldsymbol{t}_4 = \frac{1}{N} \sum_U \boldsymbol{z}_k^T \boldsymbol{z}_k - \frac{1}{N} \sum_U \boldsymbol{z}_k^T \tilde{\boldsymbol{z}}_k^{[s_U]} + \overline{\boldsymbol{z}}_N^T \frac{1}{N} \sum_U \tilde{\boldsymbol{z}}_k^{[s_U]}$$

$$\boldsymbol{t}_6 = \frac{1}{N} \sum_U \boldsymbol{z}_k^T y_k - \frac{1}{N} \sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} + \overline{\boldsymbol{z}}_N^T \frac{1}{N} \sum_U \tilde{y}_k^{[s_U]}$$

et les $\boldsymbol{t}_l$ restants peuvent être trouvés dans (15). L'existence et la continuité des dérivées de $\hat{\boldsymbol{B}}$ par rapport à $\hat{\boldsymbol{t}}_l$ et évaluées au point $\boldsymbol{t}_l$ découle du lemme 1(b) et de l'existence de l'inverse en (15), qui est supposée par A8.

Le résultat découlera d'un développement en série de Taylor d'ordre 0 si nous pouvons montrer que $\hat{\boldsymbol{t}}_l - \boldsymbol{t}_l = O_p(1/\sqrt{nh})$ pour tout l (par exemple, Fuller (1996), Corollary 5.1.5). Pour $\hat{t}_1$ et $\hat{\boldsymbol{t}}_3$, cela découle directement de A2 et A6. Les termes restants contiennent des sommes comportant des quantités lissées $\tilde{\boldsymbol{z}}_k^{[s_A]}$ et $\tilde{y}_k^{[s_A]}$. Nous démontrons le raisonnement pour l'un de ces termes dans $\hat{\boldsymbol{t}}_6$. Nous avons

$$\frac{1}{N} \sum_A \frac{\boldsymbol{z}_k^T \tilde{y}_k^{[s_A]}}{\pi_k} - \frac{1}{N} \sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} = \frac{1}{N} \sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} \left(\frac{I_k}{\pi_k} - 1\right) + \frac{1}{N} \sum_U \boldsymbol{z}_k^T (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]}) \frac{I_k}{\pi_k},$$

et le premier terme est $O_p(1/\sqrt{n})$ en vertu de A6 et du lemme 1(b), en utilisant le même argument que dans le lemme 4 de Breidt et Opsomer (2000). Pour le deuxième terme, nous utilisons l'inégalité de Schwarz

$$\left| \frac{1}{N} \sum_U \boldsymbol{z}_k^T (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]}) \frac{I_k}{\pi_k} \right| \leq \sqrt{\frac{1}{N} \sum_U \boldsymbol{z}_k^{[2]T} \frac{I_k}{\pi_k^2}} \sqrt{\frac{1}{N} \sum_U (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})^2},$$

où $\boldsymbol{z}_k^{[2]}$ dénote que les carrés sont calculés pour les composantes. Le premier terme est borné par A2 et A6, et le deuxième terme est $O_p(1/\sqrt{nh})$ en vertu du lemme 1(a) et de l'inégalité de Markov. Le résultat souhaité s'ensuit alors en appliquant le même raisonnement aux termes restants dans $\hat{\boldsymbol{t}}_2$, $\hat{\boldsymbol{t}}_4$, $\hat{t}_5$, $\hat{\boldsymbol{t}}_6$.

La limitabilité de $\boldsymbol{B}$ découle directement de l'hypothèse A8, du lemme 1(b) et de la limitabilité des z_k.

Lemme 3 *Sous les hypothèses A*1 *à A*8, *nous avons*

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Preuve du lemme 3 : Étant donné l'expression (9), nous devons montrer que

$$(\overline{\boldsymbol{z}}_N - \overline{\boldsymbol{z}}_\pi)(\boldsymbol{B} - \hat{\boldsymbol{B}}) = o_p\left(\frac{1}{\sqrt{n}}\right) \tag{16}$$

$$\frac{1}{N}\sum_U (m_k - \hat{m}_k)\left(1 - \frac{I_k}{\pi_k}\right) = o_p\left(\frac{1}{\sqrt{n}}\right). \tag{17}$$

Le lemme 2 et les hypothèses A2, A5 et A6 montrent que $(\overline{\boldsymbol{z}}_N - \overline{\boldsymbol{z}}_\pi)(\boldsymbol{B} - \hat{\boldsymbol{B}}) = O_p(1/nh)$. Afin de prouver (17), nous pouvons réécrire cette équation sous la forme

$$\begin{aligned}\frac{1}{N}\sum_U (m_k - \hat{m}_k)\left(1 - \frac{I_k}{\pi_k}\right) &= \frac{1}{N}\sum_U (\tilde{y}_k^{[s_U]} - \tilde{y}_k^{[s_A]})\left(1 - \frac{I_k}{\pi_k}\right) \\ &\quad - \frac{1}{N}\sum_U (\tilde{\boldsymbol{z}}_k^{[s_U]} - \tilde{\boldsymbol{z}}_k^{[s_A]})\left(1 - \frac{I_k}{\pi_k}\right)\boldsymbol{B} \\ &\quad - \frac{1}{N}\sum_U \tilde{\boldsymbol{z}}_k^{[s_A]}\left(1 - \frac{I_k}{\pi_k}\right)(\boldsymbol{B} - \hat{\boldsymbol{B}}).\end{aligned}$$

Breidt et Opsomer (2000) ont prouvé dans le lemme 5 que le premier terme du deuxième nombre est $o_p(1/\sqrt{n})$; ce même lemme et la limitabilité de $\boldsymbol{B}$ fournissent le même taux pour le deuxième terme. Les hypothèses A5 et A6, le lemme 1(b) et le lemme 2 montrent que le troisième terme est $O_p(1/n\sqrt{h})$ et le taux souhaité est obtenu.

Lemme 4 *Sous les hypothèses A*6 et *A*8,

$$E_p(\hat{y}_{\text{dif}}) = \overline{y}_N$$

$$\begin{aligned}\text{Var}_p(\hat{y}_{\text{dif}}) &= \frac{1}{N^2}\sum\sum_{k,l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l} \\ &= O\left(\frac{1}{n}\right).\end{aligned}$$

Preuve du lemme 4 : Les propriétés de l'estimateur par la différence sont calculées facilement. Le taux de la variance due au plan de sondage découle des hypothèses énoncées en utilisant le même raisonnement que pour le lemme 4 de Breidt et Opsomer (2000).

Lemme 5 *Sous les hypothèses A*1 *à A*8,

$$\hat{V}(\hat{y}_{\text{reg}}) = \text{Var}_p(\hat{y}_{\text{dif}}) + o_p\left(\frac{1}{n}\right).$$

Preuve du lemme 5 : Le raisonnement de cette preuve suivra étroitement celui du théorème 3 de Breidt et Opsomer (2000). Nous écrivons

$$\begin{aligned}\hat{V}(\hat{y}_{\text{reg}}) - \text{Var}_p(\hat{y}_{\text{dif}}) &= (\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) \\ &\quad + (\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}}))\end{aligned} \tag{18}$$

avec

$$\hat{V}(\hat{y}_{\text{dif}}) = \frac{1}{N^2}\sum\sum_A \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l}.$$

Puisque

$$\frac{1}{N}\sum_U (y_k - g_k)^4 < \infty.$$

en vertu des hypothèses A1 à A3 et d'après les lemmes 1(b) et 2, l'approche utilisée pour le terme A_N de Breidt et Opsomer (2000) peut servir à montrer que

$$\text{E}_p\left|\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})\right| = o\left(\frac{1}{n}\right),$$

qui fournit la convergence souhaitée par l'inégalité de Markov.

Pour le premier terme de (18), notons que

$$\begin{aligned}\hat{g}_k - g_k &= (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]}) - (\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]})\,(\hat{\boldsymbol{B}} - \boldsymbol{B}) \\ &\quad + (\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_k^{[s_U]})(\hat{\boldsymbol{B}} - \boldsymbol{B}) - (\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]})\boldsymbol{B},\end{aligned}$$

de sorte que

$$\begin{aligned}&(\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) = \\ &\frac{1}{N^2}\sum\sum_U \left\{\begin{array}{l} -2\dfrac{y_k - g_k}{\pi_k}\dfrac{\hat{g}_l - g_l}{\pi_l} \\ \qquad + \dfrac{\hat{g}_k - g_k}{\pi_k}\dfrac{\hat{g}_l - g_l}{\pi_l}\end{array}\right\}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} I_k I_l\end{aligned}$$

peut être décomposé en termes de variance comportant des lisseurs d'échantillon et de population et des estimateurs paramétriques. Il est possible de démontrer que chacun de ces termes est $o_p(1/n)$. Nous démontrons cette approche pour l'un des termes :

$$\begin{aligned}&\left|\frac{1}{N^2}\sum\sum_U \frac{y_k - g_k}{\pi_k}\frac{\hat{\boldsymbol{z}}_l - \boldsymbol{z}_l}{\pi_l}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} I_k I_l(\hat{\boldsymbol{B}} - \boldsymbol{B})\right| \\ &\le \left(\frac{C_1}{N} + C_2 \max|\pi_{kl} - \pi_k\pi_l|\right)\frac{1}{N}\sum_U |\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]}|\,|\hat{\boldsymbol{B}} - \boldsymbol{B}| \\ &= o_p\left(\frac{1}{n}\right)\end{aligned}$$

où C_1, $C_2 < \infty$ résume les termes bornés (par les hypothèses A1 à A3 et A6 et le lemme 1(b)), et le taux de convergence et le résultat de l'hypothèse A6 et des lemmes 1(a) et 2.

Preuve du théorème 3.1 : Dans le lemme 3, nous montrons que

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

où $\hat{y}_{\text{dif}}$ est l'estimateur par la différence (3). Le résultat découle immédiatement de l'hypothèse A5 et du lemme 4.

Preuve du théorème 3.2 : Notons que $\hat{y}_{\text{dif}}$ peut s'écrire comme étant la somme d'une constante de population et d'un estimateur à facteur d'extension de la forme $\overline{y}_{\pi}$ en définissant une nouvelle variable $y_k - \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U + \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_U \boldsymbol{B} - \boldsymbol{z}_k \boldsymbol{B}$ pour $k \in U$. Comme dans le cas de la variable y_k, cette nouvelle variable a un support borné en vertu du lemme 1(b) et une variance d'ordre $O(1/n)$ en vertu du lemme 4. Donc, l'existence du TLC pour $\overline{y}_{\pi}$ implique l'existence du TLC pour $\hat{y}_{\text{dif}}$. En outre, $\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p(1/\sqrt{n})$ en vertu du lemme 3, de sorte que $\sqrt{n}\,\hat{y}_{\text{reg}}$ et $\sqrt{n}\,\hat{y}_{\text{dif}}$ suivent la même loi asymptotique. L'application du théorème de Slutsky et du lemme 5 complète la preuve.

Bibliographie

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

Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1996). *Introduction to Statistical Time Series* (2ième Éd.). New York : John Wiley & Sons, Inc.

Hastie, T.J., et Tibshirani, R.J. (1990). *Generalized Additive Models*. Washington, D.C.: Chapman and Hall.

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

Larsen, D.P., Thornton, K.W., Urquhart, N.S. et Paulsen, S.G. (1993). Overview of survey design and lake selection. EMAP - Surface Waters 1991 Pilot Report. (Éds. D.P. Larsen et S.J. Christie). Rapport Technique EPA/620/R - 93/003, U.S. Environmental Protection Agency.

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73, 166-179.

Opsomer, J.D., Breidt, F.J., Moisen, G.G. et Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*. À paraître.

Opsomer, J.-D., et Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186-211.

Opsomer, J.D., et Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics*, 8, 715-732.

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

Speckman, P.E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society*, Séries B, 50, 413-436.

Stoddard, J.L., Kahl, J.S., Deviney, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, P.S., Webb, J.R. et Webster, K.E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Rapport Technique EPA/620/R-03/001, U.S. Environmental Protection Agency, Washington, DC.

U.S. National Acid Precipitation Assessment Program (1991, Novembre). 1990 Integrated Assessment Report. Rapport Technique, Washington, DC.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 53-58
Statistique Canada, N° 12-001 au catalogue

Pondération *ex post* des données de prix pour l'estimation des taux de dépréciation

Marc Tanguay et Pierre Lavallée [1]

Résumé

Pour modéliser la dépréciation économique, on utilise une base de données qui contient des informations sur les actifs dont des entreprises se départissent. On connaît les prix d'acquisition et de revente ainsi que les durées d'utilisation de ces actifs. Cependant, les actifs dont on observe les prix sont uniquement ceux qui ont fait l'objet d'une transaction. Bien que la dépréciation d'un actif soit présente de façon continue au cours de sa vie, on ne connaît donc cette valeur que lorsqu'il y a eu transaction. La présente note propose une pondération ex post afin d'atténuer, au moins en partie, cet effet dans la détermination des modèles économétriques.

Mots clés : Ratio de prix; données de survie; distribution uniforme; dépréciation des voitures.

1. Contexte

Différents modèles économétriques sont utilisés pour estimer la dépréciation économique. On utilise, à cette fin, une base de données qui contient des informations sur les actifs dont des entreprises se départissent. On connaît les prix d'acquisition et de revente ainsi que les durées d'utilisation de ces actifs. On voudrait en inférer les résultats à la population totale des actifs utilisés par les entreprises. Sur l'utilisation de prix d'actifs usagés pour estimer la dépréciation économique, on peut notamment consulter Gellatly, Tanguay et Yan (2002), ainsi que Hulten et Wykoff (1981).

On s'interroge cependant sur la représentativité de la base de données utilisée. En effet, les actifs dont on observe les prix sont uniquement ceux qui ont fait l'objet d'une transaction. On ignore dans quelle mesure les pertes de valeur observées sur eux sont représentatives de la perte de valeur pour tous les actifs en production, qu'ils aient ou non fait l'objet d'une transaction. Ceci peut constituer une source d'erreur dans l'établissement des modèles économétriques parce que ces derniers cherchent à mesurer la dépréciation des actifs au cours de leur vie, qu'il y ait eu transaction ou non.

Nous nous proposons d'atténuer cette source d'erreur, en introduisant une pondération *ex post* dans la détermination des modèles économétriques. La section 2 de la présente note décrira plus en détail la problématique. À la section 3, nous exposerons la démarche suivie pour la détermination de la pondération. Finalement, à la section 4, nous présenterons quelques résultats numériques.

2. Problématique

On cherche à décrire la relation entre les prix et l'âge des actifs. On dispose d'un échantillon de n actifs où on connaît, pour chaque actif i, le ratio de prix r_i et le temps t_i où ce ratio a été mesuré. Une fois que les prix sont exprimés en dollars réels, ce ratio est donné par $r_i = P_i^t / P_i^0$ où P_i^0 est la valeur initiale de l'investissement de l'actif i et P_i^t est son prix de revente au temps t. Ce ratio est strictement décroissant par rapport à l'axe du temps t. Au point de départ, on ignore le processus qui génère la perte de valeur et on n'a aucune spécification concernant la fonction qui décrit cette perte sinon qu'elle est strictement décroissante. Il est cependant possible d'examiner la distribution des ratios des prix entre 0 et 1. Voici un exemple construit à partir des données sur les usines de fabrication (on doit noter que les 2/3 de l'échantillon ont été exclus car il correspondent à des mises au rancard (le prix est nul) et les procédures d'estimation prennent en compte, chacune à sa façon, cette composante).

Étant donné que l'on veut utiliser les données pour inférer des statistiques sur la population des actifs en production, on souhaiterait que nos données aient des propriétés analogues à celles d'un échantillon aléatoire qui serait tiré sur cette population. Ceci n'est pas le cas, rappelons-le, parce qu'on ne dispose que des prix des actifs i qui ont fait l'objet d'une transaction au temps t_i, $i = 1, ..., n$. En effet, bien que l'on voudrait disposer de ratios des prix pour différentes périodes de l'existence d'un actif i donné, ce ratio n'est disponible que lorsqu'il y a eu transaction, ce qui survient de façon non uniforme durant la durée de vie d'un actif.

On peut donc se demander quelle forme aurait la distribution ci-dessus si elle avait été tirée d'un échantillon où le ratio des prix avait été mesuré, pour un même actif i, à différents temps t. Notre argument est qu'elle devrait converger vers une *distribution uniforme*. Nous allons donc chercher à obtenir une pondération qui nous aidera à recréer une distribution uniforme des ratios de prix. Cette pondération nous aidera à pallier le manque d'uniformité dans la distribution des observations, ce qui peut influencer les analyses statistiques comme, par exemple, la régression linéaire.

1. Marc Tanguay et Pierre Lavallée, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Courriel : marc.tanguay@statcan.ca, pierre.lavallee@statcan.ca.

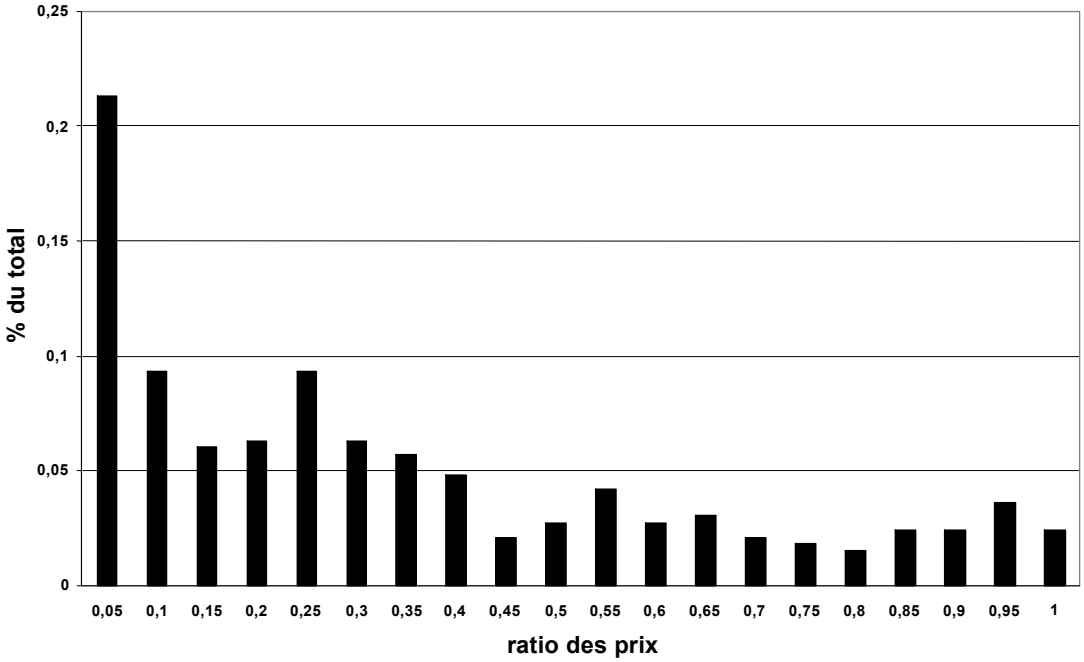


Figure 1 Distribution des observations selon le ratio des prix, usines de fabrication

3. Démarche

Notre point de départ est que les ratios de prix peuvent être considérés comme des réalisations empiriques d'une fonction de survie de forme inconnue. Dans les modèles de durée de vie, la fonction de survie exprime la probabilité qu'une entité dont la vie est limitée survive au-delà d'un certain point sur l'axe du temps. Elle fournit, par conséquent, la même information que la fonction de répartition (ou *Cumulative Distribution Function*). Soit r, une variable aléatoire qui décrit la durée de vie d'une unité de valeur incorporée dans un actif quelconque. La valeur s'épuise au fur et à mesure que le temps passe, et ce, aussi longtemps que l'actif est en service. Le ratio des prix peut donc s'interpréter comme la fraction survivante qui diminue peu à peu. On note cette fraction $S(y)$ et on a

$$S(y) = 1 - F(y)$$

où $F(y) = P(r \le y)$ est la fonction de répartition, c'est-à-dire la probabilité qu'une unité de valeur soit perdue avant le point y.

Les théorèmes des transformations fondamentales des lois de probabilité permettent de décrire la fonction inverse de $F(y)$ (Greene 1993 et Ross 2002). Soit $z = F(y)$ et supposons que la fonction inverse F^{-1} existe de sorte que $y = F^{-1}(z)$. Il y a donc une concordance directe entre l'espace de y, borné à 0 mais infini à droite, et celui de F qui est borné entre 0 et 1. La fonction de répartition de z est $F(F^{-1}(z)) = z$. La loi qui génère une telle répartition est une distribution uniforme entre 0 et 1.

Ce résultat est généralement au cœur des processus de génération de données comme les simulations Monte-Carlo puisque lors de la génération d'un échantillon aléatoire, on utilise souvent une distribution uniforme à laquelle on applique ensuite la fonction inverse (Davidson et MacKinnon 1993) . Cette approche n'est toutefois par toujours pratique ou carrément impossible, en particulier si la fonction inverse F^{-1} n'a pas de forme explicite. Ce résultat a aussi été utilisé dans les approches de résidus généralisés, notamment pour la construction de tests de spécification (Lancaster 1985).

Il en résulte que n'importe quel échantillon aléatoire construit à partir de réalisations empiriques de données de proportion de survie doit converger en distribution vers une distribution uniforme.

Dans le cas des données de prix, l'intuition est la suivante : entre l'investissement et la mise au rancard, toute la fourchette des prix relatifs doit forcément être couverte par un actif en production. À la période initiale, la valeur est perdue plus rapidement, il y a donc une plus grande quantité d'observations dont les durées sont courtes. Mais cela est compensé par le fait que la référence correspondante sur l'échelle du temps est également plus courte. Par exemple, il faut moins de temps pour passer de 100 % de la valeur initiale à 90 %, que de 15 % à 5 % de la valeur initiale.

Il est facile de vérifier ces résultats de façon numérique, à l'aide de données simulées et nous ne nous y attarderons pas. Nous allons plutôt examiner comment ce résultat peut être réintroduit dans la base de données pour lui restaurer, au moins en partie, des propriétés semblables à celle d'un échantillon aléatoire. *Pour ce faire*, *il nous suffit d'imposer ex post*, *à la distribution empirique des prix*, *une structure de poids* w_i *qui soit telle que la distribution empirique des données*, *dans l'espace des prix*, *soit uniforme.*

La distribution empirique des ratios de prix r est donnée par

$$\hat{F}_n(y) = \frac{\sum_{i=1}^{n} I_i(y)}{n} \tag{1}$$

où $I_i(y) = 1$ si la valeur mesurée r_i de l'actif *i* est inférieure ou égal à y (c'est-à-dire, $r_i \le y$), et 0 sinon, et n est le nombre total d'observations. Notons que si les n unités de l'échantillon sont indépendantes et identiquement distribuées (i.i.d.), lorsque $n \to \infty$, $\hat{F}_n(y)$ converge en probabilité vers $F(y)$, c'est-à-dire $\hat{F}_n(y) \xrightarrow{P} F(y)$ (Bickel et Doksum 1977).

Pour obtenir le poids w_i pour chaque actif *i*, on distribue simplement l'échantillon en un nombre H donné d'intervalles (ou classes) de largeur fixe sur l'échelle des ratios de prix, et on attribue la même probabilité $\pi = 1/H$ à chacun de ces intervalles. Puisque les ratios de prix sont contenus entre 0 et 1, on a alors l'intervalle $h = 1$ donné par $[0,\ H^{-1}]$, et pour $h = 2,\ ...,\ H$, les intervalles sont donnés par $](h-1)H^{-1},\ hH^{-1}]$. Un poids w_h est ensuite calculé dans chaque intervalle h par le ratio $\pi/\hat{\pi}_h$ où $\hat{\pi}_h$ est la probabilité empirique spécifique à l'intervalle *h*, c'est-à-dire

$$\hat{\pi}_h = \frac{1}{n}\sum_{i=1}^{n} \delta_i(h) = \frac{n_h}{n} \tag{2}$$

où $\delta_i(h) = 1$ si $r_i \in h$, 0 sinon. On pose alors

$$w_i = w_h = \frac{\pi}{\hat{\pi}_h} = \frac{n}{Hn_h} \tag{3}$$

pour $r_i \in h$. En utilisant ces poids, la *distribution pondérée empirique* des ratios de prix *r* est donnée par

$$\hat{F}_{n,w}(y) = \frac{\sum_{i=1}^{n} w_i\, I_i(y)}{\sum_{i=1}^{n} w_i}. \tag{4}$$

En notant que $\sum_{i=1}^{n} w_i = \sum_{h=1}^{H}\sum_{i=1}^{n_h} n/Hn_h = n$, on obtient finalement

$$\hat{F}_{n,w}(y) = \frac{\sum_{i=1}^{n} w_i\, I_i(y)}{n}. \tag{5}$$

Puisque $n_h = \sum_{i=1}^{n}\delta_i(h)$, on a

$$\begin{aligned}\hat{F}_{n,w}(y) &= \frac{\sum_{i=1}^{n} w_i\, I_i(y)}{n} \\ &= \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n}\delta_i(h)\, I_i(y) \\ &= \frac{1}{H}\sum_{h=1}^{H}\frac{\sum_{i=1}^{n}\delta_i(h)\, I_i(y)}{\sum_{i=1}^{n}\delta_i(h)} \\ &= \frac{1}{H}\sum_{h=1}^{H}\hat{F}_n(y \mid h).\end{aligned} \tag{6}$$

Lorsque $n \to \infty$, on a $(1/n)\sum_{i=1}^{n}\delta_i(h)\, I_i(y) \xrightarrow{P} P(r \in h,\ r \le y)$ et $(1/n)\sum_{i=1}^{n}\delta_i(h) \xrightarrow{P} P(r \in h)$. Donc, lorsque $n \to \infty$,

$$\begin{aligned}\hat{F}_n(y \mid h) &\xrightarrow{P} \frac{P(r \in h, r \le y)}{P(r \in h)} \\ &= P(r \le y \mid r \in h) = F(y \mid h)\end{aligned} \tag{7}$$

où $F(y \mid h)$ est la distribution des ratios de prix r à l'intérieur de l'intervalle *h*.

Pour *n* suffisamment grand, on devrait être en mesure de déterminer H de manière à construire les intervalles *h* pour que $\hat{F}_n(y \mid h)$ soit approximativement uniformément distribuée, $h = 1,\ ...,\ H$. En d'autres mots, lorsque $n \to \infty$, pour H suffisamment grand, $F(y \mid h)$ devrait suivre une distribution uniforme sur l'intervalle *h*. Notons qu'un tel argument a été utilisé par Dalenius et Hodges (1959) dans un contexte de stratification optimale. Dans ce cas, la distribution $F(y \mid h)$ est donnée par

$$F(y \mid h) = \begin{cases} 0 \text{ pour } y \le (h-1)H^{-1} \\ Hy - h + 1 \text{ pour } (h-1)H^{-1} < y \le hH^{-1} \\ 1 \text{ pour } y > hH^{-1}. \end{cases} \tag{8}$$

Puisque $F(y) = \sum_{h=1}^{H} F(y \mid h)/H$, on a $F(y) = y$, ce qui correspond à la distribution uniforme. On conclue donc que pour *n* suffisamment grand, l'utilisation de la pondération (3) devrait rendre la distribution empirique pondérée $\hat{F}_{n,w}(y)$ donnée par (5) approximativement uniformément distribuée.

Des simulations Monte-Carlo ont démontré que les estimations résultant d'un échantillon non aléatoire pouvaient être améliorées en utilisant cette approche. Ses principaux avantages résident dans :

- sa simplicité;
- le fait qu'elle peut être introduite *ex ante*, c'est-à-dire avant l'introduction du modèle économétrique comme tel. Par conséquent, elle ne requiert pas d'hypothèses de travail fortes.

Si on reprend l'histogramme exposé plus haut et divisons l'échantillon en $H = 5$ intervalles d'une largeur de 0,2 avec une valeur de $\pi = 1/5 = 0,2$, on obtient alors l'histogramme suivant qui a été pondéré *ex post*.

4. Application

Nous allons illustrer la démarche à partir d'un exemple tiré du Kelly Blue Book, une source d'information largement utilisée pour l'estimation de la dépréciation des voitures. Le tableau 1 présente les prix de deux modèles de voitures pour différents âges entre 1 et 18 ans. Pour chaque voiture, on dispose donc d'un échantillon de $n = 18$ unités. Les prix sont exprimés en valeur relative par rapport à un modèle neuf. Il est en outre nécessaire d'ajuster les ratios pour tenir compte de la probabilité de survie à chacun de ces âges. Pour chaque voiture, le ratio final utilisé r_i de l'année *i* est donc construit à partir du produit du ratio des prix par la probabilité de survie.

On s'intéresse au taux de dépréciation moyen $\bar{\tau}$ pour chaque voiture. Ce dernier pourrait être estimé à partir d'une régression des prix (ou d'une fonction de ces derniers) par rapport à l'âge (ou d'une fonction de l'âge). Toutefois, si on présume que le taux est constant et de forme géométrique, on a la relation $r_i = 1 - \bar{\tau}^i$, où r_i est le prix relatif selon l'âge *i*. Dans ce cas, un taux $\hat{\tau}_i$ peut être estimé à chaque âge *i* par $\hat{\tau}_i = 1 - r_i^{1/i}$. Une estimation du taux de dépréciation moyen est alors produite à partir de la moyenne pour tous les âges, c'est-à-dire $\hat{\bar{\tau}} = \sum_{i=1}^{18} \hat{\tau}_i / 18$.

Dans l'exemple ci-dessus, on constate que les taux de dépréciation $\hat{\tau}_i$ varient selon la fourchette d'âge et qu'ils ont tendance à augmenter avec l'âge. Par ailleurs, le fait que l'on utilise une simple moyenne des âges dans le calcul de $\hat{\bar{\tau}}$ revient à accorder de façon implicite le même poids à chacun des âges. Mais il est bien évident que ce ne serait pas la distribution que l'on obtiendrait si on tirait un échantillon aléatoire des voitures en services. La figure ci-dessous présente la distribution des cellules de prix entre les ratios de 0 et 1.

La technique de repondération consiste simplement à imposer un poids égal à chacune des fourchettes de prix relatifs. Dans cet exemple, les $n = 18$ âges sont répartis en $H = 7$ classes, ce qui répartit les âges en 18/7 à chacune d'entre elles (en réalité, la structure des cellules a été configurée pour 8 classes mais la dernière est toujours vide). Comme mentionné à la section 3, les poids individuels w_i de chaque âge *i* sont construits selon (3), c'est-à-dire en divisant 18/7 par le nombre d'observations qui se trouvent dans chaque classe, sauf pour les cellules vides dont le poids demeure nul. Le tableau 2 présente les résultats et l'impact de la repondération sur les statistiques dérivées.

Cet exemple illustre bien les problèmes de biais d'agrégation typique des régressions estimées à partir d'agrégats économiques, sans tenir compte de la distribution réelle des unités au niveau micro. Ainsi, il est assez évident que les unités de 17 et 18 ans ne sauraient avoir le même poids de régression que celles de 1 an puisque le risque de perte à 1 an concerne pratiquement toutes les voitures qui seront mises en circulation alors que très peu d'entre elles seront exposées au risque de perte de valeur à des âges avancés. Il en résulte que l'estimation non pondérée, dans cet exemple, introduit une surestimation du taux de dépréciation de l'ordre de 15 %.

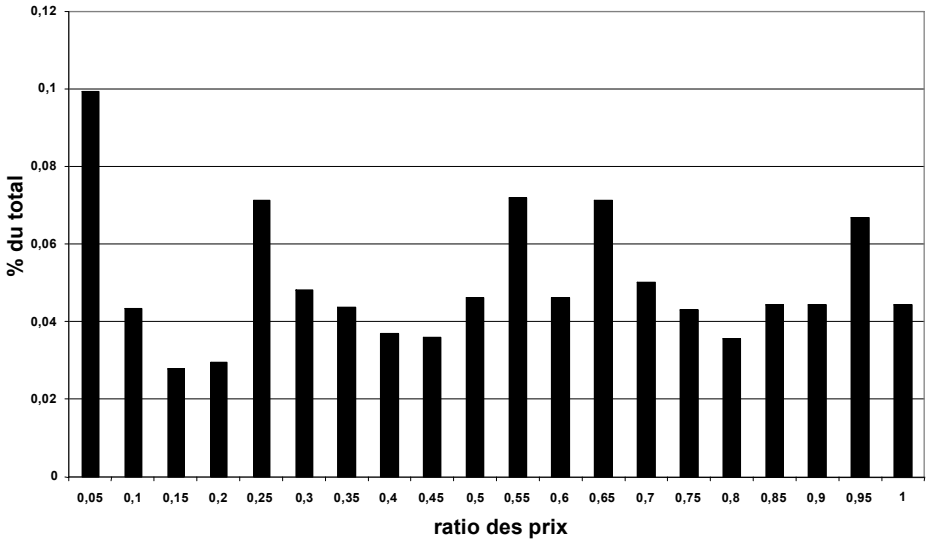


Figure 2 Distribution pondérée des observations selon le ratio des prix, usines de fabrication pondération *ex post*

Tableau 1 Prix relatifs de deux modèles de voitures selon le Kelly Blue Book et Taux de dépréciation moyen avant repondération

	Pr ($t > S$)*	Prix relatifs				Taux moyens de dépréciation	
		Excluant mises au rancard		Incluant mises au rancard		Incluant mises au rancard	
Année		*Buick*	*Chrysler*	*Buick*	*Chrysler*	*Buick*	*Chrysler*
1	0,9988	0,8633	0,8257	0,8622	0,8246	0,1367	0,1743
2	0,9901	0,7435	0,6801	0,7361	0,6734	0,1377	0,1753
3	0,9666	0,6410	0,5608	0,6195	0,5420	0,1378	0,1754
4	0,9220	0,5523	0,4621	0,5092	0,4261	0,1379	0,1755
5	0,8526	0,4740	0,3794	0,4042	0,3234	0,1387	0,1762
6	0,7582	0,4034	0,3087	0,3058	0,2341	0,1404	0,1779
7	0,6433	0,3391	0,2482	0,2181	0,1597	0,1432	0,1805
8	0,5164	0,2790	0,1953	0,1441	0,1009	0,1475	0,1846
9	0,3892	0,2227	0,1491	0,0867	0,0580	0,1537	0,1906
10	0,2731	0,1639	0,1050	0,0448	0,0287	0,1654	0,2018
11	0,1770	0,1261	0,0772	0,0223	0,0137	0,1716	0,2077
12	0,1051	0,0892	0,0523	0,0094	0,0055	0,1824	0,2180
13	0,0567	0,0614	0,0344	0,0035	0,0019	0,1932	0,2284
14	0,0276	0,0441	0,0236	0,0012	0,0007	0,1999	0,2347
15	0,0120	0,0320	0,0164	0,0004	0,0002	0,2050	0,2396
16	0,0046	0,0190	0,0093	0,0001	0,0000	0,2194	0,2534
17	0,0016	0,0088	0,0041	0,0000	0,0000	0,2432	0,2761
18	0,0005	0,0051	0,0023	0,0000	0,0000	0,2542	0,2867
					Moyenne	0,1727	0,2087

* Probabilité de Survie selon les estimations de la Division des études et de l'analyse micro-économique de Statistique Canada.

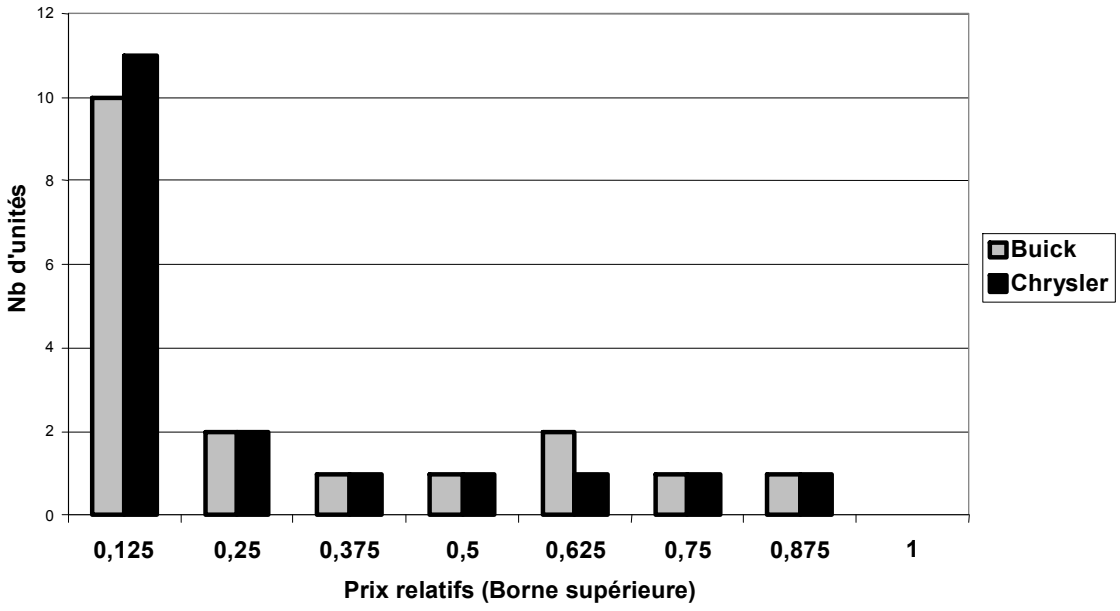


Figure 3 Distribution des cellules utilisées pour l'estimation du taux de dépréciation moyen selon les données du Kelly Blue Book avant repondération (Total = 18)

Tableau 2 Prix relatifs de deux modèles de voitures selon le Kelly Blue Book et taux de dépréciation moyen après repondération

	Prix relatifs		Taux moyens de dépréciation		Poids Ex post	
	Incluant mises au rancard		Incluant mises au rancard			
Année	*Buick*	*Chrysler*	*Buick*	*Chrysler*	*Buick*	*Chrysler*
1	0,8622	0,8246	0,1367	0,1743	2,5714	2,5714
2	0,7361	0,6734	0,1377	0,1753	2,5714	2,5714
3	0,6195	0,5420	0,1378	0,1754	1,2857	2,5714
4	0,5092	0,4261	0,1379	0,1755	1,2857	2,5714
5	0,4042	0,3234	0,1387	0,1762	2,5714	2,5714
6	0,3058	0,2341	0,1404	0,1779	2,5714	1,2857
7	0,2181	0,1597	0,1432	0,1805	1,2857	1,2857
8	0,1441	0,1009	0,1475	0,1846	1,2857	0,2338
9	0,0867	0,0580	0,1537	0,1906	0,2571	0,2338
10	0,0448	0,0287	0,1654	0,2018	0,2571	0,2338
11	0,0223	0,0137	0,1716	0,2077	0,2571	0,2338
12	0,0094	0,0055	0,1824	0,2180	0,2571	0,2338
13	0,0035	0,0019	0,1932	0,2284	0,2571	0,2338
14	0,0012	0,0007	0,1999	0,2347	0,2571	0,2338
15	0,0004	0,0002	0,2050	0,2396	0,2571	0,2338
16	0,0001	0,0000	0,2194	0,2534	0,2571	0,2338
17	0,0000	0,0000	0,2432	0,2761	0,2571	0,2338
18	0,0000	0,0000	0,2542	0,2867	0,2571	0,2338
				Moyenne pondérée	0,1479	0,1836

Remerciements

Les auteurs tiennent à remercier grandement l'arbitre anonyme de *Techniques d'enquête* qui, par ses judicieux commentaires, a contribué à améliorer la qualité de l'article.

Bibliographie

Bickel, P.J., et Doksum, K.A. (1977). *Mathematical Statistics*, Holden-Day, Oakland, CA.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Davidson, R., et MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, N.Y.

Gellatly, G., Tanguay, M. et Yan, B. (2002). An alternative methodology for estimating economic depreciation: New results using a survival model. Dans *Productivity Growth in Canada-2002*, Statistique Canada. #15-204-XPE.

Greene, W.H. (1993). *Econometric Analysis*. Deuxième édition, Prentice Hall, Englewood Cliffs, N.J.

Hulten, C.R., et Wykoff, F.C. (1981). The measurement of economic depreciation. Dans *Depreciation*, *Inflation*, *and the Taxation of Income from Capital*, (Éd. C.R. Hulten). The Urban Institute Press, Washington, D.C, 81-125.

Lancaster, T. (1985). Generalized residuals and heterogeneous duration model: With applications to the weibull model. *Journal of Econometrics*, 28, 155-69.

Ross, S.M. (2002). *Introduction to Probability Models*, 8ième Édition, Academic Press.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 59-69
Statistique Canada, N° 12-001 au catalogue

Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes-ménages

David G. Steel et Robert G. Clark [1]

Résumé

Une classe courante de plans de sondage comprend la sélection de toutes les personnes dans les ménages échantillonnés. Des estimateurs par la régression généralisée peuvent être calculés au niveau de la personne ou du ménage. L'utilisation de l'estimateur au niveau du ménage est commode parce que la même pondération d'estimation est appliquée à tous les membres du ménage. Dans le présent article, nous comparons théoriquement et empiriquement les deux approches dans le cas de l'échantillonnage aléatoire simple de ménages et la sélection de toutes les personnes présentes dans chaque ménage échantillonné. Nous constatons que l'approche au niveau du ménage est théoriquement plus efficace dans le cas de grands échantillons et que toute inefficacité empirique dans les petits échantillons est limitée.

Mots clés : Effets contextuels; estimateur par la régression généralisée; corrélation intraclasse; variance d'échantillonnage; assisté par modèle; enquêtes-ménages.

1. Introduction

De nombreuses enquêtes-ménages comprennent la sélection d'un échantillon de ménages, suivie de la sélection de toutes les personnes faisant partie du champ d'observation de l'enquête dans les ménages échantillonnés. Des données sur une ou plusieurs variables d'intérêt sont recueillies pour les personnes incluses dans l'échantillon. Il existe parfois des variables auxiliaires pour lesquelles les totaux de population et les valeurs d'échantillon sont connues; par exemple, il pourrait s'agir de chiffres de population selon les caractéristiques géographiques et démographiques. L'estimateur par la régression généralisée (GREG) est souvent utilisé pour combiner l'information auxiliaire et les données d'échantillon en vue d'estimer efficacement les totaux de population de la variable d'intérêt.

L'estimateur GREG s'appuie sur un modèle de régression reliant la variable d'intérêt aux variables auxiliaires. L'approche ordinaire consiste à ajuster ce modèle en utilisant les données recueillies pour chaque personne faisant partie de l'échantillon (par exemple Lemaître et Dufour 1987, premier paragraphe). Cet estimateur GREG au niveau de la personne est égal à une somme pondérée des valeurs d'échantillon de la variable d'intérêt, où la pondération est en général différente pour chaque personne.

Il est parfois commode d'utiliser des pondérations égales pour tous les membres d'un ménage dans le cas d'enquêtes qui recueillent des renseignements sur des variables d'intérêt au niveau du ménage ainsi qu'au niveau de la personne. Les mêmes pondérations peuvent alors être utilisées pour les deux types de variables afin d'être certain que les relations entre les variables du ménage et les variables personnelles soient reflétées dans les estimations du total. Si une variable au niveau du ménage est égale à la somme des variables au niveau de la personne (par exemple, si le revenu du ménage est égal à la somme des revenus personnels), alors le total estimé de la variable au niveau du ménage sera égal au total estimé de la variable au niveau de la personne. Cela n'est généralement pas le cas si des méthodes de pondération distinctes sont utilisées pour les variables au niveau de la personne et au niveau du ménage. De même, s'il existe une inégalité entre la variable au niveau du ménage et la somme des variables au niveau de la personne, celle-ci sera reflétée dans les estimations des deux variables. Par exemple, le nombre estimé de ménages utilisant une garderie ne devrait pas être supérieur au nombre estimé d'enfants allant en garderie.

L'estimateur GREG au niveau du ménage produit des pondérations égales au sein des ménages en ajustant le modèle de régression d'après les totaux au niveau du ménage de la variable d'intérêt et des variables auxiliaires (par exemple, Nieuwenbroek 1993). Les pondérations ayant cette propriété sont appelées pondérations intégrées.

Une autre approche consisterait à utiliser des méthodes d'estimation différentes pour les variables au niveau du ménage et celles au niveau de la personne, puis à faire une correction pour forcer les estimations qui devraient être égales à concorder. Cette approche, parfois appelée étalonnage, a surtout été utilisée pour obtenir la cohérence entre les estimations calculées d'après des enquêtes-entreprises annuelles et infra-annuelles (par exemple, Cholette 1984). L'application de l'approche d'étalonnage aux variables au niveau du ménage et au niveau de la personne des enquêtes-ménages nécessiterait l'identification explicite des variables aux niveaux de la personne et du ménage pour lesquelles les totaux de population devraient être égaux. Dans le présent

1. David G. Steel et Robert G. Clark, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australie. Courriel : David_Steel@uow.edu.au.

article, nous nous concentrons sur la pondération intégrée et n'envisageons pas les approches d'étalonnage.

Luery (1986), Alexander (1987), Heldal (1992), ainsi que Lemaître et Dufour (1987) ont discuté d'un certain nombre de méthodes qui produisent des pondérations intégrées pour les estimations au niveau de la personne et au niveau du ménage. Toutefois, aucun de ces auteurs n'a évalué l'effet sur la variance d'échantillonnage du calcul de l'estimateur par la régression généralisée au niveau du ménage plutôt qu'au niveau de la personne. Cette question est importante en pratique, car il convient de trouver le juste équilibre entre l'avantage cosmétique des pondérations intégrées et tout effet sur l'efficacité d'échantillonnage.

Le présent article donne une comparaison de la variance sous le plan, qui est la variance calculée par échantillonnage probabiliste répété à partir d'une population fixe, des estimateurs par la régression généralisée au niveau de la personne et au niveau du ménage. À la section 2, nous prouvons que la variance en grand échantillon de l'estimateur au niveau du ménage est inférieure ou égale à celle de l'estimateur au niveau de la personne, en montrant que le premier est optimal dans une grande classe d'estimateur GREG. Nous montrons qu'il en est ainsi parce que l'estimateur au niveau du ménage modélise efficacement les effets contextuels, tandis que l'estimateur au niveau de la personne ne le fait pas. À la section 3, nous recourrons à la simulation pour comparer les deux estimateurs en nous appuyant sur une gamme de variables. À la section 4, nous discutons des résultats. Les preuves des trois théorèmes sont présentées en annexe.

2. Comparaison théorique des estimateurs GREG aux niveaux de la personne et du ménage

2.1 L'estimateur par la régression généralisée

À la présente sous-section, nous décrivons l'estimateur par la régression généralisée pour le cas général de l'échantillonnage probabiliste à partir de toute population d'unités. Soit U une population finie d'unités et $s \subseteq U$, l'échantillon. Les probabilités de sélection sont $\pi_i = \Pr[i \in s]$ pour les unités $i \in U$. Soit y_i la variable d'intérêt qui est observée pour les unités $i \in s$. Soit $\boldsymbol{z_i}$ le vecteur de variables auxiliaires pour l'unité i, qui sont observées pour chaque unité de la population. Les totaux de population de ces variables sont T_Y et $\boldsymbol{T_Z}$, respectivement.

L'estimateur par la régression généralisée de T_Y est basé sur un modèle reliant la variable d'intérêt aux variables auxiliaires :

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\beta}^T \boldsymbol{z_i} \\ &\mathrm{var}_M[y_i] = v_i \sigma^2 \\ &y_i,\ y_j \text{ indépendantes pour } i \neq j \end{aligned}\right\} \qquad (1)$$

où v_i représente les paramètres de variance connus. L'indice « M » désigne les espérances sous un modèle et l'indice « p » désigne les espérances sous le plan, qui sont les espérances calculées par échantillonnage probabiliste répété à partir d'une population fixe. Dans le cas des enquêtes-entreprises, qui recueillent des données sur des variables continues telles que les revenus et les dépenses de l'entreprise, v_i est souvent modélisée sous forme d'une fonction de la taille de l'entreprise. Dans le cas des enquêtes-ménages, la variable d'intérêt est fréquemment dichotomique, auquel cas v_i est habituellement fixée à 1, ce qui correspond à un modèle homoscédastique.

Habituellement, $\boldsymbol{z_i}$ a la propriété qu'il existe un vecteur $\boldsymbol{\lambda}$ tel que $\boldsymbol{\lambda}^T \boldsymbol{z_i} = 1$ pour tout $i \in U$. Par exemple, cela est vrai si le modèle de régression (1) contient un paramètre d'ordonnée à l'origine.

Définition 1. Estimateur par la régression généralisée

L'estimateur par la régression généralisée pour le modèle (1) est défini comme étant

$$\hat{T}_r = \hat{T}_\pi + \hat{\boldsymbol{\beta}}^T (\boldsymbol{T_Z} - \hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}}) \qquad (2)$$

où

$$\hat{T}_\pi = \sum_{i \in s} \pi_i^{-1} y_i$$

$$\hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}} = \sum_{i \in s} \pi_i^{-1} \boldsymbol{z_i}.$$

et $\hat{\boldsymbol{\beta}}$ est une solution de

$$\sum_{i \in s} c_i \pi_i^{-1} \left(y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{z_i} \right) \boldsymbol{z_i} = \boldsymbol{0}$$

où c_i représente les poids de régression. (Souvent, c_i est fixé à $c_i = v_i^{-1}$.)

Les coefficients $\hat{\boldsymbol{\beta}}$ sont calculés d'après une régression par les moindres carrés pondérés de y_i sur $\boldsymbol{z_i}$ pour $i \in s$. L'estimateur GREG possède une variance sous le plan faible si le modèle est approximativement vrai, mais est convergent sous le plan indépendamment de la vérité du modèle (par exemple Särndal et coll. 1992, chapitre 6).

Pour de grands échantillons, la variance sous le plan de $\hat{T}_r$ est approximativement égale à

$$\mathrm{var}_p[\hat{T}_r] \approx \mathrm{var}_p[\tilde{T}_r] \qquad (3)$$

où

$$\tilde{T}_r = \hat{T}_\pi + \boldsymbol{B}^T(\boldsymbol{T}_Z - \hat{\boldsymbol{T}}_{Z\pi})$$

et $\boldsymbol{B}$ est une solution de

$$\sum_{i\in U} c_i\,(y_i - \boldsymbol{B}^T\boldsymbol{z}_i)\,\boldsymbol{z}_i = \boldsymbol{0}$$

(Särndal et coll., 1992, Résultat 6.6.1, page 235). Les coefficients $\boldsymbol{B}$ sont calculés d'après une régression par les moindres carrés pondérés de y_i sur $\boldsymbol{z}_i$ pour $i \in U$. Les coefficients de régression d'échantillon $\hat{\boldsymbol{\beta}}$ sont convergents sous le plan pour $\boldsymbol{B}$.

2.2 Estimateurs GREG aux niveaux de la personne et du ménage

Considérons maintenant le cas particulier de l'échantillonnage de ménages, où l'unité de base, i, est la personne. Soit $\boldsymbol{x}_i$ le vecteur de dimension p de variables auxiliaires observées pour toutes les personnes $i \in U$. Les éléments de $\boldsymbol{x}_i$ peuvent renvoyer à des caractéristiques de la personne ou du ménage auquel ces personnes appartiennent. La population et l'échantillon de ménages seront dénotés U_1 et s_1, respectivement. La population de personnes dans le ménage $g \in U_1$ sera dénotée U_g qui est de taille N_g. Soit $y_{g1} = \sum_{i\in U_g} y_i$ et $\boldsymbol{x}_{g1} = \sum_{i\in U_g} \boldsymbol{x}_i$ les totaux au niveau du ménage de y_i et $\boldsymbol{x}_i$. Soit $\bar{\boldsymbol{x}}_g = \boldsymbol{x}_{g1}/N_g$ la moyenne au niveau du ménage de $\boldsymbol{x}_i$.

Considérons le cas courant où les ménages sont sélectionnés par échantillonnage probabiliste et où toutes les personnes faisant partie des ménages échantillonnés sont sélectionnées, de sorte que $s = \bigcup_{g\in s_1} U_g$. Soit $\pi_{g1} = P[g \in s_1] > 0$ la probabilité de sélection pour le ménage g. Il s'ensuit que $\pi_i = \pi_{g1}$ pour $i \in U_g$.

L'estimateur GREG au niveau de la personne, $\hat{T}_P$, est l'estimateur GREG sous le modèle suivant :

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\beta}^T\boldsymbol{x}_i \\ &\mathrm{var}_M[y_i] = v_i\sigma^2 \\ &y_i,\ y_j \text{ indépendantes pour } i \neq j. \end{aligned}\right\} \qquad (4)$$

Donc, l'estimateur GREG au niveau de la personne, $\hat{T}_P$, est donné par substitution de $\boldsymbol{x}_i$ à $\boldsymbol{z}_i$ dans (2). Le modèle (4) ignore toute corrélation entre y_i et y_j pour les personnes i et j dans le même ménage. Ces corrélations étaient égales ou inférieures à 0,3 pour la plupart des variables considérées par Clark et Steel (2002), quoiqu'ils aient observé des valeurs plus élevées pour les variables associées à l'ethnicité, comme l'auto-identification en tant qu'Autochtone. Des corrélations de valeur 1 pourraient survenir pour des variables environnementales. Tam (1995) montre que l'estimateur assisté par modèle optimal pour l'échantillonnage en grappes est robuste à l'erreur de spécification des corrélations intra-grappe. Une interprétation de ce résultat serait que les corrélations à l'intérieur du ménage ne sont pas pertinentes en ce qui concerne l'estimation des totaux de population, parce que tous les membres des ménages échantillonnés sont sélectionnés. Donc, les corrélations à l'intérieur des ménages ne facilitent pas l'estimation pour les personnes non échantillonnées, puisque les personnes échantillonnées et non échantillonnées appartiennent à des ménages distincts.

Un certain nombre de méthodes ont été proposées pour l'estimation de type GREG avec pondérations égales dans les ménages. Nieuwenbroek (1993) a proposé un estimateur dont la motivation était l'agrégation du modèle (4) au niveau du ménage :

$$\left.\begin{aligned} &E_M[y_{g1}] = \boldsymbol{\beta}^T\boldsymbol{x}_{g1} \\ &\mathrm{var}_M[y_{g1}] = v_{g1}\sigma^2 \\ &y_{g1},\ y_{k1} \text{ indépendantes pour } g \neq k. \end{aligned}\right\} \qquad (5)$$

où $v_{g1} = \sum_{i\in U_g} v_i$. L'estimateur GREG calculé en utilisant les données d'échantillon y_{g1} pour $g \in s_1$ basé sur ce modèle est $\hat{T}_H$:

$$\hat{T}_H = \hat{T}_\pi + \hat{\boldsymbol{\beta}}_H^T\,(\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi}) \qquad (6)$$

où $\hat{\boldsymbol{\beta}}_H$ est une solution de

$$\sum_{g\in s_1} \pi_{g1}^{-1} a_g\,(y_{g1} - \hat{\boldsymbol{\beta}}_H^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1} = \boldsymbol{0}. \qquad (7)$$

Le coefficient de régression $\hat{\boldsymbol{\beta}}_H$ est une régression par les moindres carrés pondérés au niveau du ménage des valeurs d'échantillon de y_{g1} sur $\boldsymbol{x}_{g1}$ avec les pondérations $\pi_{g1}^{-1}a_g$. Les valeurs de a_g pourraient être fixées à v_{g1}^{-1}. Si $v_i = 1$ alors $v_{g1} = N_g$, de sorte que $a_g = N_g^{-1}$. Sinon, $a_g = 1$ pourrait également être utilisé.

Plusieurs autres méthodes à pondération intégrée équivalentes ont été utilisées. Lemaître et Dufour (1987) ont construit un estimateur par la régression généralisée au niveau de la personne en utilisant $\bar{\boldsymbol{x}}_g$ au lieu de $\boldsymbol{x}_i$ comme variables auxiliaires. Nieuwenbroek (1993) a fait remarquer que cela est équivalent à (6) si $c_i = a_g N_g$ pour $i \in U_g$. Alexander (1987) a élaboré des méthodes de pondération étroitement reliées en utilisant un critère de distance minimale.

Les estimateurs GREG aux niveaux de la personne et du ménage peuvent tous deux s'écrire sous la forme pondérée $\sum_{i\in s} w_i Y_i$. Les pondérations pour les deux estimateurs peuvent s'écrire $w_i = \pi_i^{-1} g_i$ où

$$g_i = 1 + (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})^T \left(\sum_{i\in s} c_i\pi_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} c_i\,\boldsymbol{x}_i$$

pour $\hat{T}_P$ et

$$g_i = 1 + (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})^T \left(\sum_{g \in s_1} a_g \, \pi_{g1}^{-1} \, \boldsymbol{x}_{g1} \, \boldsymbol{x}_{g1}^T \right)^{-1} a_g \, \pi_g^{-1} \, \boldsymbol{x}_{g1}$$

pour $\hat{T}_H$, où la personne i appartient au ménage g. (L'indice supérieur « - » indique l'inverse généralisée d'une matrice).

2.3 Résultats théoriques

À la présente section, nous montrons que $\hat{T}_H$ possède la variance en grand échantillon la plus faible possible dans une classe d'estimateurs qui comprend aussi $\hat{T}_P$ pour le plan de sondage où les ménages sont sélectionnés par échantillonnage aléatoire simple sans remise. Puis, nous expliquons ce résultat en montrant que $\hat{T}_H$ est équivalent à un estimateur par la régression calculé en utilisant les données au niveau de la personne, où le modèle contient des effets contextuels.

Pour les grands échantillons, $\hat{T}_P$ et $\hat{T}_H$ peuvent être approximés par

$$\tilde{T}_P = \hat{T}_\pi + \boldsymbol{B}_P^T (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi});$$

et

$$\tilde{T}_H = \hat{T}_\pi + \boldsymbol{B}_H^T (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})$$

respectivement, où $\boldsymbol{B}_P$ et $\boldsymbol{B}_H$ sont les solutions de

$$\left.\begin{aligned} &\sum_{i \in U} c_i (y_i - \boldsymbol{B}_P^T \boldsymbol{x}_i) \, x_i = \boldsymbol{0} \\ &\sum_{g \in U_1} a_g (y_{g1} - \boldsymbol{B}_H^T \boldsymbol{x}_{g1}) \, \boldsymbol{x}_{g1} = \boldsymbol{0} \end{aligned}\right\} \qquad (8)$$

(Särndal et coll., 1992, Résultat 6.6.1, page 235). Le théorème 1 énonce l'estimateur à variance minimale dans une classe incluant $\tilde{T}_P$ et $\tilde{T}_H$.

Théorème 1. Estimateur optimal pour l'échantillonnage en grappes simple

Supposons que m ménages soient sélectionnés par échantillonnage aléatoire simple sans remise à partir d'une population de M ménages, et que tous les membres des ménages échantillonnés soient sélectionnés. Considérons l'estimateur de T donné par

$$\tilde{T} = \hat{T}_\pi + \boldsymbol{h}^T (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})$$

où $\boldsymbol{h}$ est un vecteur de dimension p constant. Nous supposons qu'il existe un vecteur $\boldsymbol{\lambda}$ tel que $\boldsymbol{\lambda}^T \boldsymbol{x}_i = 1$ pour tout $i \in U$. La variance de cet estimateur est minimisée par les valeurs h^* qui sont les solutions de

$$\sum_{g \in s_1} (y_{g1} - \boldsymbol{h}^T \boldsymbol{x}_{g1}) \, \boldsymbol{x}_{g1} = \boldsymbol{0}.$$

Donc, $\tilde{T}_H$ avec $a_g = 1$ pour tous g est le choix optimal pour $\tilde{T}$.

Le théorème 1 a l'implication peut-être étonnante que $\hat{T}_H$ (avec $a_g = 1$ pour tout g) a une variance plus faible que $\hat{T}_P$ pour les grands échantillons, et cela en dépit du fait que $\hat{T}_H$ écarte une partie de l'information contenue dans l'échantillon, parce qu'il utilise les sommes de $\boldsymbol{x}_i$ et y_i sur l'ensemble des ménages. Le théorème donne à penser que $\hat{T}_H$ est l'estimateur GREG approprié pour le plan d'échantillonnage en grappes supposé ici et que l'information écartée par sommation au niveau du ménage n'est pas pertinente quand on utilise ce plan. Pour expliquer pourquoi $\hat{T}_H$ peut donner de meilleurs résultats que $\hat{T}_P$, nous utiliserons un « modèle contextuel linéaire » qui est un modèle plus général de $E_M[Y_i]$ que (4). Ce modèle est :

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\gamma}_1^T \bar{\boldsymbol{x}}_g + \boldsymbol{\gamma}_2^T \boldsymbol{x}_i \;\; (i \in U_g) \\ &\mathrm{var}_M[y_i] = \sigma^2 \\ &y_i, \, y_j \text{ indépendantes pour } i \neq j. \end{aligned}\right\} \qquad (9)$$

Nous utilisons $\bar{\boldsymbol{x}}_g$ ainsi que $\boldsymbol{x}_i$ comme variables explicatives de y_i parce que la moyenne au niveau du ménage des variables auxiliaires au niveau de la personne peut refléter une partie de l'effet du contexte des ménages (Lazarfeld et Menzel 1961). Par exemple, si les éléments de $\boldsymbol{x}_i$ sont des variables indicatrices qui résument l'âge et le sexe de la personne i, alors $\bar{\boldsymbol{x}}_g$ représente les proportions de personnes dans le ménage qui appartiennent à différentes catégories âge-sexe. Si la population d'intérêt comprend des adultes ainsi que des enfants, alors $\bar{\boldsymbol{x}}_g$ comprend la proportion d'enfants dans le ménage, laquelle pourrait être en rapport avec la situation d'activité des adultes faisant partie du ménage.

Le théorème 2 montre que l'amélioration de la variance due à l'utilisation de $\tilde{T}_H$ avec $a_g = 1$ plutôt que $\tilde{T}_P$ peut être expliquée par le modèle contextuel linéaire.

Théorème 2. Explication de la différence entre les variances asymptotiques

Supposons que les ménages soient sélectionnés par échantillonnage aléatoire simple sans remise et que tous les membres des ménages échantillonnés soient sélectionnés. Soit $r_i = y_i - \boldsymbol{B}_P^T \boldsymbol{x}_i$, et soit $\boldsymbol{B}_C$ le résultat de la régression de r_i en fonction de $\bar{\boldsymbol{x}}_g$ sur les $i \in U$ par la méthode des moindres carrés pondérés en utilisant N_g comme pondération. Alors

$$\text{var}_p[\tilde{T}_P] - \text{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\boldsymbol{B}_C^{\boldsymbol{T}}\left(\sum_{g\in U_1}\boldsymbol{x}_{\boldsymbol{g1}}\boldsymbol{x}_{\boldsymbol{g1}}^{\boldsymbol{T}}\right)\boldsymbol{B}_C$$

où $\tilde{T}_H$ est calculé en utilisant $a_g = 1$ pour tout g.

Le résultat montre que la réduction de la variance due à l'utilisation de $\tilde{T}_H$ (avec $a_g = 1$) plutôt que $\tilde{T}_P$ est une forme quadratique en $\boldsymbol{B}_C$. Donc, l'importance de l'amélioration dépend de la mesure dans laquelle $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ aide à prédire y_i quand on a déjà neutralisé $\boldsymbol{x}_{\boldsymbol{i}}$, c'est-à-dire la mesure dans laquelle un effet contextuel linéaire aide à prédire r_i sur les $i \in U$, en utilisant une régression par les moindres carrés pondérés avec N_g comme pondération.

Les preuves de théorèmes 1 et 2 dépendent fortement de l'hypothèse d'échantillonnage en grappes. On ne s'attendrait pas à ce que les résultats soient applicables en cas de sous-échantillonnage dans les ménages.

Les théorèmes 1 et 2 s'appliquent uniquement quand $a_g = 1$ dans la régression par les moindres carrés pondérés pour $\hat{T}_H$. D'autres choix de a_g sont souvent utilisés; par exemple, il serait souvent raisonnable de supposer que $v_{g1} = N_g$ dans le modèle (5), auquel cas il serait logique d'utiliser $a_g = N_g^{-1}$. Le théorème 3 montre que $\hat{T}_H$ est équivalent à un estimateur GREG au niveau de la personne ajusté sous le modèle contextuel linéaire pour d'autres choix de a_g.

Théorème 3. L'estimateur GREG contextuel linéaire

Pour les plans d'échantillonnage où toutes les personnes sont sélectionnées dans les ménages échantillonnés et $\pi_{g1} > 0$ pour tout $g \in U_1$, $\hat{T}_H$ avec un choix donné de a_g est l'estimateur par la régression généralisée pour le modèle (9) où $c_i = a_g N_g$ pour $i \in U_g$.

Le théorème 3 signifie que $\hat{T}_H$ est l'estimateur GREG sous un modèle plus général que $\hat{T}_P$. Nieuwenbroek (1993) a montré que $\hat{T}_H$ est égal à un estimateur GREG au niveau de la personne dérivé par régression de y_i sur $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$. Le théorème 3 énonce qu'il est également égal à l'estimateur GREG au niveau de la personne provenant de la régression de y_i sur $\boldsymbol{x}_{\boldsymbol{i}}$ ainsi que sur $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$, donc qu'il intègre automatiquement tous effets contextuels du ménage. Par conséquent, $\hat{T}_H$ devrait en principe avoir une variance plus faible que $\hat{T}_P$ pour les grands échantillons. (Dans le cas $a_g = 1$, le théorème 1 énonçait que cela est toujours le cas). Pour les petits échantillons, par contre, un modèle plus général est parfois contreproductif. Silva et Skinner (1997) ont montré, pour l'échantillonnage à un seul degré, que l'ajout de paramètres au modèle peut accroître la variance de l'estimateur GREG, quoique cet effet soit négligeable pour les grands échantillons. Il est possible que les effets contextuels n'aient que peu de pouvoir prédictif, voire aucun, pour certaines variables. Le cas échéant, on s'attendrait à ce que $\hat{T}_H$ donne d'un peu moins bons résultats que $\hat{T}_P$ pour les petits échantillons, et à peu près les mêmes résultats pour les grands échantillons.

Le modèles contextuel (9) contient tous les éléments de $\boldsymbol{x}_{\boldsymbol{i}}$ et tous les éléments de $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$. Une autre solution consisterait à utiliser uniquement les éléments de $\boldsymbol{x}_{\boldsymbol{i}}$ et $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ qui sont significatifs, ou qui produisent des améliorations de la variance estimée d'un estimateur GREG. Un estimateur GREG basé sur ce genre de modèle aurait probablement une variance plus faible que les estimateurs examinés dans le présent article, mais ne donnerait pas de pondérations intégrées, à moins d'utiliser les mêmes éléments de $\boldsymbol{x}_{\boldsymbol{i}}$ et $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$.

3. Étude empirique

3.1 Méthodologie

Nous avons entrepris une étude par simulation en vue de comparer les estimateurs GREG aux niveaux de la personne et du ménage, $\hat{T}_P$ et $\hat{T}_H$, pour une gamme de variables d'enquête. Nous avons utilisé deux populations, constituées de 187 178 ménages sélectionnés aléatoirement à partir du recensement de la population de l'Australie de 2001 et de 210 132 ménages provenant de l'enquête nationale sur la santé de la population australienne de 1995. Tous les adultes et enfants faisant partie de ces ménages ont été inclus dans l'étude. La taille moyenne des ménages était de 2,5 environ.

Nous avons tirés des échantillons en grappes à partir de ces populations, en sélectionnant d'abord des ménages par échantillonnage aléatoire simple sans remise, puis tous les membres des ménages échantillonnés. Nous avons simulés des échantillons de taille m = 500, 1 000, 2 000, 5 000 et 10 000 ménages. Dans chaque cas, 5 000 échantillons ont été sélectionnés. Les variables auxiliaires $\boldsymbol{x}_{\boldsymbol{i}}$ correspondaient à des variables indicatrices du sexe selon le groupe d'âge (12 catégories). (Ce choix de $\boldsymbol{x}_{\boldsymbol{i}}$ signifie que l'estimation GREG équivaut à une poststratification.) Nous avons calculé l'estimateur GREG au niveau de la personne avec $c_i = 1(\hat{T}_P)$, l'estimateur GREG au niveau du ménage avec $a_g = N_g^{-1}(\hat{T}_{H1})$, et l'estimateur GREG au niveau du ménage avec $a_g = 1(\hat{T}_{H2})$. Nous avons également inclus l'estimateur de Hájek

$$\hat{T}_1 = N\left(\sum_{i\in s}\pi_i^{-1}y_i\right)\Big/\left(\sum_{i\in s}\pi_i^{-1}\right)$$

qui est égal à $N/n \sum_{g \in s_1} \sum_{i \in U_g} y_i$ pour l'échantillonnage en grappes avec échantillonnage aléatoire simple des ménages où n est la taille réalisée de l'échantillon de personnes.

Les variables comprennent la situation d'activité, l'état de santé et d'autres caractéristiques. Toutes les variables sont dichotomiques, sauf celle du revenu (revenu annuel en dollars australiens, basé sur les données déclarées pour les fourchettes de revenu du recensement). « Occupée (F) » est une variable indicatrice dont la valeur est 1 si une personne est occupée et de sexe féminin, et 0 autrement. Les six premières variables sont tirées du recensement de la population et les cinq autres, de l'enquête sur la santé de la population.

3.2 Résultats

Le tableau 1 donne la racine de l'erreur quadratique moyenne relative (REQMR) de $\hat{T}_1$, $\hat{T}_P$, $\hat{T}_{H1}$ et $\hat{T}_{H2}$, pour une taille d'échantillon de 1 000 ménages. Les REQMR sont exprimés en pourcentage du total réel de population. Les biais n'ont pas été tabulés, parce qu'ils représentaient une composante négligeable de l'EQM dans tous les cas. Le pourcentage d'amélioration de l'EQM de $\hat{T}_{H1}$ et de $\hat{T}_{H2}$ relativement à $\hat{T}_P$ est également présenté. Les chiffres entre crochets sont les erreurs-types de simulation de ces pourcentages d'amélioration.

Pour la taille d'échantillon susmentionnée, $\hat{T}_{H1}$ et $\hat{T}_{H2}$ donnent des résultats un peu moins bons que $\hat{T}_P$ pour les variables relatives à la santé et un peu meilleurs pour la plupart des autres variables. Nous observons le gain le plus important pour l'estimation du nombre de parents seuls; cette variance a été réduite de 10,8 % et de 16,3 %, respectivement, en utilisant les deux estimateurs GREG au niveau du ménage. Pour toutes les autres variables, l'amélioration est faible ou bien l'estimateur GREG au niveau du ménage donne d'un peu moins bons résultats que l'estimateur GREG au niveau de la personne. L'inefficacité due à l'utilisation d'un estimateur GREG au niveau du ménage plutôt que $\hat{T}_P$ n'est jamais supérieure à 2,2 %.

Le tableau 2 montre le pourcentage d'amélioration de l'EQM résultant de l'utilisation de $\hat{T}_{H1}$ plutôt que $\hat{T}_P$ pour diverses tailles d'échantillon. Pour chaque chiffre, l'erreur-type de simulation est indiquée entre crochets. Le tableau 3 donne le pourcentage d'amélioration résultant de l'utilisation de $\hat{T}_{H2}$ plutôt que $\hat{T}_P$. Sont également présentés les pourcentages d'amélioration asymptotiques $(m = \infty)$ basés sur l'approximation en grand échantillon de la variance d'un estimateur GREG. Pour les deux estimateurs GREG au niveau du ménage, le pourcentage d'amélioration augmente généralement avec la taille d'échantillon. Pour $m = 500$, les estimateurs GREG au niveau du ménage sont généralement pires que l'estimateur GREG au niveau de la personne, quoique jamais de plus de 5 %. Pour $m = 10\,000$, nous observons une amélioration pour plus de la moitié des variables. Les améliorations les plus importantes sont celles constatées pour les estimations du nombre de parents seuls (11,5 %) et de femmes occupées (4,2 %); toutes les autres améliorations sont faibles. $\hat{T}_{H1}$ et $\hat{T}_{H2}$ n'ont jamais une variance surpassant de plus de 0,2 % celle de $\hat{T}_P$ pour $m = 10\,000$. En général, $\hat{T}_{H2}$ donne de meilleurs résultats que $\hat{T}_{H1}$ pour les échantillons de plus grande taille, comme on s'y attendrait d'après le théorème 1, mais l'inverse est également vrai pour les petites tailles d'échantillon.

En pratique, les estimations des totaux de sous-population présentent souvent autant d'intérêt que les totaux de population. Le tableau 4 montre les propriétés des divers estimateurs pour les domaines âge-sexe (12 catégories d'âge) et les domaines régionaux, pour une taille d'échantillon de 1 000 ménages. L'ensemble de données du recensement comportait 49 régions. L'ensemble de données de l'enquête sur la santé de la population ne contenait aucune variable de région semblable, de sorte que nous avons utilisé à la place le quintile socioéconomique du district de collecte (une unité géographique constituée d'environ 200 ménages contigus). Pour produire les estimateurs de domaine, nous avons calculé les pondérations à partir de chaque estimateur, puis pris la somme pondérée sur l'ensemble de l'échantillon dans le domaine. Cela équivaut à l'estimateur par le ratio pour le domaine décrit au cas 1, section 2.1 de Hidiroglou et Patak (2004). Nous avons suivi cette méthode parce qu'elle est la plus utilisée en pratique, car elle permet d'estimer tous les totaux de domaine et de population à l'aide d'un seul ensemble de pondérations, quoique des estimateurs de domaine plus efficaces existent (Hidiroglou et Patak 2004, cas 2 à 6).

Dans chaque cas, nous présentons la REQMR médiane sur les domaines. Le tableau montre que les différences entre les trois estimateurs GREG sont faibles. Pour les domaines âge-sexe, les estimateurs GREG au niveau du ménage donnent d'un peu de meilleurs résultats que l'estimateur GREG au niveau de la personne pour les variables de recensement, et d'un peu moins bons pour les variables de l'enquête sur la santé de la population. Pour les estimations régionales, les estimateurs GREG au niveau du ménage sont un peu moins bons dans tous les cas. Le tableau 5 montre que les propriétés de l'estimateur GREG au niveau du ménage sont fort semblables à celles de $\hat{T}_P$ pour une taille d'échantillon de 10 000 ménages. Il convient de souligner que les théorèmes 1 et 2 ne s'appliquent pas aux estimateurs de domaine que nous avons utilisés.

Tableau 1 REQMR relative pour une taille d'échantillon de 1 000 ménages

Variable	REQMR en %				% d'amélioration de l'EQM	
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
Occupé(e)	2,62	2,09	2,09	2,10	0,20 (0,26)	-0,28 (0,27)
Occupée F	3,78	3,05	3,01	3,02	2,63 (0,33)	2,09 (0,33)
Revenu	2,56	2,20	2,19	2,19	1,04 (0,25)	0,75 (0,24)
Faible revenu	5,04	4,87	4,89	4,90	-0,62 (0,20)	-1,12 (0,22)
Heures travaillées	3,08	2,54	2,53	2,53	0,94 (0,28)	0,70 (0,28)
Parent seul	12,50	12,73	12,02	11,65	10,84 (0,62)	16,31 (0,49)
Arthrite	5,52	4,50	4,53	4,53	-1,38 (0,17)	-1,57 (0,18)
Fumeur(se)	4,73	4,57	4,60	4,61	-1,64 (0,18)	-1,81 (0,20)
PA élevée	6,80	5,30	5,35	5,36	-1,70 (0,17)	-2,06 (0,18)
Santé passable/mauvaise	9,79	9,42	9,47	9,47	-1,16 (0,16)	-1,07 (0,18)
Alcool	4,81	4,66	4,70	4,71	-1,77 (0,16)	-2,15 (0,18)

Tableau 2 Amélioration de l'EQM de l'estimateur GREG au niveau du ménage $\hat{T}_{H1}$ comparativement à $\hat{T}_P$

Variable	% d'amélioration de l'EQM					
	m = 500	1 000	2 000	5 000	10 000	∞
Occupé(e)	-0,65 (0,31)	0,20 (0,26)	1,02 (0,24)	0,90 (0,21)	2,17 (0,21)	1,85
Occupée F	1,22 (0,37)	2,63 (0,33)	2,59 (0,33)	3,53 (0,31)	4,24 (0,31)	4,13
Revenu	-1,53 (0,31)	1,04 (0,25)	0,48 (0,24)	0,61 (0,19)	1,43 (0,19)	1,07
Faible revenu	-2,45 (0,27)	-0,62 (0,20)	0,02 (0,18)	0,18 (0,15)	0,00 (0,00)	0,65
Heures travaillées	-0,26 (0,34)	0,94 (0,28)	1,72 (0,27)	1,61 (0,24)	2,64 (0,24)	2,12
Parent seul	7,81 (0,69)	10,84 (0,62)	10,74 (0,61)	10,23 (0,57)	11,50 (0,58)	11,21
Arthrite	-3,01 (0,24)	-1,38 (0,17)	-0,34 (0,12)	-0,08 (0,09)	-0,13 (0,07)	0,08
Fumeur(se)	-3,91 (0,25)	-1,64 (0,18)	-1,02 (0,12)	-0,26 (0,08)	-0,06 (0,07)	0,16
PA élevée	-2,93 (0,24)	-1,70 (0,17)	-0,86 (0,12)	-0,31 (0,08)	-0,04 (0,06)	0,08
Santé passable/mauvaise	-3,67 (0,25)	-1,16 (0,16)	-0,71 (0,12)	-0,05 (0,08)	0,03 (0,06)	0,10
Alcool	-4,22 (0,23)	-1,77 (0,16)	-0,77 (0,12)	-0,31 (0,08)	-0,21 (0,07)	0,14

Tableau 3 Amélioration de l'EQM de l'estimateur GREG au niveau du ménage $\hat{T}_{H2}$ comparativement à $\hat{T}_P$

Variable	% d'amélioration de l'EQM					
	m = 500	1 000	2 000	5 000	10 000	∞
Occupé(e)	-1,85 (0,35)	-0,28 (0,27)	1,25 (0,25)	1,05 (0,21)	2,22 (0,21)	1,98
Occupée F	0,28 (0,39)	2,09 (0,33)	2,71 (0,33)	3,55 (0,29)	4,50 (0,30)	4,31
Revenu	-2,64 (0,31)	0,75 (0,24)	0,71 (0,22)	0,90 (0,17)	1,30 (0,16)	1,37
Faible revenu	-3,15 (0,30)	-1,12 (0,22)	-0,15 (0,18)	0,06 (0,15)	0,00 (0,00)	0,94
Heures travaillées	-1,51 (0,35)	0,70 (0,28)	1,98 (0,25)	1,79 (0,21)	2,57 (0,22)	2,26
Parent seul	14,70 (0,53)	16,31 (0,49)	16,39 (0,47)	15,41 (0,44)	16,44 (0,44)	16,35
Arthrite	-3,31 (0,26)	-1,57 (0,18)	-0,05 (0,13)	-0,12 (0,09)	-0,10 (0,07)	0,16
Fumeur(se)	-3,82 (0,28)	-1,81 (0,20)	-0,69 (0,14)	0,21 (0,11)	0,28 (0,10)	0,57
PA élevée	-3,20 (0,26)	-2,06 (0,18)	-1,12 (0,13)	-0,40 (0,09)	-0,05 (0,07)	0,12
Santé passable/mauvaise	-4,02 (0,28)	-1,07 (0,18)	-0,57 (0,13)	-0,09 (0,09)	0,00 (0,07)	0,15
Alcool	-5,00 (0,26)	-2,15 (0,18)	-0,82 (0,13)	-0,49 (0,09)	-0,29 (0,08)	0,18

Tableau 4 REQMR relative médiane pour les estimateurs de domaine pour une taille d'échantillon $m = 1\,000$

Variable	Domaines âge-sexe				Domaines régionaux			
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
Occupé(e)	12,74	7,92	7,93	7,90	29,89	29,92	30,20	30,34
Occupée F	13,12	8,32	8,36	8,34	34,64	34,65	35,03	35,16
Revenu	13,25	8,43	8,49	8,47	28,04	28,12	28,43	28,51
Faible revenu	21,17	18,77	18,96	18,94	42,71	42,85	43,24	43,33
Heures travaillées	14,56	10,69	10,76	10,72	31,24	31,23	31,52	31,63
Parent seul	96,20	96,33	97,64	96,69	92,99	93,30	94,37	93,50
Arthrite	24,94	20,94	21,12	21,11	13,31	12,94	13,02	13,04
Fumeur(se)	32,10	29,25	29,39	29,37	12,32	12,27	12,35	12,38
PA élevée	27,01	23,80	23,97	23,95	15,83	15,31	15,44	15,45
Santé passable/mauvaise	39,64	37,73	38,05	38,08	22,38	22,30	22,51	22,55
Alcool	25,58	21,42	21,53	21,58	12,73	12,70	12,80	12,82

Tableau 5 REQMR relative médiane pour les estimateurs de domaine pour une taille d'échantillon $m = 10\,000$

Variable	Domaines âge-sexe				Domaines régionaux			
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
Occupé(e)	3,77	2,35	2,32	2,31	8,85	8,85	8,87	8,88
Occupée F	3,86	2,43	2,43	2,42	10,30	10,26	10,25	10,25
Revenu	3,91	2,53	2,51	2,51	8,24	8,23	8,23	8,24
Faible revenu	6,31	5,63	5,62	5,61	12,67	12,68	12,69	12,69
Heures travaillées	4,29	3,15	3,15	3,12	9,26	9,25	9,27	9,27
Parent seul	28,40	28,26	28,29	28,23	27,11	27,14	27,16	27,11
Arthrite	7,40	6,26	6,27	6,27	3,98	3,85	3,85	3,85
Fumeur(se)	9,53	8,58	8,58	8,57	3,69	3,67	3,68	3,67
PA élevée	8,07	7,02	7,01	7,01	4,66	4,48	4,49	4,49
Santé passable/mauvaise	11,69	11,02	11,02	11,01	6,75	6,69	6,69	6,69
Alcool	7,74	6,43	6,43	6,43	3,87	3,85	3,85	3,85

4. Discussion

L'estimateur GREG au niveau de la personne standard produit des pondérations inégales à l'intérieur des ménages. Les estimateurs GREG au niveau du ménage peuvent être utilisés pour obtenir des pondérations intégrées au niveau du ménage et de la personne, ce qui est avantageux pour les enquêtes recueillant de l'information sur des variables au niveau du ménage ainsi qu'au niveau de la personne. Nous avons démontré dans le présent article que l'avantage pratique de la pondération intégrée découlant de l'utilisation d'un estimateur GREG au niveau du ménage est associé à une perte faible, voire nulle. Pour les grands échantillons, l'estimateur GREG au niveau du ménage a une variance sous le plan plus faible que l'estimateur GREG au niveau de la personne. Pour les échantillons plus petits, l'utilisation de l'estimateur GREG au niveau du ménage produit, au plus, un faible accroissement de la variance pour certaines variables, parce que cet estimateur est équivalent à l'utilisation d'un modèle de régression contenant un plus grand nombre de paramètres. Par conséquent, si les pondérations intégrées améliorent la cohérence des données de sortie d'une enquête-ménage, l'adoption de l'estimateur GREG au niveau du ménage ne causera qu'un accroissement faible, voire nul, de la variance et du biais des estimateurs.

Remerciements

Les présents travaux ont été financés conjointement par l'Australian Research Council et l'Australian Bureau of Statistics. Les opinions exprimées ici ne reflètent pas forcément celles de ces organismes. Les auteurs remercient Julian England, Frank Yu et Ray Chambers de leurs commentaires constructifs.

Annexe

Preuve des théorèmes

Preuve du théorème 1

Soit $\bar{Y}_1 = T_Y/M$ et $\bar{\boldsymbol{X}}_1 = \boldsymbol{T}_X/M$ les moyennes de population de y_{g1} et $\boldsymbol{x}_{g1}$ respectivement. La variance de $\tilde{T}$ est

$$\begin{aligned}\text{var}_p[\tilde{T}] &= \text{var}[\hat{T}_\pi + \boldsymbol{h}^T(\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})] \\ &= \text{var}\left[\frac{M}{m}\sum_{g\in s_1}(y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1})\right] \\ &= \frac{M^2}{m}\left(1-\frac{m}{M}\right)S_r^2\end{aligned}$$

où $S_r^2 = (M-1)^{-1}\sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}^2$. Pour minimiser par rapport à $\boldsymbol{h}$, nous fixons la dérivée de S_r^2 à zéro :

$$\begin{aligned}0 &= (M-1)^{-1}\sum_{g\in U_1} \\ &\quad \{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\,(\boldsymbol{x}_{g1} - \bar{\boldsymbol{X}}_1) \\ 0 &= \sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\,\boldsymbol{x}_{g1} \\ &\quad -\sum_{g\in U_1}\{(y_{g1} - \bar{Y}_1) - \boldsymbol{h}^T(\boldsymbol{x}_{g1} - \bar{\boldsymbol{X}}_1)\}\bar{\boldsymbol{X}}_1 \\ 0 &= \sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\boldsymbol{x}_{g1} \\ 0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\,\boldsymbol{T}_X. \quad (10)\end{aligned}$$

Nous montrons maintenant que (10) est satisfaite par $\boldsymbol{h}^*$. Par hypothèse, $\boldsymbol{h}^*$ satisfait

$$\boldsymbol{0} = \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*)\,\boldsymbol{x}_{g1}. \quad (11)$$

Donc, la première somme dans le deuxième membre de (10) est égale à zéro pour $\boldsymbol{h} = \boldsymbol{h}^*$. La prémultiplication des deux membres de (11) par $\boldsymbol{\lambda}^T$ donne

$$\begin{aligned}0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*)\,\boldsymbol{\lambda}^T\boldsymbol{x}_{g1} \\ 0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*) \\ 0 &= T_Y - \boldsymbol{T}_X^T\boldsymbol{h}^*.\end{aligned}$$

La division par M donne $\bar{Y}_1 - \bar{\boldsymbol{X}}_1^T\boldsymbol{h}^* = 0$. Donc, le reste du deuxième membre de (10) est égal à zéro. Donc, $\boldsymbol{h}^*$ satisfait (10).

Preuve du théorème 2

Dénotons par « - » l'inverse généralisée d'une matrice. Alors, $\boldsymbol{B}_C$ est égal à

$$\begin{aligned}\boldsymbol{B}_C &= \left\{\sum_{g\in U_1}\sum_{i\in U_g} N_g\bar{\boldsymbol{x}}_g\bar{\boldsymbol{x}}_g^T\right\}^-\sum_{g\in U_1}\sum_{i\in U_g} N_g\bar{\boldsymbol{x}}_g r_i \\ &= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}r_{g1}. \quad (12)\end{aligned}$$

Or, $r_i = y_i - \boldsymbol{B}_P^T\boldsymbol{x}_i$ de sorte que $r_{g1} = y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1}$. Donc, (12) devient

$$\begin{aligned}\boldsymbol{B}_C &= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}(y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1}) \\ &= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}y_{g1} \\ &\quad -\left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_P \\ &= \boldsymbol{B}_H - \boldsymbol{B}_P \quad (13)\end{aligned}$$

puisque $\boldsymbol{B}_H = \{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\}\sum_{g\in U_1}\sum_{i\in U_g}\boldsymbol{x}_{g1}y_{g1}$. La différence entre les variances est donnée par

$$\begin{aligned}&\text{var}_p[\tilde{T}_P] - \text{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1-\frac{m}{M}\right)(M-1)^{-1} \\ &\qquad\left\{\sum_{g\in U_1}(y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1})^2 - \sum_{g\in U_1}(y_{g1} - \boldsymbol{B}_H^T\boldsymbol{x}_{g1})^2\right\}\end{aligned}$$

qui devient

$$\begin{aligned}&\{\text{var}_p[\tilde{T}_P] - \text{var}_p[\tilde{T}_H]\}\Big/\left\{\frac{M^2}{m}\left(1-\frac{m}{M}\right)(M-1)^{-1}\right\} \\ &= \sum_{g\in U_1} r_{g1}^2 - \sum_{g\in U_1}(r_{g1} + \boldsymbol{B}_P^T\boldsymbol{x}_{g1} - \boldsymbol{B}_H^T\boldsymbol{x}_{g1})^2 \\ &= \sum_{g\in U_1} r_{g1}^2 - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\ &= \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1} + \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\ &= \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 + \sum_{g\in U_1}(\boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\ &\quad + 2\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1}^T\boldsymbol{B}_C^T \\ &\quad - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\ &= \sum_{g\in U_1}\boldsymbol{B}_C^T\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_C + 2\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1}^T\boldsymbol{B}_C. \quad (14)\end{aligned}$$

Or, $\boldsymbol{B_C}$ est une régression par les moindres carrés ordinaires de r_{g1} sur $\boldsymbol{x_{g1}}$, de sorte que

$$\sum_{g\in U_1} (r_{g1} - \boldsymbol{B_C^T x_{g1}})\, \boldsymbol{x_{g1}} = \boldsymbol{0}.$$

Donc, (14) devient

$$\text{var}_p[\tilde{T}_P] - \text{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\boldsymbol{B_C^T}\sum_{g\in U_1}\boldsymbol{x_{g1}x_{g1}^T B_C}.$$

Preuve du théorème 3

L'estimateur GREG est invariant sous des transformations linéaires inversibles des variables auxiliaires. Donc, le modèle (9) peut être reparamétrisé pour donner

$$E_M[y_i] = \boldsymbol{\phi_1^T \bar{x}_g} + \boldsymbol{\phi_2^T}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g}) \qquad (15)$$

ou, de manière équivalente,

$$E_M[y_i] = \boldsymbol{\phi^T z_i}$$

où

$$\boldsymbol{z_i} = \begin{pmatrix} \boldsymbol{\bar{x}_g} \\ \boldsymbol{x_i} - \boldsymbol{\bar{x}_g} \end{pmatrix}$$

et

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi_1} \\ \boldsymbol{\phi_2} \end{pmatrix}.$$

Les paramètres du modèle (15) sont reliés à ceux du modèle (9) par $\boldsymbol{\phi_1} = \boldsymbol{\gamma_1} + \boldsymbol{\gamma_2}$ et $\boldsymbol{\phi_2} = \boldsymbol{\gamma_2}$.

Partant de la définition 1, en notant que

$$s = \bigcup_{g\in s_1} U_g$$

pour le plan supposé, l'estimateur par la régression généralisée sous le modèle (15) est

$$\begin{aligned}\hat{T} &= \hat{T}_\pi + \sum_{i\in U}\hat{\boldsymbol{\phi}}^{\boldsymbol{T}}\boldsymbol{z_i} - \sum_{i\in s}\pi_i^{-1}\hat{\boldsymbol{\phi}}^{\boldsymbol{T}}\boldsymbol{z_i} \\ &= \hat{T}_\pi + \sum_{g\in U_1}\sum_{i\in U_g}\{\hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} + \hat{\boldsymbol{\phi}}_{\boldsymbol{2}}^{\boldsymbol{T}}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g})\} \\ &\quad - \sum_{g\in s_1}\sum_{i\in U_g}\pi_i^{-1}\{\hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} + \hat{\boldsymbol{\phi}}_{\boldsymbol{2}}^{\boldsymbol{T}}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g})\}. \qquad (16)\end{aligned}$$

Cependant, $\sum_{i\in U_g}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g}) = \boldsymbol{0}$ pour chaque g. Donc (16) devient

$$\begin{aligned}\hat{T} &= \hat{T}_\pi + \sum_{g\in U_1}\sum_{i\in U_g}\hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} - \sum_{g\in s_1}\sum_{i\in U_g}\pi_i^{-1}\hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\sum_{g\in U_1}\sum_{i\in U_g}\boldsymbol{\bar{x}_g} - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\sum_{g\in s_1}\pi_{g1}^{-1}\sum_{i\in U_g}\boldsymbol{\bar{x}_g} \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\sum_{g\in U_1}\boldsymbol{\bar{x}_{g1}} - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\sum_{g\in s_1}\pi_{g1}^{-1}\boldsymbol{\bar{x}_{g1}} \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}(\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}}). \qquad (17)\end{aligned}$$

Remarquons que (17) n'inclut pas l'estimateur de $\boldsymbol{\phi_2}$. Les estimateurs par les moindres carrés

$$\hat{\boldsymbol{\phi}} = \begin{pmatrix} \hat{\boldsymbol{\phi}}_{\boldsymbol{1}} \\ \hat{\boldsymbol{\phi}}_{\boldsymbol{2}} \end{pmatrix}$$

sont la solution de :

$$\sum_{i\in s}\pi_i^{-1}c_i(y_i - \hat{\boldsymbol{\phi}}^{\boldsymbol{T}}\boldsymbol{z_i})\,\boldsymbol{z_i} = \boldsymbol{0}$$

qui est équivalente à :

$$\sum_{i\in s}\pi_i^{-1}c_i\{y_i - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g})\}\begin{pmatrix} \boldsymbol{\bar{x}_g} \\ \boldsymbol{x_i} - \boldsymbol{\bar{x}_g} \end{pmatrix} = \boldsymbol{0}.$$

Par hypothèse, $c_i = a_g N_g$ de sorte que les p premiers éléments de cette équation sont :

$$\begin{aligned}\boldsymbol{0} &= \sum_{g\in s_1}\sum_{i\in U_g}\pi_i^{-1}a_g N_g\,\boldsymbol{\bar{x}_g}\{y_i - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} - \hat{\boldsymbol{\phi}}_{\boldsymbol{2}}^{\boldsymbol{T}}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g})\} \\ \boldsymbol{0} &= \sum_{g\in s_1}\pi_{g1}^{-1}a_g N_g\,\boldsymbol{\bar{x}_g}\sum_{i\in U_g}\{y_i - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{\bar{x}_g} - \hat{\boldsymbol{\phi}}_{\boldsymbol{2}}^{\boldsymbol{T}}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_g})\} \\ \boldsymbol{0} &= \sum_{g\in s_1}\pi_{g1}^{-1}a_g\,\boldsymbol{x_{g1}}\{y_{g1} - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{x_{g1}} - \hat{\boldsymbol{\phi}}_{\boldsymbol{2}}^{\boldsymbol{T}}(\boldsymbol{x_{g1}} - \boldsymbol{x_{g1}})\} \\ \boldsymbol{0} &= \sum_{g\in s_1}\pi_{g1}^{-1}a_g\,\boldsymbol{x_{g1}}(y_{g1} - \hat{\boldsymbol{\phi}}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{x_{g1}}).\end{aligned}$$

Donc, $\hat{\boldsymbol{\phi}}_{\boldsymbol{1}}$ est une solution de (7). Par conséquent, l'estimateur GREG pour le modèle (9) est égal à $\hat{T}_H$ à condition que $c_i = a_g N_g$.

Bibliographie

Alexander, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

Cholette, P. (1984). L'ajustement des séries infra-annuelles aux répères annuels. *Techniques d'enquête*, 10, 39-53.

Clark, R.G., et Steel, D.G. (2002). The effect of using household as a sampling unit. *Revue Internationale de Statistique*, 70 (2), 289-314.

Heldal, J. (1992). A method for calibration of weights in sample surveys. Dans *Workshop on uses of auxiliary information in surveys*. University of Orebro, Suède.

Hidiroglou, M., et Patak, Z. (2004). Estimation par domaine par la régression linéaire. *Techniques d'enquête*, 30, 73-85.

Lazarfeld, P.F., et Menzel, H. (1961). On the relation between individual and collective properties. Dans *Complex Organizations*: *A Sociological Reader*. Holt, Reinhart and Winston. 422-440.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

Luery, D.M. (1986). Weighting sample survey data under linear constraints on the weights. Dans *Proceedings of the Social Statistics Section*, American Statistical Association, (Alexandria, VA), 325-330.

Nieuwenbroek, N. (1993). *An integrated method for weighting characteristics of persons and households using the linear regression estimator*. Netherlands Central Bureau of Statistics.

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

Silva, P.L.N., et Skinner, C. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.

Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 71-79
Statistique Canada, N° 12-001 au catalogue

Bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases

Hiroshi Saigo [1]

Résumé

L'échantillonnage à deux phases est un plan utile lorsque l'on ne dispose pas de variables auxiliaires a priori. L'estimation de la variance sous ce plan est toutefois compliquée, particulièrement si les fractions d'échantillonnage sont grandes. Le présent article décrit une méthode bootstrap simple pour l'échantillonnage aléatoire simple à deux phases sans remise à chaque phase avec fraction d'échantillonnage élevée. Elle est applicable à l'estimation des fonctions de répartition et des quantiles, puisqu'aucune remise à l'échelle n'est effectuée. La méthode peut être étendue à l'échantillonnage à deux phases stratifié en répétant indépendamment la procédure proposée dans diverses strates. L'estimation de la variance de certains estimateurs classiques, comme les estimateurs par le ratio et par la régression, est étudiée à titre d'exemple. Une étude par simulation est réalisée pour comparer la méthode proposée aux estimateurs de la variance existants pour l'estimation des fonctions de répartition et des quantiles.

Mots clés : Échantillonnage double; rééchantillonnage; estimation de la variance.

1. Introduction

L'échantillonnage à deux phases ou échantillonnage double est un outil puissant pour l'estimation efficace dans les sondages. Habituellement, on tire un grand échantillon de première phase où les variables auxiliaires, corrélées aux caractéristiques d'intérêt et relativement faciles à obtenir, sont observées. Puis, on sélectionne un petit sous-échantillon à partir de l'échantillon de première phase pour mesurer les caractéristiques d'intérêt qui sont plus difficiles à obtenir. À l'étape de l'estimation, les variables auxiliaires de la première phase sont utilisées pour obtenir un estimateur efficace.

Une formule explicite de la variance d'échantillon d'un estimateur peut être compliquée, voire même inexistante sous échantillonnage à deux phases. Par conséquent, les méthodes de rééchantillonnage, comme le jackknife et le bootstrap, sont séduisantes dans ces conditions. Rao et Sitter (1995) et Sitter (1997) ont étudié l'approche du jackknife avec suppression d'une unité pour les estimateurs par le ratio et par la régression sous échantillonnage à deux phases et constaté que la méthode produit des estimations de la variance convergentes par rapport au plan ayant des propriétés conditionnelles désirables sachant les variables auxiliaires.

Une faiblesse du jackknife avec suppression d'une unité est qu'il ne permet pas de traiter l'estimation des quantiles. De surcroît, l'intégration de la correction pour population finie dans l'estimation de la variance par le jackknife sous échantillonnage à deux phases n'est pas une question triviale (voir Lee et Kim 2002 et Berger et Rao 2006). Le bootstrap, par contre, élimine ces problèmes s'il est formulé convenablement.

Plusieurs méthodes bootstrap ont été proposées et étudiées pour l'échantillonnage à deux phases. Schreuder, Li et Scott (1987), Biemer et Atkinson (1993) et Sitter (1997) ont considéré des méthodes bootstrap similaires qui fournissent une estimation de la variance convergente lorsque les fractions d'échantillonnage sont négligeables. Rao et Sitter (1997) ont proposé un bootstrap avec rééchelonnage pour les fractions d'échantillonnage élevées.

Un inconvénient de l'approche de rééchelonnage est qu'elle ne permet pas de traiter l'estimation des fonctions de répartition ni des quantiles. Dans le présent article, nous proposons un bootstrap corrigé sur la moyenne pour l'échantillonnage à deux phases qui permet l'estimation des fonctions de répartition et des quantiles. La méthode est simple et englobe les méthodes existantes pour les fractions d'échantillonnage négligeables à titre de cas particuliers. Récemment, Kim, Navarro et Fuller (2006) ont étudié l'estimation de la variance par rééchantillonnage sans rééchelonnage pour l'échantillonnage à deux phases dans un cadre plus généralisé que celui du présent article. Toutefois, notre méthode diffère en ce que la correction pour population finie y est intrinsèque.

La présentation de l'article est la suivante. La section 2 décrit le bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases. La section 3 illustre le fonctionnement de la méthode proposée pour certains estimateurs classiques. La section 4 décrit l'exécution d'une simulation pour l'estimation des fonctions de répartition et des quantiles. La section 5 comprend la discussion d'autres applications du bootstrap avec moyenne ajustée. Enfin, les conclusions sont présentées à la section 6.

1. Hiroshi Saigo, Faculty of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku Tokyo 169-8050, Japon.

2. Bootstrap avec moyenne ajustée

Pour simplifier la notation, nous supposons ici qu'il n'existe qu'une seule strate. Pour étendre notre méthode à l'échantillonnage stratifié, il suffit de répéter la même procédure indépendamment dans diverses strates pour obtenir un échantillon bootstrap (voir Rao et Sitter 1997, pages 759 à 762).

Soit P l'ensemble d'étiquettes d'unité dans une population de taille N. Supposons que l'on sélectionne un échantillon aléatoire simple sans remise (EASSR) de taille n_{A+B} à partir de P et dénotons les étiquettes échantillonnées par $A+B$. La variable auxiliaire (le vecteur de variables auxiliaires) x_i est observée pour $i \in A+B$. Puis, nous tirons un EASSR de deuxième phase de taille $n_A < n_{A+B}$ à partir de $A+B$ et dénotons les étiquettes échantillonnées par A. La caractéristique (le vecteur de caractéristiques) y_i est mesurée pour $i \in A$. Soit $B = (A+B) - A$, $n_B = n_{A+B} - n_A$, $\mathbf{y}_A = \{y_i : i \in A\}$, $\mathbf{x}_A = \{x_i : i \in A\}$, et $\mathbf{x}_B = \{x_j : j \in B\}$. Nous supposons qu'un estimateur approximativement sans biais par rapport au plan du paramètre θ peut s'écrire sous la forme $\hat{\theta} = t(\mathbf{y}_A, \mathbf{x}_A, \mathbf{x}_B)$.

Sous la méthode proposée, nous construisons un échantillon bootstrap comme il suit.

1. Considérer A comme un EASSR de taille n_A tiré de P. Choisir n_A unités à partir de A par une méthode bootstrap appropriée pour un EASSR de taille n_A tiré de P. Dénoter les étiquettes échantillonnées par A^*.
2. Considérer B comme un EASSR de taille n_B tiré de $P - A$, sachant que A a été sélectionné. Choisir n_B unités à partir de B par une méthode bootstrap appropriée pour un EASSR de taille n_B tiré de $P - A$. Dénoter les étiquettes échantillonnées B^*.
3. Pour $j \in B^*$, définir l'ajustement de la moyenne comme étant $\tilde{x}_j$, où

$$\tilde{x}_j = x_j + f_A(\bar{x}_A - \bar{x}_{A^*})/(1 - f_A), \qquad (1)$$

avec $\bar{x}_A = n_A^{-1}\sum_{i\in A} x_i$, $\bar{x}_{A^*} = n_A^{-1}\sum_{i\in A^*} x_i$, et $f_A = n_A/N$.

4. Soit $\mathbf{y}_{A^*} = \{y_i : i \in A^*\}$, $\mathbf{x}_{A^*} = \{x_i : i \in A^*\}$, et $\tilde{\mathbf{x}}_{B^*} = \{\tilde{x}_j : j \in B^*\}$. L'analogue bootstrap de $\hat{\theta}$ est alors donné par $\hat{\theta}^* = t(\mathbf{y}_{A^*}, \mathbf{x}_{A^*}, \tilde{\mathbf{x}}_{B^*})$.

Pour les méthodes bootstrap applicables à une population finie, voir Shao et Tu (1995, chapitre 6). Le bootstrap de Bernoulli (BBE) proposé par Funaoka, Saigo, Sitter et Toida (2006) convient pour notre méthode, pour une raison que nous mentionnerons plus loin. Pour obtenir un échantillon bootstrap A^* dans le BBE, nous procédons à un remplacement aléatoire de chaque i dans A : garder le couple (x_i, y_i) dans l'échantillon bootstrap avec une probabilité $p = \{1 - (1 - n_A^{-1})^{-1}(1 - f_A)\}^{1/2}$ ou le remplacer par celui sélectionné aléatoirement à partir de A. Pour le cas où $p \notin [0, 1]$, voir Funaoka et coll. (2006).

Pour estimer la variance de $\hat{\theta}$, répéter les étapes 1 à 4 un grand nombre K de fois et utiliser

$$v_{\text{boot}}(\hat{\theta}) = K^{-1}\sum_{k=1}^{K}(\hat{\theta}^*_{(k)} - \hat{\theta}^*_{(\cdot)})^2, \qquad (2)$$

où $\hat{\theta}^*_{(k)}$ est la valeur de $\hat{\theta}^*$ dans le $k^{\text{ième}}$ échantillon bootstrap et $\hat{\theta}^*_{(\cdot)} = K^{-1}\sum_k \hat{\theta}^*_{(k)}$.

Quand f_A est négligeable, l'ajustement de la moyenne (1) est inutile. La méthode susmentionnée se réduit alors pour un grand n_A à celle de Schreuder et coll. (1987) et de Sitter (1997).

La méthode bootstrap proposée est motivée par les deux observations qui suivent. En premier lieu, posons que les plans d'échantillonnage I et II sont $[P \to A+B,\ A+B \to A]$ et $[P \to A,\ P - A \to B]$, respectivement, où $\to$ signifie que « le deuxième membre est un EASSR provenant du premier membre ». Alors, I et II implémentent le plan de sondage identique. En fait, la probabilité d'échantillonnage attribué à un échantillon particulier $\{\mathbf{i} = (i_1, i_2, \ldots, i_{n_A}) \in A,\ \mathbf{j} = (j_1, j_2, \ldots, j_{n_B}) \in B\}$ dans I est $\Pr\{\mathbf{i} \in A,\ \mathbf{j} \in B\} = [{}_N C_{n_{A+B}} \times {}_{n_{A+B}} C_{n_A}]^{-1} = n_A!\, n_B!(N - n_{A+B})!/N!$, tandis qu'elle est $\Pr\{\mathbf{i} \in A,\ \mathbf{j} \in B\} = [{}_N C_{n_A} \times {}_{N-n_A} C_{n_B}]^{-1} = n_A!\, n_B!(N - n_{A+B})!/N!$ dans II. De toute évidence, la distribution d'échantillonnage d'un estimateur sous échantillonnage répété dépend du plan d'échantillonnage. Donc, il est commode de supposer que II est réalisé, même si I est employé.

En deuxième lieu, pour justifier l'ajustement de la moyenne (1), observons que la moyenne de x de l'ensemble $P - A$, ou l'espérance conditionnelle de $\bar{x}_B$ sous échantillonnage répété sachant A, est $\bar{X}_{P-A} = (\bar{X} - f_A\bar{x}_A)/(1 - f_A)$. La valeur bootstrap de $\bar{X}_{P-A}$ est donnée par $\bar{X}_{P-A^*} = (\bar{X} - f_{A^*}\bar{x}_{A^*})/(1 - f_{A^*})$. Donc, l'équation (1) équivaut à $\tilde{x}_j = x_j - \bar{X}_{P-A} + \bar{X}_{P-A^*}$, un ajustement de la moyenne semblable à celui proposé par Rao et Shao (1992) dans le contexte de l'imputation hot deck sous mécanisme de réponse uniforme. Cet ajustement de la moyenne fait en sorte qu'existent les corrélations appropriées entre x dans A^* et x dans B^* nécessaires pour que l'estimation de la variance soit convergente lorsque les fractions d'échantillonnage sont élevées (voir Rao et Sitter 1997, page 760). Notons que la condition $n_A = n_{A^*}$ ou $f_A = f_{A^*}$ est essentielle à l'annulation de $\bar{X}$ dans l'ajustement de la moyenne. Par conséquent, le bootstrap avec moyenne ajustée requiert une méthode bootstrap pour l'EASSR qui retient la taille originale d'échantillon, telle que le BBE.

Nous montrons à l'annexe A que la méthode bootstrap proposée produit une estimation de la variance convergente par rapport au plan pour la classe d'estimateurs étudiés par Rao et Sitter (1997). Puisqu'aucun rééchelonnement n'est effectué, la méthode s'applique aussi à l'estimation des fonctions de répartition. Sous certaines conditions de régularité pour la fonction de répartition de population, elle produit des estimateurs de la variance convergent par rapport au plan pour les quantiles.

3. Illustrations

3.1 Estimateur par le ratio

En guise d'illustration, commençons par considérer l'estimateur par le ratio $\bar{y}_r = r_A \bar{x}_{A+B}$, où $r_A = \bar{y}_A / \bar{x}_A$, $w_A = n_A / n_{A+B}$, et $\bar{x}_{A+B} = w_A \bar{x}_A + (1 - w_A) \bar{x}_B$. Soit $\bar{y}_r^* = (\bar{y}_{A^*} / \bar{x}_{A^*})\{w_A \bar{x}_{A^*} + (1 - w_A)\bar{x}_{B^*}\}$, l'analogue bootstrap de $\bar{y}_r$. En utilisant les résultats de l'annexe A avec $h(\bar{y}_A, \bar{x}_A, \bar{x}_B) = (\bar{y}_A / \bar{x}_A)\{w_A \bar{x}_A + (1 - w_A)\bar{x}_B\}$, nous pouvons approximer la variance de $\bar{y}_r^*$ sous la méthode bootstrap proposée $V_*(\bar{y}_r^*)$ par

$$V_*(\bar{y}_r^*) \doteq (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1 - f_A)}{n_A} \hat{S}_{dA}^2 + 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)} \hat{S}_{xA}^2 + \frac{(1 - w_A)}{(1 - f_A)} \hat{S}_{xB}^2 \right], \quad (3)$$

où $\hat{S}_{dA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)^2$, $\hat{S}_{dxA} = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)(x_i - \bar{x}_A)$, $\hat{S}_{xA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^2$, et $\hat{S}_{xB}^2 = (n_B - 1)^{-1} \sum_{i \in B} (x_i - \bar{x}_B)^2$. Le deuxième membre de (3) peut être décrit comme un estimateur de la variance par « bootstrap-linéarisation ». Nous le dénotons par $v_{\mathrm{BL}}(\bar{y}_r)$. Soulignons que $v_{\mathrm{BL}}(\bar{y}_r)$ est presque identique à l'estimateur jackknife-linéarisation de la variance de Rao et Sitter (1995),

$$v_{\mathrm{JL}}(\bar{y}_r) = (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1 - f_A)}{n_A} \hat{S}_{dA}^2 + 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \hat{S}_{xA+B}^2, \quad (4)$$

où $\hat{S}_{xA+B}^2 = (n_{A+B} - 1)^{-1} \sum_{A+B} (x_i - \bar{x}_{A+B})^2$, qui concorde avec l'équation 4.8 de Demnati et Rao (2004), page 25. Puisqu'ils sont proches de $v_{\mathrm{JL}}(\bar{y}_r)$, $V_*(\bar{y}_{lr})$, son approximation de Monte Carlo $v_{\mathrm{boot}}(\bar{y}_r^*)$ et $v_{\mathrm{BL}}(\bar{y}_{lr})$ devraient donner de bons résultats non seulement inconditionnellement, mais conditionnellement à $(\bar{x}_{A+B} / \bar{x}_A)$ également. Il est intéressant de souligner que la linéarisation de Taylor dans la dérivation de $v_{\mathrm{BL}}(\bar{y}_r)$ est effectuée autour des moyennes d'échantillon et non des moyennes de population (voir le commentaire fait par Demnati et Rao 2004, page 21).

3.2 Estimateur par la régression

Nous considérons maintenant l'estimateur par la régression. L'estimateur de la moyenne de population est $\bar{y}_{lr} = \bar{y}_A + b_A(\bar{x}_{A+B} - \bar{x}_A) = \bar{y}_A + (1 - w_A) b_A(\bar{x}_B - \bar{x}_A)$, où $b_A = \hat{S}_{xyA} / \hat{S}_{xA}^2$ avec $\hat{S}_{xyA} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)(y_i - \bar{y}_A)$. Soit $\bar{y}_{lr}^* = \bar{y}_{A^*} + (1 - w_A) b_{A^*} (\bar{x}_{B^*} - \bar{x}_{A^*})$. En utilisant les résultats de l'annexe A (voir aussi l'annexe B), nous avons

$$V_*(\bar{y}_{lr}^*) \doteq \frac{(1 - f_A)}{n_A} m_{02} + \frac{(1 - f_{A+B})}{n_{A+B}} b_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)} \hat{S}_{xA}^2 + \frac{(1 - w_A)}{(1 - f_A)} \hat{S}_{xB}^2 \right] + z_A^2 \frac{(1 - f_A)}{n_A} m_{22} + 2 z_A \frac{(1 - f_A)}{n_A} m_{12} + 2 z_A \frac{(1 - f_{A+B})}{n_{A+B}} b_A m_{21} + 4 z_A^2 \frac{(1 - f_A)}{n_A} a_A b_A \bar{x}_A \hat{S}_{xA}^2, \quad (5)$$

où $z_A = n_A(\bar{x}_{A+B} - \bar{x}_A) / \{(n_A - 1) \hat{S}_{xA}^2\}$, $m_{pq} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^p e_i^q$, $e_i = y_i - \bar{y}_A - b_A(x_i - \bar{x}_A)$, et $a_A = \bar{y}_A - b_A \bar{x}_A$. Nous appelons le deuxième membre de (5) un estimateur bootstrap-linéarisation de la variance de $\bar{y}_{lr}$ et le dénotons par $v_{\mathrm{BL}}(\bar{y}_{lr})$. L'estimateur jackknife-linéarisation de la variance de $\bar{y}_{lr}$ (Sitter 1997, page 781) est

$$v_{\mathrm{JL}}(\bar{y}_{lr}) = \frac{(1 - f_A)}{n_A} m_{02} + \frac{(1 - f_{A+B})}{n_{A+B}} b_A^2 \hat{S}_{xA+B}^2 + \frac{z_A^2}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A)^2 e_i^2}{(1 - c_i)^2} + \frac{2 z_A}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A) e_i^2}{(1 - c_i)} + \frac{2 z_A b_A}{n_A (n_{A+B} - 1)} \sum_{i \in A} \frac{(x_i - \bar{x}_A)(x_i - \bar{x}_{A+B}) e_i}{(1 - c_i)}, \quad (6)$$

où $c_i = n_A^{-1} + (x_i - \bar{x}_A)^2 / \{(n_A - 1) \hat{S}_{xA}^2\}$, les valeurs d'effet de levier. Partant de (5) et (6), $v_{\mathrm{boot}}(\bar{y}_{lr})$, $v_{\mathrm{BL}}(\bar{y}_{lr})$ et $v_{\mathrm{JL}}(\bar{y}_{lr})$ donnent des résultats similaires à condition que $f_{A+B} \doteq 0$, que n_A soit suffisamment grand pour que tous les c_i soient presque nuls et que le dernier terme du deuxième membre de (5) soit négligeable.

3.3 Estimation des fonctions de répartition

À titre d'exemple, prenons l'estimateur du maximum de vraisemblance pseudo-empirique calé sur un modèle (ME)

sous échantillonnage à deux phases proposé par Wu et Luan (2003) et défini par

$$\hat{F}_{\mathrm{ME}}(t) = \sum_{i \in A} \hat{p}_i I(y_i \le t), \qquad (7)$$

où $\hat{p}_i$ maximise la fonction de pseudo-vraisemblance $\hat{l}(p) = \sum_A (N/n_A) \log p_i$ sous les contraintes a) $\sum_A p_i = 1 (0 < p_i < 1)$; et b) $\sum_A p_i g_i = n_{A+B}^{-1} \sum_{A+B} g_i$ où $g_i = g(x_i, t) = P(y \le t \mid x_i)$ sous un certain modèle de travail. Par exemple, nous pouvons supposer que $\log(g_i/(1 - g_i)) = x_i' \theta$ avec une fonction de variance $V(g) = g(1 - g)$. Chen, Sitter et Wu (2002) ont montré un algorithme simple pour le calcul de $\hat{p}_i$. Il peut être démontré (voir Wu et Luan 2003) que, sous l'échantillonnage à deux phases considéré dans le présent article,

$$\hat{F}_{\mathrm{ME}}(t) = n_A^{-1} \sum_{i \in A} I(y_i \le t) + \left\{ n_{A+B}^{-1} \sum_{i \in A+B} g_i - n_A^{-1} \sum_{i \in A} g_i \right\} \beta + o_p(n_A^{-1/2}),$$

où $\beta = \sum_P (g_i - \bar{g}) I(y \le t) / \sum_P (g_i - \bar{g})^2$ avec $\bar{g} = N^{-1} \sum_P g_i$. Notons que cette équation n'est pas utilisée dans l'estimation, mais elle montre que la variance de $\hat{F}_{\mathrm{ME}}(t)$ peut être estimée par le bootstrap avec moyenne ajustée, puisque $\hat{F}_{\mathrm{ME}}(t)$ est approximé par un estimateur de type régression.

3.4 Estimation des quantiles

L'estimation des quantiles peut être obtenue directement en inversant $\hat{F}(t)$ par $\hat{F}^{-1}(\alpha) = \inf\{t : \hat{F}(t) \ge \alpha\}$ pour un certain $\alpha \in (0, 1)$. Par exemple, si on utilise (7), alors une estimation des quantiles est donnée par $y_{(k)}$, où $y_{(k)}$ est la statistique de $k^{\text{ième}}$ ordre de y telle que $\sum_{i=1}^{k-1} \hat{p}_{(i)} < \alpha$ et $\sum_{i=1}^{k} \hat{p}_{(i)} \ge \alpha$ (Chen et Wu 2002). Sous certaines conditions spécifiées dans Chen et Wu (2002), une représentation de type Bahadur de $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ peut être établie. Donc, l'estimateur de la variance par le bootstrap avec moyenne ajustée pour $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ est convergent par rapport au plan. Notons qu'il n'existe aucune forme explicite de l'estimateur de la variance pour $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$, mais qu'on peut appliquer un estimateur de la variance convergent basé sur l'estimation d'intervalle de Woodruff (Woodruff 1952).

4. Simulation

4.1 Population et échantillonnage

Nous avons réalisé une étude par simulation afin d'examiner l'estimation de la variance par le bootstrap avec moyenne ajustée pour les estimateurs de la section 3. Nous présentons ici les résultats pour l'estimation des fonctions de répartition et des quantiles. Les résultats pour les estimateurs par le ratio et par la régression peuvent être obtenus auprès de l'auteur sur demande.

Pour commencer, nous avons généré la variable auxiliaire x pour une population finie P de taille $N = 2\,000$ en utilisant une loi Gamma(1, 1). La variable dépendante y a ensuite été générée au moyen de $y_i = x_i + \sqrt{x_i} v_i$, où $v_i \sim N(0, 0{,}5^2)$. Un EASSR $A + B$ de taille $n_{A+B} = 800$ a été sélectionné à partir de la population, puis un EASSR A de taille $n_A = 200$ a été sélectionné à partir de $A + B$. La population est demeurée fixe au cours des exécutions de la simulation, puisque nous nous concentrons sur les propriétés de l'échantillonnage répété par rapport au plan.

4.2 Estimation des fonctions de répartition

Pour l'estimation des fonctions de répartition, nous avons pris $\hat{F}_{\mathrm{ME}}(t)$ comme exemple. D'autres estimateurs, comme ceux de Chambers et Dunstan (1986) et de Rao, Kovar et Mantel (1990) peuvent être traités de la même façon quand un estimateur approximativement sans biais par rapport au plan. Nous avons supposé que le modèle de travail pour g dans $\hat{F}_{\mathrm{ME}}(t)$ était le logit avec variance binomiale. L'estimateur bootstrap de la variance $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ a été calculé avec $K = 200$. Nous avons utilisé le BBE pour construire un échantillon bootstrap. Le nombre total de simulations était $M = 5\,000$, tandis que l'EQM réelle de $\hat{F}_{\mathrm{ME}}(t)$ à un temps t donné a été estimée sur 50 000 exécutions.

Nous avons comparé $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ à trois estimateurs de la variance : l'estimateur analytique de Wu et Luan (2003), le jackknife avec suppression d'une unité standard et le jackknife avec suppression d'une unité et une correction pour population finie *ad hoc*. L'estimateur de Wu et Luan (2003) est

$$v_a(\hat{F}_{\mathrm{ME}}(t)) = (n_{A+B}^{-1} - N^{-1}) \hat{S}_I^2 + (n_A^{-1} - n_{A+B}^{-1}) \hat{S}_D^2,$$

où les deux composantes $\hat{S}^2$ sont estimées respectivement par

$$\hat{S}^2 = s^2 + \left[\frac{1}{n_{A+B}(n_{A+B} - 1)} \sum_{j>i:i, j \in A+B} u_{ij} - \frac{1}{n_A(n_A - 1)} \sum_{j>i:i, j \in A} u_{ij} \right] \hat{\beta}_F,$$

où $s^2 = \{n_A(n_A - 1)\}^{-1} \sum_{i<j:i, j \in A} v_{ij}$, et $\hat{\beta}_F = \sum_{i<j:i, j \in A} u_{ij} v_{ij} / \sum_{i<j:i, j \in A} u_{ij}^2$ avec u_{ij} et v_{ij} spécifiés comme suit : Pour $\hat{S}_I^2$, $v_{ij} = (I_i - I_j)^2$ et $u_{ij} = (\hat{g}_i - \hat{g}_j)^2$ avec $I_i = I(y_i \le t)$ et $\hat{g}_i = \hat{g}(x_i, t)$ estimé en A; pour $\hat{S}_D^2$, $v_{ij} = (\hat{D}_i - \hat{D}_j)^2$ et $u_{ij} = \hat{g}_i(1 - \hat{g}_i) + \hat{g}_j(1 - \hat{g}_j)$ avec $\hat{D}_i = I_i - \hat{g}_i \hat{\beta}$, $\hat{\beta} = \sum_{i \in A} I_i(\hat{g}_i - \bar{\hat{g}}_A) / \sum_{i \in A} (\hat{g}_i - \bar{\hat{g}}_A)^2$ et $\bar{\hat{g}}_A = n_A^{-1} \sum_{i \in A} \hat{g}_i$.

La formule du jackknife avec suppression d'une unité standard est donnée par

$$v_J(\hat{\theta}) = \frac{(n_{A+B}-1)}{n_{A+B}} \sum_{j \in A+B} (\hat{\theta}_{(-j)} - \hat{\theta}_{(\cdot)})^2,$$

où $\hat{\theta} = \hat{F}_{\mathrm{ME}}(t)$, $\hat{\theta}_{(-j)}$ est la j^{e} pseudo-estimation par le jackknife et $\hat{\theta}_{(\cdot)} = n_{A+B}^{-1} \sum_{j \in A+B} \hat{\theta}_{(-j)}$. Notons que, pour $j \in A$, y_j et x_j sont toutes deux éliminées de l'échantillon, tandis que pour $j \in B$, seul x_j est éliminée (voir Rao et Sitter 1995 et Sitter 1997). La formule avec correction pour population finie *ad hoc* est $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t)) = (1 - f_{A+B}) v_J(\hat{F}_{\mathrm{ME}}(t))$.

Le tableau 1 présente le biais relatif (%biais) et le coefficient de variation (CV) des quatre estimateurs de la variance pour $\hat{F}_{\mathrm{ME}}(t_\alpha)(\alpha = 0{,}10,\ 0{,}25,\ 0{,}50,\ 0{,}75,\ 0{,}90)$, où $F(t_\alpha) = \alpha$. Ici, %Biais et CV ont été calculés sous la forme $\%\mathrm{Biais} = 100 \times (M^{-1} \sum_{m=1}^{M} v^{(m)} - \mathrm{EQM})/\mathrm{EQM}$ et $\mathrm{CV} = [M^{-1} \sum_{m=1}^{M} (v^{(m)} - \mathrm{EQM})^2]^{1/2}/\mathrm{EQM}$, respectivement, où $v^{(m)}$ est une estimation de la variance dans la m^{e} exécution de la simulation. Le tableau 1 démontre que $v_J(\hat{F}_{\mathrm{ME}}(t))$ présente un biais par excès, puisque les fractions d'échantillonnage ne sont pas négligeables, que $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t))$ présente un biais par défaut puisque le facteur d'ajustement *ad hoc* $(1 - f_{A+B})$ est trop faible, et que $v_a(\hat{F}_{\mathrm{ME}}(t))$ et $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ sont tous deux approximativement sans biais, quoique le dernier soit un peu plus instable, ce qui est typique d'une méthode de rééchantillonnage.

Tableau 1 Estimation de la variance pour l'EMV pseudo-empirique $\hat{F}_{\mathrm{ME}}(t_\alpha)$

		α				
Estimateur		**0,10**	**0,25**	**0,50**	**0,75**	**0,90**
$v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Biais	0,27	-0,22	0,64	0,83	2,73
	CV	0,19	0,14	0,14	0,15	0,24
$v_a(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Biais	-2,29	-2,03	-0,47	-1,95	-3,26
	CV	0,17	0,11	0,09	0,11	0,19
$v_J(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Biais	14,24	17,29	22,98	23,80	24,97
	CV	0,24	0,21	0,25	0,27	0,36
$v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Biais	-31,45	-29,63	-26,21	-25,72	-25,02
	CV	0,33	0,30	0,27	0,27	0,30

En nous inspirant de Royall et Cumberland (1981a, 1981b), nous avons ordonné les $M = 5\,000$ échantillons simulés sur les valeurs de $\bar{x}_{A+B} - \bar{x}_A$, nous les avons classés en vingt groupes consécutifs de $G = 250$ dans chacun desquels l'EQM conditionnelle (EQM_c) simulée et la moyenne conditionnelle de $v(E_c(v))$ ont été calculées. La figure 1 montre la représentation graphique d' EQM_c et d' $E_c(v)$ en fonction des moyennes de groupe de $\bar{x}_{A+B} - \bar{x}_A$ pour $t_{0,10}$ et $t_{0,90}$. On constate que $v_a(\hat{F}_{\mathrm{ME}}(t))$ et $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ ont tous deux le même comportement conditionnellement à $\bar{x}_{A+B} - \bar{x}_A$. Les estimateurs jackknife de la variance, $v_J(\hat{F}_{\mathrm{ME}}(t))$ et $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t))$, quoique biaisés, suivent une tendance de l' EQM_c.

4.3 Estimation des quantiles

Par inversion directe de $\hat{F}_{\mathrm{ME}}(t)$, nous estimons le quantile α. Pour obtenir $\hat{p}_i$ pour $\hat{F}_{\mathrm{ME}}(t)$, nous avons fixé t à la valeur $\hat{t}_\alpha$, où $\hat{t}_\alpha = \inf\{t: n_A^{-1} \sum_A I(y_i \le t) \ge \alpha\}$, un estimateur utilisant uniquement $\{y_i : i \in A\}$. Pour l'estimation de la variance, nous avons créé $K = 1\,000$ échantillons bootstrap. En vue de comparaison, nous avons également calculé l'estimateur de la variance de Woodruff (Woodruff 1952 et Shao et Tu 1995, page 238),

$$v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha)) = \left[\frac{\hat{F}_{\mathrm{ME}}^{-1}(\alpha + \zeta_{1-\kappa/2}\hat{\sigma}_{\hat{F}}) - \hat{F}_{\mathrm{ME}}^{-1}(\alpha - \zeta_{1-\kappa/2}\hat{\sigma}_{\hat{F}})}{2\zeta_{1-\kappa/2}} \right]^2,$$

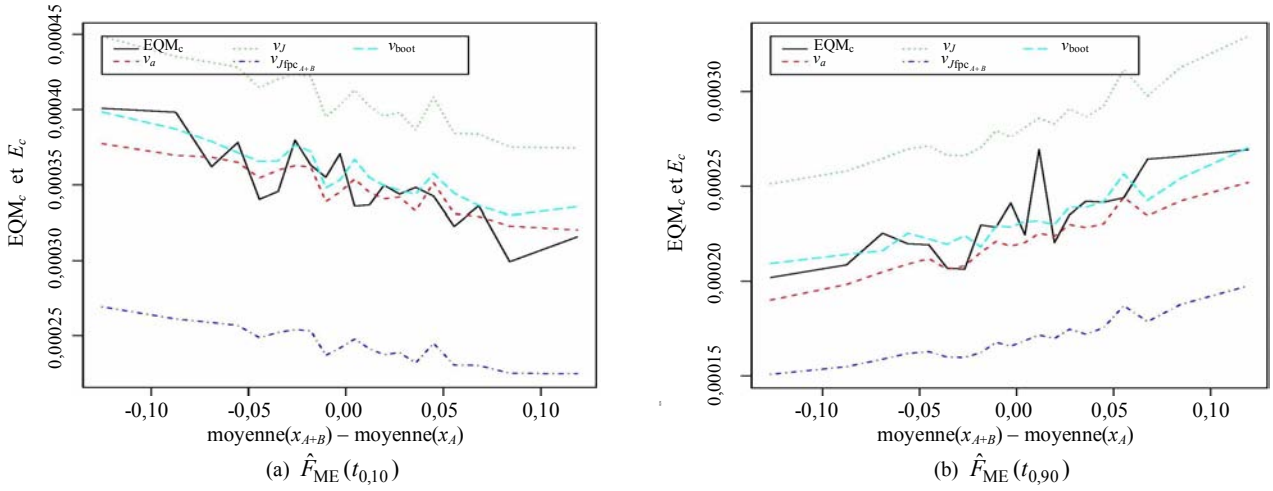
où $\hat{\sigma}_{\hat{F}}^2 = v(\hat{F}_{\mathrm{ME}}(t))$ avec $t = \hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ et $\zeta_{1-\kappa/2}$ est le $(1 - \kappa/2)$ quantile de $N(0, 1)$. Soit $\kappa = 0{,}05$, quoique le meilleur choix de κ soit inconnu. Les mesures de performance, %Biais et CV, ont été calculées sur $M = 5\,000$ exécutions, tandis que l'EQM réelle a été estimée sur 50 000 exécutions de la simulation.

Le tableau 2 résume les résultats pour l'estimation des quantiles. Il démontre que le bootstrap avec moyenne ajustée produit un biais par excès dans l'estimation de $V(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$, mais un biais négligeable dans l'estimateur de la variance de Woodruff.

Tableau 2 Estimation de la variance pour les quantiles

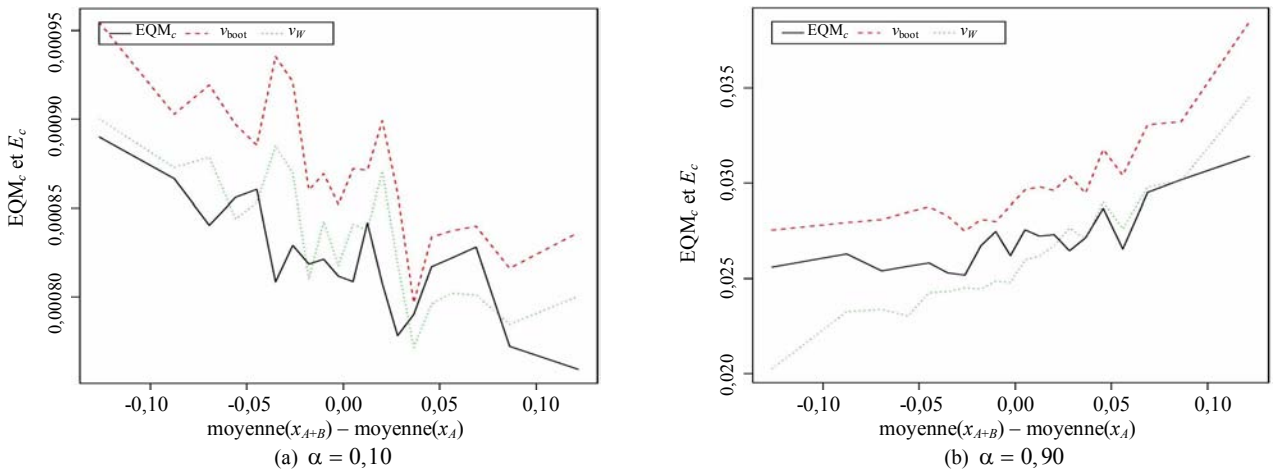
		α				
Estimateur		**0,10**	**0,25**	**0,50**	**0,75**	**0,90**
$v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$	%Biais	6,27	14,32	10,05	10,02	10,28
	CV	0,53	0,53	0,51	0,52	0,61
$v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$	%Biais	1,64	3,75	2,92	0,70	-3,67
	CV	0,50	0,45	0,45	0,46	0,52

La figure 2 montre les propriétés conditionnelles de $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ et de $v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ pour $\alpha = 0{,}10,\ 0{,}90$. Nous voyons que $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ et $v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ suivent tous deux l' EQM_c de la même façon, quoique le premier possède uniformément un biais par excès.



(a) $\hat{F}_{\mathrm{ME}}(t_{0,10})$ (b) $\hat{F}_{\mathrm{ME}}(t_{0,90})$

Figure 1 EQM$_c$ et $E_c(v)$ pour $\hat{F}_{\mathrm{ME}}(t_\alpha)$



(a) $\alpha = 0,10$ (b) $\alpha = 0,90$

Figure 2 EQM$_c$ et $E_c(v)$ pour l'estimation des quantiles

5. Remarques supplémentaires

5.1 Échantillonnage à deux phases stratifié

Supposons qu'une population doit être stratifiée en H strates, mais qu'on ne dispose d'aucune information pour la stratification. Une solution possible dans cette situation est de commencer par obtenir un EASSR de taille n' à partir de la population, d'observer les variables auxiliaires, y compris celles pour la stratification, de stratifier l'échantillon en H strates et, dans chaque strate, de tirer un EASSR de taille n_h à partir de n'_h unités appartenant à la strate h dans l'échantillon. Voir, par exemple, Cochran (1977, section 12.2) pour les détails.

Soit N_h la taille de la strate h dans la population. Sous la condition $n'_h > 0$, l'échantillonnage de première phase dans la strate h décrit plus haut est équivalent à l'échantillonnage aléatoire simple sans remise de taille n'_h dans la strate h réalisé de façon indépendante dans chacune des strates. Donc, sachant n'_h $(h = 1,...,H)$, le bootstrap avec moyenne ajustée peut être appliqué indépendamment dans diverses strates pour obtenir un échantillon bootstrap. Quand N_h est inconnu, comme cela est habituellement le cas pour l'échantillonnage à deux phases stratifié, on peut utiliser un estimateur sans biais $\hat{N}_h = N(n'_h/n')$ dans le bootstrap avec moyenne ajustée. Dans ce cas, la fraction d'échantillonnage n'/N est habituellement utilisée dans toutes les strates.

Notons toutefois que la présente discussion est légitime pour l'estimation sachant les tailles d'échantillon de première phase. La variance due à la variable n'_h peut être grande. Pour l'estimation non conditionnelle de la variance, voir Kim et coll. (2006).

5.2 Non-réponse

Le commentaire qui précède s'applique aux données d'enquête imputées sous le mécanisme de réponse uniforme. Supposons qu'une population est stratifiée en S_h $(h = 1,...,H)$ où l'échantillonnage aléatoire simple est

réalisé indépendamment. Un échantillon est divisé en classes d'imputation C_l $(l = 1,..., L)$ dans chacune desquelles on suppose que le taux de réponse est uniforme et on procède à l'imputation. Une classe d'imputation peut recouper les strates. Nous supposons aussi que la classe d'imputation à laquelle appartient une unité échantillonnée est identifiée correctement avant l'imputation. Dénotons les nombres d'unités échantillonnées et de répondants dans $S_h \cap C_l$ comme étant n_{hl} et r_{hl}, respectivement. Alors, on voit que, sachant n_{hl} et r_{hl}, le plan correspondant dans $S_h \cap C_l$ est le même que celui discuté dans le présent article si nous considérons les n_{hl} unités et les r_{hl} répondants comme étant $A + B$ et A, respectivement. Par conséquent, le bootstrap avec moyenne ajustée peut être exécuté indépendamment dans différents $S_h \cap C_l$ $(h = 1, ..., H; l = 1, ..., L)$. La taille de $S_h \cap C_l$, dénotée par N_{hl}, peut être estimée par $\hat{N}_{hl} = N_h(n_{hl}/n_h)$. Notons qu'il s'agit d'une méthode bootstrap conditionnée sur le nombre de répondants.

6. Conclusion

Dans le présent article, nous avons proposé le bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases. La méthode requiert un simple ajustement de la moyenne et permet de traiter l'estimation des fonctions de répartition et de quantiles, car elle ne nécessite pas de rééchelonnage. Le développement en série de Taylor montre que la méthode a de bonnes propriétés conditionnelles pour les estimateurs par le ratio et par la régression. Une étude par simulation démontre qu'elle a aussi des propriétés conditionnelles similaires lors de l'estimation des fonctions de répartition et des quantiles. Une extension à l'échantillonnage à deux phases stratifiées est simple. Conditionnellement aux tailles d'échantillon de première phase, la méthode permet de traiter l'échantillonnage à deux phases stratifié et l'imputation sous le mécanisme de réponse uniforme. Nous étudions à l'heure actuelle une extension de la méthode proposée à des plans d'échantillonnage multiphases plus généralisés.

Remerciements

Cette étude a été financée par une bourse de la Société japonaise de promotion de la science. L'auteur remercie le professeur Randy R. Sitter, le rédacteur en chef, le rédacteur adjoint et deux examinateurs de leurs commentaires et suggestions utiles.

Annexe A

Dans la présente annexe, nous montrons que la méthode bootstrap proposée fournit des estimations de la variance convergentes pour une classe d'estimateurs considérés par Rao et Sitter (1997). Nous utilisons les mêmes conditions que dans Rao et Sitter (1997) avec une notation légèrement différente. Pour simplifier, nous supposons qu'il n'existe qu'une seule strate, mais une extension à l'échantillonnage à deux phases stratifié est simple.

Considérons une classe d'estimateurs, $\theta = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$, d'un paramètre de population $\theta = h(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$, où $\bar{\boldsymbol{Y}}$ et $\bar{\boldsymbol{X}}$ sont les moyennes de population des vecteurs $\boldsymbol{y}$ et $\boldsymbol{x}$, c'est-à-dire $\bar{\boldsymbol{Y}} = N^{-1}\sum_{i\in P}\boldsymbol{y}_i$ et $\bar{\boldsymbol{X}} = N^{-1}\sum_{i\in P}\boldsymbol{x}_i$. Ici, $\boldsymbol{x}$ est observé dans l'échantillon de première phase $A + B$, tandis que $\boldsymbol{y}$ est mesuré uniquement dans l'échantillon de deuxième phase A. Les moyennes d'échantillon $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A)$ et $\bar{\boldsymbol{x}}_B$ sont calculées dans A et B, respectivement, c'est-à-dire $\bar{\boldsymbol{y}}_A = n_A^{-1}\sum_{i\in A}\boldsymbol{y}_i$, $\bar{\boldsymbol{x}}_A = n_A^{-1}\sum_{i\in A}\boldsymbol{x}_i$, et $\bar{\boldsymbol{x}}_B = n_B^{-1}\sum_{i\in B}\boldsymbol{x}_i$.

Par un développement en série de Taylor, nous obtenons

$$\hat{\theta} = \theta + \nabla h'(\Delta\bar{\boldsymbol{y}}_A', \Delta\bar{\boldsymbol{x}}_A', \Delta\bar{\boldsymbol{x}}_B')' + o_p(n_A^{-1/2}),$$

où ∇h est le vecteur de gradients de h évalué à $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$, $\Delta\bar{\boldsymbol{y}}_A = \bar{\boldsymbol{y}}_A - \bar{\boldsymbol{Y}}$, $\Delta\bar{\boldsymbol{x}}_A = \bar{\boldsymbol{x}}_A - \bar{\boldsymbol{X}}$, $\Delta\bar{\boldsymbol{x}}_B = \bar{\boldsymbol{x}}_B - \bar{\boldsymbol{X}}$, et $'$ dénote une matrice transposée (voir l'équation 33.7 de Rao et Sitter 1997, page 757 et les conditions requises). Alors, la variance de $\hat{\theta} = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$ est approximée par

$$V(\hat{\theta}) \doteq \nabla h' \sum_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'} \nabla h,$$

où $\sum_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'}$ est la matrice de variance-covariance de $(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A', \bar{\boldsymbol{x}}_B')'$ sous échantillonnage à deux phases répété. Comme A et B sont des EASSR de taille n_A et n_B tirés de la population P, respectivement, nous voyons que $\sum_{(\bar{\boldsymbol{y}}_A', \bar{\boldsymbol{x}}_A')'} = (1 - f_A)S^2_{(\boldsymbol{y}', \boldsymbol{x}')'}/n_A$ et $\sum_{\bar{\boldsymbol{x}}_B} = (1 - f_B)S^2_{\boldsymbol{x}}/n_B$, où $S^2_{\boldsymbol{u}} = (N-1)^{-1}\sum_{i\in P}(\boldsymbol{u}_i - \bar{\boldsymbol{U}})(\boldsymbol{u}_i - \bar{\boldsymbol{U}})'$ est la variance de population de $\boldsymbol{u} = (\boldsymbol{y}', \boldsymbol{x}')'$ ou $\boldsymbol{x}$ et $f_B = n_B/N$. Pour $\mathrm{Cov}(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_B)$, soit E_A et $E_{B|A}$ les espérances pour la sélection d'un EASSR A à partir de P et pour le choix d'un EASSR B à partir de $P - A$ sachant A, respectivement. Notons que $E_{B|A}(\boldsymbol{x}_B) = (\bar{\boldsymbol{X}} - f_A\bar{\boldsymbol{x}}_A)/(1 - f_A)$. Donc, nous avons

$$\begin{aligned}\mathrm{Cov}(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_B) &= E(\bar{\boldsymbol{y}}_A\bar{\boldsymbol{x}}_B') - E(\bar{\boldsymbol{y}}_A)E(\bar{\boldsymbol{x}}_B') \\ &= E_A(\bar{\boldsymbol{y}}_A E_{B|A}(\bar{\boldsymbol{x}}_B)) - \bar{\boldsymbol{Y}}\bar{\boldsymbol{X}}' \\ &= -S_{\boldsymbol{yx}}/N,\end{aligned}$$

où $S_{\boldsymbol{yx}} = (N-1)^{-1}\sum_{i\in P}(\boldsymbol{y} - \bar{\boldsymbol{Y}})(\boldsymbol{x} - \bar{\boldsymbol{X}})'$. De même, $\mathrm{Cov}(\bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B) = -S^2_{\boldsymbol{x}}/N$.

Maintenant, considérons un développement en série de Taylor de $\hat{\theta}^* = h(\bar{\boldsymbol{y}}_{A^*}, \bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$ avec $\bar{\tilde{\boldsymbol{x}}}_{B^*} = \bar{\boldsymbol{x}}_{B^*} + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)$, l'analogue bootstrap de $\hat{\theta} = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$. Soit E_* et V_* l'espérance et la variance sous la procédure bootstrap proposée, respectivement. Pour commencer, observons que $E_*(\bar{\boldsymbol{y}}_{A^*}) = \bar{\boldsymbol{y}}_A$, $E_*(\bar{\boldsymbol{x}}_{A^*}) = \bar{\boldsymbol{x}}_A$ et

$$\begin{aligned} E_*(\bar{\tilde{\boldsymbol{x}}}_{B^*}) &= E_{*A^*}(E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*})) \\ &= E_{*A^*}(\bar{\boldsymbol{x}}_B + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)) \\ &= \bar{\boldsymbol{x}}_B, \end{aligned}$$

où E_{*A^*} et $E_{*B^*|A^*}$ sont, respectivement, l'espérance par rapport à l'échantillonnage A^* et l'espérance conditionnelle par rapport à l'échantillonnage B^* sachant A^* sous la méthode bootstrap proposée. Alors, $\hat{\theta}^* = h(\bar{\boldsymbol{y}}_{A^*}, \bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$ est approximé par

$$\hat{\theta}^* = \hat{\theta} + \nabla h^{*'}(\Delta\bar{\boldsymbol{y}}'_{A^*}, \Delta\bar{\boldsymbol{x}}'_{A^*}, \Delta\bar{\tilde{\boldsymbol{x}}}'_{B^*})' + o_p(n_A^{-1/2}),$$

où ∇h^* est le gradient de h évalué à $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$, $\Delta\bar{\boldsymbol{y}}_{A^*} = \bar{\boldsymbol{y}}_{A^*} - \bar{\boldsymbol{y}}_A$, $\Delta\bar{\boldsymbol{x}}_{A^*} = \bar{\boldsymbol{x}}_{A^*} - \bar{\boldsymbol{x}}_A$, et $\Delta\bar{\tilde{\boldsymbol{x}}}_{B^*} = \bar{\tilde{\boldsymbol{x}}}_{B^*} - \bar{\boldsymbol{x}}_B$ (voir l'équation 33.A.1 de Rao et Sitter 1997, page 767 et les conditions requises connexes). Par conséquent, $V_*(\hat{\theta}^*)$ est approximé par

$$V_*(\hat{\theta}^*) \doteq \nabla h^{*'} \sum\nolimits^*_{(\bar{\boldsymbol{y}}'_{A^*}, \bar{\boldsymbol{x}}'_{A^*}, \bar{\tilde{\boldsymbol{x}}}'_{B^*})'} \nabla h^*,$$

où $\sum^*_{(\bar{\boldsymbol{y}}'_{A^*}, \bar{\boldsymbol{x}}'_{A^*}, \bar{\tilde{\boldsymbol{x}}}'_{B^*})'}$ est la matrice de variance-covariance de $(\bar{\boldsymbol{y}}_{A^*}, \bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})'$ sous l'échantillonnage bootstrap proposé.

L'estimation convergente de la variance sous la méthode proposée est prouvée en montrant que ∇h^* et $\sum^*_{(\bar{\boldsymbol{y}}'_{A^*}, \bar{\boldsymbol{x}}'_{A^*}, \bar{\tilde{\boldsymbol{x}}}'_{B^*})'}$ sont convergents pour ∇h et $\sum_{(\bar{\boldsymbol{y}}'_A, \bar{\boldsymbol{x}}'_A, \bar{\boldsymbol{x}}'_B)'}$, respectivement. La convergence de ∇h^* pour ∇h découle de la convergence de $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)'$ pour $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$ et de la continuité de h.

La convergence de $\sum^*_{(\bar{\boldsymbol{y}}'_{A^*}, \bar{\boldsymbol{x}}'_{A^*}, \bar{\tilde{\boldsymbol{x}}}'_{B^*})'}$ peut être montrée comme il suit. Pour commencer, puisque nous utilisons une méthode bootstrap appropriée pour l'échantillonnage aléatoire simple sans remise dans le sous-échantillonnage A^*, nous avons $\sum^*_{(\bar{\boldsymbol{y}}'_{A^*}, \bar{\boldsymbol{x}}'_{A^*})'} = (1-f_A)\hat{S}^2_{(\boldsymbol{y}',\boldsymbol{x}')'A}/n_A$, où $\hat{S}^2_{\boldsymbol{u}A} = (n_A-1)^{-1}\sum_{i\in A}(\boldsymbol{u}_i - \bar{\boldsymbol{u}}_A)(\boldsymbol{u}_i - \bar{\boldsymbol{u}}_A)'$ avec $\boldsymbol{u} = (\boldsymbol{y}', \boldsymbol{x}')'$. Deuxièmement, parce que

1. $\sum_{\bar{\tilde{\boldsymbol{x}}}_{B^*}} = E_{*A^*}(V_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*})) + V_{*A^*}(E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*}))$, où V_{*A^*} et $V_{*B^*|A^*}$ sont, respectivement, la variance par rapport à l'échantillonnage A^* et la variance conditionnelle par rapport à l'échantillonnage B^* sachant A^*,
2. $V_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*}) = (1-f_{B|A})\hat{S}^2_{\boldsymbol{x}B}/n_B$, où $\hat{S}^2_{\boldsymbol{x}B} = (n_B-1)^{-1}\sum_{i\in B}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_B)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_B)'$ et $f_{B|A} = n_B/(N-n_A)$, et
3. $E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*}) = \bar{\boldsymbol{x}}_B + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)$, nous avons $\sum^*_{\bar{\tilde{\boldsymbol{x}}}_{B^*}} = (1-f_{B|A})\hat{S}^2_{\boldsymbol{x}B}/n_B + f_A\hat{S}^2_{\boldsymbol{x}A}/(N-n_A)$. Puisque $\hat{S}^2_{\boldsymbol{x}A}$ et $\hat{S}^2_{\boldsymbol{x}B}$ sont tous deux convergents pour $S^2_{\boldsymbol{x}}$, $\sum^*_{\bar{\tilde{\boldsymbol{x}}}_{B^*}}$ est convergent pour $\sum_{\bar{\boldsymbol{x}}_B} = (1-f_B)S^2_{\boldsymbol{x}}/n_B$. Enfin, nous calculons $\mathrm{Cov}_*(\bar{\boldsymbol{y}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$ et $\mathrm{Cov}_*(\bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*})$. Pour la première, nous avons

$$\begin{aligned} \mathrm{Cov}_*(\bar{\boldsymbol{y}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*}) &= E_*(\bar{\boldsymbol{y}}_{A^*}\bar{\tilde{\boldsymbol{x}}}'_{B^*}) - E_*(\bar{\boldsymbol{y}}_{A^*})\,E_*(\bar{\tilde{\boldsymbol{x}}}_{B^*})' \\ &= E_{*A^*}(\bar{\boldsymbol{y}}_{A^*}E_{*B^*|A^*}(\bar{\tilde{\boldsymbol{x}}}_{B^*})) - \bar{\boldsymbol{y}}_A\bar{\boldsymbol{x}}'_B \\ &= E_{*A^*}(\bar{\boldsymbol{y}}_{A^*}\{\bar{\boldsymbol{x}}_B + f_A(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_{A^*})/(1-f_A)\}) - \bar{\boldsymbol{y}}_A\bar{\boldsymbol{x}}'_B \\ &= -\hat{S}_{\boldsymbol{y}\boldsymbol{x}A}/N, \end{aligned}$$

où $\hat{S}_{\boldsymbol{y}\boldsymbol{x}A} = (n_A-1)^{-1}\sum_{i\in A}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_A)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_A)'$. De même, $\mathrm{Cov}_*(\bar{\boldsymbol{x}}_{A^*}, \bar{\tilde{\boldsymbol{x}}}_{B^*}) = -\hat{S}^2_{\boldsymbol{x}A}/N$. Ceci complète la preuve de la convergence de $\sum^*_{(\bar{\boldsymbol{y}}'_{A^*}, \bar{\boldsymbol{x}}'_{A^*}, \bar{\tilde{\boldsymbol{x}}}'_{B^*})'}$ pour $\sum_{(\bar{\boldsymbol{y}}'_A, \bar{\boldsymbol{x}}'_A, \bar{\boldsymbol{x}}'_B)'}$.

Annexe B

Dans cette annexe nous dérivons $v_{\mathrm{BL}}(\bar{y}_{lr})$. Sous le bootstrap avec moyenne ajustée,

$$\bar{y}^*_{lr} = \bar{y}_{A^*} + (1-w_A)b_{A^*}\left\{-\frac{(\bar{x}_{A^*} - \bar{x}_A)}{(1-f_A)} + (\bar{x}_{B^*} - \bar{x}_B) + (\bar{x}_B - \bar{x}_A)\right\}.$$

Définissons

$$\hat{\xi}^*_{pq} = n_A^{-1}\sum\nolimits_{i\in A^*} x_i^p\, y_i^q,$$
$$\hat{\xi}^* = [\hat{\xi}^*_{10}, \hat{\xi}^*_{01}, \hat{\xi}^*_{11}, \hat{\xi}^*_{20}, \bar{x}_{B^*}]'$$

et

$$\xi = \left[\bar{x}_A, \bar{y}_A, n_A^{-1}\sum\nolimits_{i\in A} x_i\, y_i, n_A^{-1}\sum\nolimits_{i\in A} x_i^2, \bar{x}_B\right]' = E_*(\hat{\xi}^*).$$

Notons que $b_{A^*} = (\hat{\xi}^*_{11} - \hat{\xi}^*_{10}\hat{\xi}^*_{01})/(\hat{\xi}^*_{20} - \hat{\xi}^{*2}_{10})$. Soit $\bar{y}^*_{lr} = h(\hat{\xi}^*)$. Cette expression est légèrement différente de celle de l'annexe A, mais nous pouvons exploiter le sous-échantillonnage indépendant de A^* et B^*. Alors, par linéarisation de Taylor de $\bar{y}^*_{lr} = h(\hat{\xi}^*)$ autour de ξ, nous obtenons $\bar{y}^*_{lr} \doteq \bar{y}_{lr} + \nabla h^{*'}(\hat{\xi}^* - \xi)$ et $V_*(\bar{y}^*_{lr}) \doteq \nabla h^{*'}\sum^*_{\hat{\xi}^*}\nabla h^*$, où

$$\nabla h^* = [-b_A(1-w_A)/(1-f_A) - z_A(\bar{y}_A - 2b_A\bar{x}_A), 1 - z_A\bar{x}_A, z_A, -z_A b_A, b_A(1-w_A)]'$$

et $\sum^*_{\hat{\xi}^*} = [v_{ij}]$ avec

$$\begin{aligned} v_{11} &= c_A\hat{S}^2_{xA}, \\ v_{21} &= c_A\hat{S}_{xyA}, \\ v_{22} &= c_A\hat{S}^2_{yA}, \\ v_{31} &= c_A(n_A-1)^{-1}\sum\nolimits_{i\in A}(x_i\,y_i - \xi_{11})(x_i - \bar{x}_A), \\ v_{32} &= c_A(n_A-1)^{-1}\sum\nolimits_{i\in A}(x_i\,y_i - \xi_{11})(y_i - \bar{y}_A), \\ v_{33} &= c_A(n_A-1)^{-1}\sum\nolimits_{i\in A}(x_i\,y_i - \xi_{11})^2, \\ v_{41} &= c_A(n_A-1)^{-1}\sum\nolimits_{i\in A}(x_i^2 - \xi_{20})(x_i - \bar{x}_A), \\ v_{42} &= c_A(n_A-1)^{-1}\sum\nolimits_{i\in A}(x_i^2 - \xi_{20})(y_i - \bar{y}_A), \end{aligned}$$

$$v_{43} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i y_i - \xi_{11}),$$

$$v_{44} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})^2,$$

$$v_{51} = v_{52} = v_{53} = v_{54} = 0,$$

$$v_{55} = \{n_B^{-1} - (N - n_A)^{-1}\} \hat{S}_{xB}^2,$$

$v_{ij} = v_{ji}$, et $c_A = (1 - f_A)/n_A$. En réécrivant les moments par rapport à l'origine sous forme de moments centrés, en notant que $y_i - \bar{y}_A = b_A(x_i - \bar{x}_A) + e_i$ et en utilisant les propriétés de e_i en tant que résidus des moindres carrés, nous obtenons le deuxième membre de (5) après certains calculs algébriques.

Bibliographie

Berger, Y.G., et Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society*, B, 68, 531-547.

Biemer, P.P., et Atkinson, D. (1993). Estimation de l'erreur systématique de mesure par la prédiction modéliste. *Techniques d'enquête*, 19, 137-146.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.

Chen, J., et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Cochran, W.G. (1977). *Sampling Techniques*. 3ième Édition. New York : John Wiley & Sons, Inc.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

Funaoka, F., Saigo, H., Sitter, R.R. et Toida, T. (2006). Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*. 32, 169-175.

Kim, J.-K., Navarro, A. et Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Lee, H., et Kim, J.-K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. *Proceedings of ASA Section on Survey Research Methods*, 2024-2028.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data. *Biometrika*, 77, 365-375.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K., et Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. Dans *Survey Measurement and Process Quality*: *Wiley Series in Probability and Statistics*. (Éds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York. 753-768.

Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Royall, R.M., et Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Schreuder, H.T., Li, H.G. et Scott, C.T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33, 676-689.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag : New York.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Woodruff, R.S. (1952). Confidence intervals for median and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., et Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19, 119-131.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 81-92
Statistique Canada, N° 12-001 au catalogue

De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle

Nicholas Tibor Longford [1]

Résumé

Nous dérivons un estimateur de l'erreur quadratique moyenne (EQM) de l'estimateur de Bayes empirique et composite de la moyenne locale dans les conditions standard de petits domaines. L'estimateur de l'EQM est un composite de l'estimateur établi, fondé sur l'espérance conditionnelle de l'écart aléatoire associé au domaine, et d'un estimateur naïf de l'EQM fondé sur le plan de sondage. Nous évaluons ses propriétés par simulation. Enfin, nous examinons des variantes de cet estimateur de l'EQM et décrivons certaines extensions.

Mots clés : Estimation composite; estimation empirique bayésienne; rétrécissement; estimation pour petits domaines.

1. Introduction

Au fil des ans, les méthodes fondées sur le plan de sondage se sont avérées inefficaces pour l'estimation pour petits domaines, parce que, contrairement aux méthodes empiriques bayésiennes et connexes, elles ne permettent pas d'utiliser efficacement l'information auxiliaire. Cependant, les hypothèses associées aux modèles utilisés demeurent une faiblesse des méthodes fondées sur un modèle, parce que les inférences qui en découlent ont le défaut généralisé de dépendre de la validité du modèle. Dans l'application des modèles empiriques bayésiens à l'estimation pour petits domaines, les zones locales (districts) sont associées à des effets aléatoires. Sous l'approche fondée sur le plan de sondage, cette hypothèse n'est pas valide, car, lors d'une répétition hypothétique de l'enquête, les mêmes districts seraient réalisés (à l'exception de certains qui ne sont pas représentés dans l'échantillon tiré), et les grandeurs cibles associées à ces districts seraient également les mêmes. Autrement dit, les districts devraient être associés à des effets fixes. Le manque de validité de cet aspect des modèles empiriques bayésiens n'a aucun effet indésirable sur l'estimation des grandeurs pour petits domaines (moyennes, totaux, proportions, *etc.*). L'association des petits domaines à des effets aléatoires est un élément essentiel à l'emprunt d'information aux autres domaines ou à l'exploitation des similarités entre les domaines, ainsi qu'entre les variables, les points dans le temps, les enquêtes et d'autres sources de données, mais elle fausse l'évaluation de la précision des estimateurs. Certains estimateurs composites et les estimateurs de leur erreur quadratique moyenne ont le même défaut.

À la section suivante, nous diagnostiquons le problème en détail et à la section 3, nous proposons une solution, que nous illustrons et évaluons ensuite à la section 4 par simulation en utilisant une série d'exemples. Ceux-ci varient du plus simple et favorable (en accord avec la plupart des hypothèses formulées) au plus complexe et le moins favorable, afin d'explorer la robustesse de la méthode. Nous discutons de son potentiel de manière plus complète à la dernière section.

2. Fixe et aléatoire

Par variance d'échantillonnage d'un estimateur général $\hat{\theta}$ fondé sur un processus de génération de données (échantillonnage) donné χ, nous entendons la variation des valeurs de $\hat{\theta}(\mathbf{X})$ dans les répétitions des processus qui génèrent les ensembles de données $\mathbf{X}$ et qui leur appliquent $\hat{\theta}$. Dans l'approche fondée sur le plan de sondage, la répétition d'une enquête à l'échelle d'un pays et sa division en D districts produit les mêmes quantités de population au niveau du district $\theta_d, d = 1, ..., D$; ces D quantités sont *fixes*. En revanche, dans l'approche fondée sur un modèle, chaque répétition en utilisant un modèle empirique bayésien débute par la génération d'un nouvel ensemble de D valeurs de θ_d, indépendamment des répétitions antérieures.

Nous considérons l'approche fondée sur le plan comme appropriée, parce qu'en principe, chaque quantité θ_d pourrait être établie avec précision et qu'une répétition hypothétique de l'enquête correspondrait au tirage d'un échantillon à partir de la même population, avec la même division du pays en ses districts et les mêmes valeurs des variables enregistrées pour chaque membre de la population. La plupart des méthodes fondées sur le plan de sondage établies sont valides quand l'enquête porte sur une base de sondage parfaite, qui ne contient aucun enregistrement en double et s'applique exclusivement à la population étudiée, et que le plan d'échantillonnage est exécuté parfaitement, sans aucun écart par rapport au protocole établi. Autrement dit, les estimateurs qu'elles produisent sont (approximativement) sans biais, les expressions pour les variances d'échantillonnage sont correctes, ou le sont quasiment, et ces variances sont estimées avec un biais faible ou nul.

1. Nicholas Tibor Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Espagne. Courriel : NTL@SNTL.co.uk.

Par contre, les méthodes fondées sur un modèle sont assorties d'un beaucoup plus grand nombre d'hypothèses qui, souvent, ne peuvent être vérifiées. Diverses méthodes de diagnostic en vue d'évaluer la qualité de la modélisation sont disponibles, mais elles sont toutes teintées d'une incertitude. Interpréter la non-découverte d'une contradiction comme une preuve de l'absence de toute contradiction est une incohérence logique commise fréquemment. Elle ne peut être évitée qu'en mentionnant les propriétés des estimateurs quand les hypothèses ne sont pas valides, mais les méthodes de ce genre sont difficiles à élaborer, à cause de la vaste gamme de violations du modèle dont il faudrait tenir compte. Pourtant, malgré ces inconvénients, il s'est avéré que les méthodes d'estimation pour petits domaines fondées sur un modèle ont du mérite et sont aujourd'hui considérées, à juste titre, comme indispensables (Ghosh et Rao 1994; Rao 2003; Longford 2005).

Le projet EURAREA (EURAREA Consortium 2004) a mené une étude par simulation à grande échelle comportant l'échantillonnage de populations artificiellement générées ressemblant aux populations humaines de plusieurs pays européens et l'application de plusieurs classes d'estimateurs. L'étude a confirmé la supériorité des estimateurs fondés sur un modèle, avec plusieurs réserves, mais a produit des résultats plutôt décevants en ce qui concerne les estimateurs de leurs erreurs-types. Nous imputons le problème à l'application d'un calcul de moyenne dans la dérivation des erreurs-types des estimateurs de rétrécissement.

Supposons qu'une population est divisée en D districts, chacun ayant une taille de population pouvant, à toutes fins utiles, être considérée infinie, et que des plans d'échantillonnage aléatoire simple indépendants sont appliqués dans les districts. Nous supposons que, dans chaque district d, la variable de résultat Y suit une loi normale de moyenne μ_d et de même variance σ_W^2, $N(\mu_d, \sigma_W^2)$. Pour les moyennes de population intra-district μ_d, nous supposons le modèle de population $\mu_d \sim N(\mu^*, \sigma_{B^*}^2)$, mais nous voulons faire des inférences au sujet d'un ensemble fixe de moyennes (réalisées) $\{\mu_d\}$. À la section 5, nous discutons des conditions de régression plus générales définies par les modèles intra-district

$$(Y|d) \sim N(\mathbf{X}_d\boldsymbol{\beta} + \delta_d, \sigma_W^2),$$

dans lesquels $\mathbf{X}_d$ sont les matrices de régression intra-district, $\boldsymbol{\beta}$ est l'ensemble de paramètres de régression correspondants communs à tous les districts, et δ_d est l'écart de la régression intra-district par rapport à la régression typique définie par $\delta_d = 0$. Dans la superpopulation, δ_d représentent un échantillon aléatoire tiré d'une loi $N(0, \sigma_{B^*}^2)$, mais nous voulons faire des inférences au sujet de l'ensemble fixe (réalisé) $\{\delta_d\}$. Donc, nous utilisons des estimateurs fondés sur un modèle, mais nous évaluons leurs propriétés en fonction de critères fondés sur le plan.

Dénotons par μ la moyenne (nationale) des grandeurs μ_d et par σ_B^2 la variance entre les districts, $\sigma_B^2 = D^{-1}\sum_d (\mu_d - \mu)^2$. Notons que ces paramètres diffèrent de leurs équivalents en superpopulation μ^* et $\sigma_{B^*}^2$. Nous supposons pour commencer que σ_B^2, σ_W^2 et μ sont connues. Soit $\hat{\mu}_d$ et $\hat{\mu}$ les moyennes d'échantillon de la variable d'intérêt dans le district d et dans le domaine complet (pays). Elles sont fondées sur des échantillons de tailles n_d et $n = n_1 + \cdots + n_D$. Si l'on n'utilise aucune covariable, l'estimateur empirique bayésien (de rétrécissement) de μ_d est

$$\tilde{\mu}_d = \left(1 - \frac{1}{1 + n_d\omega}\right)\hat{\mu}_d + \frac{1}{1 + n_d\omega}\hat{\mu}, \qquad (1)$$

où $\omega = \sigma_B^2/\sigma_W^2$ est le ratio des variances. La variance conditionnelle fondée sur le modèle de μ_d, sachant les données, μ, σ_W^2 et σ_B^2, égale à $\sigma_B^2/(1 + n_d\omega)$, est souvent considérée comme la variance d'échantillonnage de $\tilde{\mu}_d$; les origines de cette pratique remontent à l'application de l'algorithme EM. Une dérivation plus minutieuse tient compte du fait que, dans l'approche fondée sur le plan de sondage, $\tilde{\mu}_d$ est biaisé pour μ_d,

$$\mathrm{E}(\tilde{\mu}_d|\mu_d) - \mu_d = -\frac{\mu_d - \mu}{1 + n_d\omega},$$

et que sont erreur quadratique moyenne est

$$\begin{aligned}\mathrm{EQM}(\tilde{\mu}_d; \mu_d) &= \left(1 - \frac{1}{1 + n_d\omega}\right)^2 \mathrm{var}(\hat{\mu}_d) + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2} \\ &= \sigma_W^2 \frac{n_d\omega^2}{(1 + n_d\omega)^2} + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2}, \qquad (2)\end{aligned}$$

en supposant, pour simplifier, que $\hat{\mu} \equiv \mu$. Afin de souligner que l'EQM dépend de la cible, nous incluons l'estimateur ainsi que la cible dans son argument. En particulier, $\mathrm{EQM}(\hat{\mu}; \mu) \neq \mathrm{EQM}(\hat{\mu}; \mu_d)$, à moins que $\mu_d = \mu$. Une caractéristique gênante de l'identité (2) est qu'elle contient μ_d, la cible de l'estimation. Si nous remplaçons $(\mu_d - \mu)^2$ par son espérance sur l'ensemble des districts, σ_B^2, nous obtenons l'identité plus connue

$$\overline{\mathrm{EQM}}(\hat{\mu}_d; \mu_d) = \frac{\sigma_B^2}{1 + n_d\omega}, \qquad (3)$$

la variance fondée sur un modèle conditionnel reliée à EM de μ_d. La barre au-dessus de EQM indique l'espérance (moyenne) de $(\mu_d - \mu)^2$, le numérateur du dernier terme de (2), sur l'ensemble des districts, en gardant les tailles d'échantillon n_d intactes. Tout au long de l'exposé, nous conditionnons sur les tailles d'échantillon intra-district $n_d, d = 1, \ldots, D$, même si, dans le plan d'échantillonnage,

chacune d'elle peut varier. $\overline{\text{EQM}}$ peut être interprétée comme l'espérance du modèle, quoique l'espérance ou la moyenne des écarts quadratiques $(\mu_d - \mu)^2$ puisse être considérée et estimée pour un ensemble donné de districts sans aucune référence à un modèle. Dans (3), la variance conditionnelle est appropriée pour les districts pour lesquels μ_d est dans la distance « typique », σ_B, par rapport à la moyenne nationale μ. Si $|\mu_d - \mu| \neq \sigma_B$, un estimateur sans biais de la variance conditionnelle $\sigma_B^2/(1+n_d\omega)$ est biaisé pour $\text{EQM}(\tilde{\mu}_d; \mu_d)$. Comme le biais est relié à la grandeur de population $\mu_d - \mu$, il n'est pas réduit par l'accroissement de la taille d'échantillon n_d.

3. Estimations composites de l'EQM

Pour estimer $\text{EQM}(\tilde{\mu}_d; \mu_d)$, nous réutilisons l'idée du rétrécissement et combinons deux estimateurs possibles, c'est-à-dire $\sigma_B^2/(1+n_d\omega)$ et un estimateur naïf de l'EQM donné par (2). Cet estimateur composite peut être justifié comme il suit. Si $n_d = 0$, et par conséquent $\tilde{\mu}_d = \hat{\mu}$, nous ne possédons aucune information directe au sujet de μ_d, de sorte que nous ne pouvons pas améliorer $\sigma_B^2/(1+n_d\omega)$ en tant qu'estimateur de $\text{EQM}(\tilde{\mu}_d; \mu_d)$. Quand n_d est grand, μ_d est estimée avec une précision suffisante pour utiliser $(\tilde{\mu}_d - \hat{\mu})^2$, éventuellement avec une correction du biais, comme estimateur de $(\mu_d - \mu)^2$. Pour les tailles d'échantillons intermédiaires, nous recherchons une combinaison (compromis) de ces deux alternatives qui sont appropriées dans les conditions extrêmes, c'est-à-dire quand $n_d = 0$ et quand $n_d \to +\infty$. Par conséquent, nous dérivons des expressions pour leurs EQM, puis pour l'EQM de leur combinaison.

Nous considérons la constante $\sigma_B^2/(1+n_d\omega)$ comme un estimateur et la nommons estimateur moyenné de l'EQM. Bien qu'il n'ait pas de variance, cet estimateur est biaisé, et son erreur quadratique moyenne est

$$\begin{aligned}
&\text{EQM}\left\{\frac{\sigma_B^2}{1+n_d\omega}; \text{EQM}(\tilde{\mu}_d; \mu_d)\right\} \\
&= \left\{\frac{\sigma_B^2}{1+n_d\omega} - \frac{\sigma_W^2 n_d \omega^2}{(1+n_d\omega)^2} - \frac{(\mu_d - \mu)^2}{(1+n_d\omega)^2}\right\}^2 \\
&= \left\{\frac{\sigma_B^2 - (\mu_d - \mu)^2}{(1+n_d\omega)^2}\right\}^2. \qquad (4)
\end{aligned}$$

L'écart quadratique $(\mu_d - \mu)^2$, qui intervient dans (2), est estimé naïvement par $(\hat{\mu}_d - \hat{\mu})^2$ avec un biais égal à $\sigma_W^2(n_d^{-1} - n^{-1}) \doteq \sigma_W^2/n_d$ et, en supposant que $\hat{\mu}_d$ suit une loi normale,

$$\begin{aligned}
&\text{EQM}\{(\hat{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} \\
&= \text{var}\{(\hat{\mu}_d - \hat{\mu})^2 | \mu_d\} \\
&\quad + [\text{E}\{(\hat{\mu}_d - \hat{\mu})^2 - (\mu_d - \mu)^2 | \mu_d\}]^2 \\
&\doteq \frac{2\sigma_W^4}{n_d^2} + 4(\mu_d - \mu)^2 \frac{\sigma_W^2}{n_d} + \frac{\sigma_W^4}{n_d^2} \\
&= \frac{\sigma_W^2}{n_d}\left\{\frac{3\sigma_W^2}{n_d} + 4(\mu_d - \mu)^2\right\}, \qquad (5)
\end{aligned}$$

dérivée d'après les propriétés de la loi du χ^2 non centré et d'une approximation en permettant que $n \to +\infty$. Une autre option consiste à utiliser $\tilde{\mu}_d$ au lieu de $\hat{\mu}_d$; des opérations élémentaires donnent les approximations

$$\text{E}\{(\tilde{\mu}_d - \hat{\mu})^2 | \mu_d\} \doteq (1-b_d)^2\left\{\frac{\sigma_W^2}{n_d} + (\mu_d - \mu)^2\right\}$$

$$\text{var}\{(\tilde{\mu}_d - \hat{\mu})^2 | \mu_d\} \doteq \frac{(1-b_d)^4}{n_d^2}\sigma_W^2\{2\sigma_W^2 + 4n_d(\mu_d - \mu)^2\},$$

où $b_d = 1/(1+n_d\omega)$, et donc

$$\begin{aligned}
&\text{EQM}\{(\tilde{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} \\
&= \text{var}\{(\tilde{\mu}_d - \hat{\mu})^2 | \mu_d\} + [\text{E}\{(\tilde{\mu}_d - \hat{\mu})^2 - (\mu_d - \mu)^2 | \mu_d\}]^2 \\
&\doteq (1-b_d)^4 \frac{3\sigma_W^4}{n_d^2} \\
&\quad + 2(1-b_d)^2(2 - 6b_d + 3b_d^2)\frac{\sigma_W^2(\mu_d - \mu)^2}{n_d} \\
&\quad + b_d^2(2-b_d)^2(\mu_d - \mu)^4. \qquad (6)
\end{aligned}$$

Cette approximation est valide uniquement pour $b_d = 1/(1+n_d\omega)$, de sorte qu'une approximation supplémentaire intervient quand nous substituons un choix éventuellement sous-optimal ou une estimation de b_d fondée sur une estimation de ω. En général, le coefficient b_d qui minimise l'EQM dans (6) diffère de $1/(1+n_d\omega)$ parce qu'avec $b_d = 1/(1+n_d\omega)$, le rétrécissement est optimal uniquement pour les cibles qui sont des transformations linéaires de μ_d (Shen et Louis 1998). Nous ne poursuivons pas cette route car, étant une fonction compliquée des paramètres, la solution est vraisemblablement sensible à l'erreur dans l'estimation de ces derniers. L'estimateur $(\hat{\mu}_d - \hat{\mu})^2$ pourrait être corrigé de son biais en estimant $(\mu_d - \mu)^2$, quoique l'on risque d'obtenir une estimation négative, surtout si n_d est faible.

Enfin, nous combinons les deux estimateurs (biaisés) de $\text{EQM}(\tilde{\mu}_d; \mu_d)$, l'estimateur moyenné $\sigma_B^2/(1+n_d\omega)$ et l'estimateur naïf dérivé de l'identité (2), en utilisant $(\hat{\mu}_d - \hat{\mu})^2$ comme estimateur de $(\mu_d - \mu)^2$. Les EQM de ces deux estimateurs dépendent de $(\mu_d - \mu)^2$, de sorte que nous remplaçons les termes pertinents par leurs espérances

sur l'ensemble des districts *d*. Nous remplaçons $(\mu_d - \mu)^2$ par σ_B^2, et $(\mu_d - \mu)^4$ par $3\sigma_B^4$ ou, en général, par $\kappa\sigma_B^4$, où κ est le aplatissement de la distribution (au niveau du district) de μ_d. Bien qu'il puisse sembler, à première vue, que nous n'ayons rien gagné, parce que nous devons encore éliminer la dépendance de l'EQM à l'égard de $(\mu_d - \mu)^2$ en utilisant σ_B^2 à la place, nous procédons maintenant à cette étape à un stade ultérieur. Dans les simulations présentées à la section 4, nous montrons que cela réduit l'effet indésirable du calcul d'une moyenne, ou moyennage.

Donc, nous recherchons le coefficient c_d qui minimise l'EQM espérée de l'estimateur composite de l'EQM,

$$\begin{aligned}&\widetilde{\mathrm{EQM}}(\tilde{\mu}_d; \mu_d)\\ &= (1-c_d)\widehat{\mathrm{EQM}}(\tilde{\mu}_d; \mu_d) + c_d\,\overline{\mathrm{EQM}}(\tilde{\mu}_d; \mu_d)\\ &= (1-c_d)\left\{(1-b_d)^2\,\frac{\sigma_W^2}{n_d} + b_d^2(\hat{\mu}_d - \hat{\mu})^2\right\} + c_d b_d \sigma_B^2. \quad (7)\end{aligned}$$

Pour évaluer l'EQM de l'estimateur de l'EQM, sous la forme d'une fonction de c_d, nous utilisons les expressions

$$\begin{aligned}&\overline{\mathrm{MSE}}\{b_d\sigma_B^2; \mathrm{MSE}(\tilde{\mu}_d; \mu_d)\} = 2b_d^4\sigma_B^4,\\ &\overline{\mathrm{MSE}}\{(\hat{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} \doteq \frac{\sigma_W^4}{n_d^2}(3 + 4n_d\omega),\\ &\overline{\mathrm{MSE}}\{(\tilde{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\}\\ &\qquad \doteq \frac{\sigma_W^4}{n_d^2}\{3(1-b_d)^4 + 3b_d^2(2-b_d)^2 n_d^2\omega^2\\ &\qquad\qquad + 2(1-b_d)^2(2 - 6b_d + 3b_d^2)n_d\omega\},\end{aligned}$$

obtenues en calculant la moyenne des équations respectives 4), 5) et 6); $(\mu_d - \mu)^2$ est remplacé par σ_B^2 et $(\mu_d - \mu)^4$, par $3\sigma_B^4$.

En supposant que les cibles au niveau du district μ_d suivent une loi normale, l'EQM de l'estimateur composite (7) est

$$\begin{aligned}&\mathrm{E}\{(1-c_d)(1-b_d)^2\,\frac{\sigma_W^2}{n_d} + (1-c_d)b_d^2(\hat{\mu}_d - \hat{\mu})^2\\ &\quad + c_d b_d\sigma_B^2 - b_d^2\sigma_W^2 n_d\omega^2 - b_d^2(\mu_d - \mu_d)^2\}^2\\ &= b_d^4\,\mathrm{E}\{(1-c_d)\sigma_B^2 n_d\omega + (1-c_d)(\hat{\mu}_d - \hat{\mu})^2\\ &\qquad + c_d\sigma_B^2(1 + n_d\omega) - \sigma_B^2 n_d\omega - (\mu_d - \mu)^2\}^2\\ &= b_d^4\,\mathrm{E}\{(1-c_d)(\hat{\mu}_d - \hat{\mu})^2 + c_d\sigma_B^2 - (\mu_d - \mu)^2\}^2\\ &\doteq b_d^4\left[(1-c_d)^2\left\{\frac{2\sigma_W^4}{n_d^2} + \frac{4\sigma_W^2}{n_d}(\mu_d - \mu)^2\right\}\right]\\ &\quad + b_d^4\left[(1-c_d)\frac{\sigma_W^2}{n_d} + c_d\{\sigma_B^2 - (\mu_d - \mu)^2\}\right]^2,\end{aligned}$$

en utilisant les identités $(1-b_d)^2 = b_d^2 n_d^2\omega^2$ et $\sigma_B^2 = \sigma_W^2\omega$ pour extraire le facteur b_d^4. En calculant l'espérance sur l'ensemble des districts en gardant intactes les tailles d'échantillons, nous obtenons

$$\begin{aligned}&\overline{\mathrm{EQM}}\{\widetilde{\mathrm{EQM}}(\tilde{\mu}_d; \mu_d)\}\\ &\qquad \doteq \frac{b_d^4}{n_d^2}\{(1-c_d)^2(3 + 4n_d\omega)\sigma_W^4 + 2c_d^2 n_d^2\sigma_B^4\}.\end{aligned}$$

Le minimum de cette fonction quadratique de c_d est atteint pour

$$c_d^* = \frac{3 + 4n_d\omega}{3 + 4n_d\omega + 2n_d^2\omega^2}.$$

Ce choix d'un coefficient c_d concorde avec nos attentes. Pour $n_d = 0, c_d^* = 1$ et nous nous appuyons uniquement sur l'estimateur de l'EQM moyenné, égal à σ_B^2. En outre, c_d^* est une fonction décroissante de n_d, qui converge vers zéro à mesure que n_d diverge vers $+\infty$; pour les grandes valeurs de n_d, nous utilisons l'estimateur naïf de l'EQM. Il s'agit également d'une fonction décroissante de ω; pour $\omega = 0$, c'est-à-dire $\sigma_B^2 = 0, c_d^* = 1$ pour chaque district *d*, ce qui confirme que $\mu_d \equiv \mu$ et que μ_d serait estimée avec précision *si* μ était connue. À mesure que ω augmente, $\sigma_B^2/(1 + n_d\omega)$ devient de moins en moins utile, parce que les écarts quadratiques $(\mu_d - \mu)^2$ sont fortement étalés (autour de σ_B^2).

Si nous corrigeons $(\hat{\mu}_d - \hat{\mu})^2$ de son biais en estimant $(\mu_d - \mu)^2$, l'EQM espérée de l'estimateur de rétrécissement est minimisée pour

$$c_d^\dagger = \frac{1 + 2n_d\omega}{(1 + n_d\omega)^2}.$$

Il est facile de vérifier que

$$c_d^* - c_d^\dagger = \frac{n_d^2\omega^2}{(1 + n_d\omega)^2}\,\frac{1}{3 + 4n_d\omega + 2n_d^2\omega^2},$$

de sorte qu'il soit assigné à l'estimateur corrigé du biais dérivé de (2) un poids plus grand (égal à $1 - c_d^\dagger$) qu'à l'estimateur naïf. Toutefois, la différence est faible pour toutes les valeurs de $n_d\omega$.

L'estimateur composite de l'EQM basé sur $(\tilde{\mu}_d - \hat{\mu})^2$ s'obtient de façon similaire, mais l'expression résultante est nettement plus compliquée. Le coefficient de rétrécissement optimal est

$$\begin{aligned}c_d^{*\prime} &= 3(1-b_d)^4 + 2(1-b_d)^2 f(b_d) n_d\omega - b_d(2-b_d) f(b_d) n_d^2\omega^2\\ &\quad \times [3(1-b_d)^4 + 2(1-b_d)^2 f(b_d) n_d\omega -\\ &\qquad \{2 - 4b_d(2-b_d) + 3b_d^2 f(b_d)\} n_d^2\omega^2],\end{aligned}$$

où $f(b_d) = 2 - 6b_d + 3b_d^2$. La dépendance à l'égard de b_d est particulièrement préoccupante, car en pratique, b_d est estimé et les propriétés de l'estimateur de l'EQM fondé sur le coefficient $c_d^{*\prime}$ estimé sont obligatoirement affectées par

l'incertitude au sujet de b_d. Dans les dérivations, nous avons utilisé l'identité $b_d = 1/(1 + n_d\omega)$, de sorte que cette expression ne pourrait pas être utilisée si les valeurs de b_d étaient fixées a priori.

4. Simulations

Les propriétés de l'estimateur composite de l'EQM ne pouvant être dérivées analytiquement, nous recourrons à la simulation. Nous considérons les conditions artificielles d'une enquête nationale menée selon un plan d'échantillonnage stratifié, dont les strates coïncident avec les 100 districts du pays pour lesquels ont souhaite estimer la moyenne d'une variable Y. Un échantillonnage aléatoire simple est appliqué à chaque strate, en supposant que la taille de leur population respective est pratiquement infinie. Nous avons généré les valeurs des moyennes à partir de la loi normale $N(\mu = 20, \sigma_B^2 = 8)$, et les tailles d'échantillon n_d, à partir de lois bêta conditionnelles normées, sachant les moyennes μ_d, de façon à injecter un minimum de dépendance des moyennes à l'égard des tailles d'échantillon. Avec cet ajustement, l'hypothèse sur laquelle s'appuie l'estimateur moyenné de l'EQM est fausse, mais cela risque de ne pas être décelé par une méthode diagnostique ou un test d'hypothèse, même dans le cas où μ_d est connue. La taille d'échantillon de l'un des districts a été modifiée de façon qu'elle soit beaucoup plus grande que les autres, afin de représenter la capitale du pays factice. Les lois intra-strate de Y sont $N(\mu_d, \sigma_W^2 = 100)$. Les moyennes au niveau du district et les tailles d'échantillon sont fixées dans les répétitions. En guise d'orientation, elles sont représentées graphiquement à la figure 1. Des numéros d'ordre allant de 1 à 100 sont affectés aux districts par ordre croissant de taille d'échantillon. La taille d'échantillon la plus faible est $n_1 = 15$ et la taille d'échantillon globale est 3 698.

Dans les simulations, qui comprennent 1 000 répétitions, nous générons les estimations directes $\hat{\mu}_d$ par des tirages aléatoires indépendants à partir de $N(\mu_d, \sigma_W^2/n_d)$ et les sommes corrigées des carrés intra-district, par des tirages indépendants à partir des lois du χ^2 normées appropriées avec $n_d - 1$ degrés de liberté. Puis, nous évaluons l'estimateur de rétrécissement $\tilde{\mu}_d$ pour chaque district d, et ensuite, l'estimateur moyenné, naïf et les deux estimateurs composites de l'EQM en utilisant les coefficients c_d^* et $c_d^\dagger$ ou les estimations naïves.

Dans le premier ensemble de répétitions, nous supposons que μ, σ_W^2 et σ_B^2 sont connues, de sorte que la simulation reproduit les résultats dérivés théoriquement et nous permet d'évaluer la qualité des estimateurs composites de l'EQM sans l'interférence de l'incertitude au sujet du coefficient de rétrécissement $b_d = 1/(1 + n_d\omega)$. Les résultats sont résumés graphiquement à la figure 2. Les biais empiriques (valeurs absolues) des quatre estimateurs de l'EQM sont tracés dans le panneau de gauche. Des cercles et des points noirs sont utilisés pour les estimateurs moyenné et naïf, respectivement, et les biais des estimateurs composites sont reliés par des traits pleins. Les valeurs absolues des biais empiriques sont tracées afin de mettre en relief leur forte association à la taille d'échantillon dans le cas de l'estimateur naïf. Pour 60 districts (60 %), l'estimateur composite de l'EQM présente un biais positif. Pour l'estimateur naïf, ce chiffre ou pourcentage est plus élevé (78), et pour l'estimateur moyenné, il est plus faible (52). Partout, le contributeur principal au biais de l'estimateur moyenné de l'EQM est l'écart de la distance quadratique $(\mu_d - \mu)^2$ par rapport à la variance au niveau du district σ_B^2. Les deux estimateurs composites, fondés sur $(\hat{\mu}_d - \hat{\mu})^2$ et sur sa version corrigée du biais, diffèrent si peu qu'il est impossible de faire la distinction entre leurs biais dans le tracé. Le diagramme montre que l'estimateur moyenné de l'EQM comporte un biais important pour quelques districts, y compris plusieurs dont la taille d'échantillon est grande. Les biais des estimateurs naïf et composite ne présentent pas de tels extrêmes.

Dans le panneau de droite, les racines de l'EQM des estimateurs de l'EQM sont représentées en utilisant les mêmes symboles et disposition. Le diagramme montre que l'estimateur naïf est inefficace, surtout pour les districts ayant les tailles d'échantillon les plus faibles, tandis que l'estimateur moyenné est très efficace pour certains, mais inefficace pour d'autres, sans aucune relation apparente avec leur taille d'échantillon. En fait, mise à part la taille d'échantillon, la grande efficacité est associée à la proximité de $(\mu_d - \mu)^2$ par rapport à σ_B^2 et la faible efficacité, aux valeurs les plus petites et les plus grandes de $(\mu_d - \mu)^2$. Par exemple, la racine de l'EQM empirique de l'estimateur moyenné de l'EQM pour le district 1, avec $n_1 = 15$, est 2,63, tandis que son équivalent pour le district 11 $(n_{11} = 16)$ est 0,049. Les moyennes de population sont $\mu_1 = 24,55$, surpassant $\mu + \sigma_B$ de 1,72, et $\mu_{11} = 22,87$, ne différant de $\mu + \sigma_B$ que de 0,04. Les racines de l'EQM pour l'estimateur naïf sont 5,08 et 3,51, et celles pour l'estimateur composite, 2,10 et 1,00 pour les districts 1 et 11, respectivement. L'estimateur composite de l'EQM donne des résultats nettement plus uniformes, atténuant les défauts des estimateurs moyenné et naïf.

Les trois estimateurs sont prudents (ont un biais positif) pour les districts pour lesquels l'EQM de $\tilde{\mu}_d$ est relativement faible. L'estimateur moyenné possède un biais négatif lorsque les EQM sont relativement grandes. L'estimateur composite est également entaché d'un biais négatif pour certains districts, mais celui-ci a tendance à être plus faible en valeur absolue. Pour les districts ayant les tailles d'échantillon les plus faibles, l'estimateur composite n'est pas très

efficace, parce que l'estimateur naïf est très inefficace. Pour quelques-uns de ces districts, la combinaison des estimateurs est contreproductive, à cause du moyennage, mais il est impossible d'identifier ces districts d'après une seule réalisation de l'enquête.

Nous étudions maintenant des conditions moins favorables, sous lesquelles les hypothèses de normalité de μ_d sur l'ensemble des districts et des observations élémentaires y_{id} dans les districts sont encore satisfaites, mais les paramètres globaux, μ, σ_W^2 et σ_B^2 sont inconnus et estimés. Nous utilisons les mêmes moyennes μ_d et tailles d'échantillon n_d qu'à la figure 1. Les résultats des simulations sont résumés à la figure 3. Dans le panneau de gauche, nous représentons les moyennes empiriques des estimateurs de l'EQM en utilisant les mêmes symboles qu'à la figure 2, ainsi que les EQM empiriques (représentées par des croix « + ») des estimateurs de rétrécissement $\tilde{\mu}_d$. Les moyennes empiriques des estimateurs moyennés présentent une courbe régulière, parce que dans chaque répétition les estimations dépendent uniquement de la taille d'échantillon n_d et du rapport des variances estimées $\hat{\omega}$. En ce qui concerne le biais, les estimateurs naïfs présentent une courbe régulière, semblable à celle de la figure 2. Les estimateurs naïfs sont entachés d'un biais positif qui diminue avec la taille d'échantillon. Les estimateurs moyennés sont beaucoup trop prudents; leur moyenne ne s'écarte pas de la tendance lissée. Les estimateurs composites de l'EQM s'écartent de cette tendance dans la direction appropriée, mais non pleinement. Leur biais moyen est positif, égal à 0,22, ou 10 % (2,42 *vs* 2,20), et ils surestiment l'EQM cible pour 70 des 100 districts.

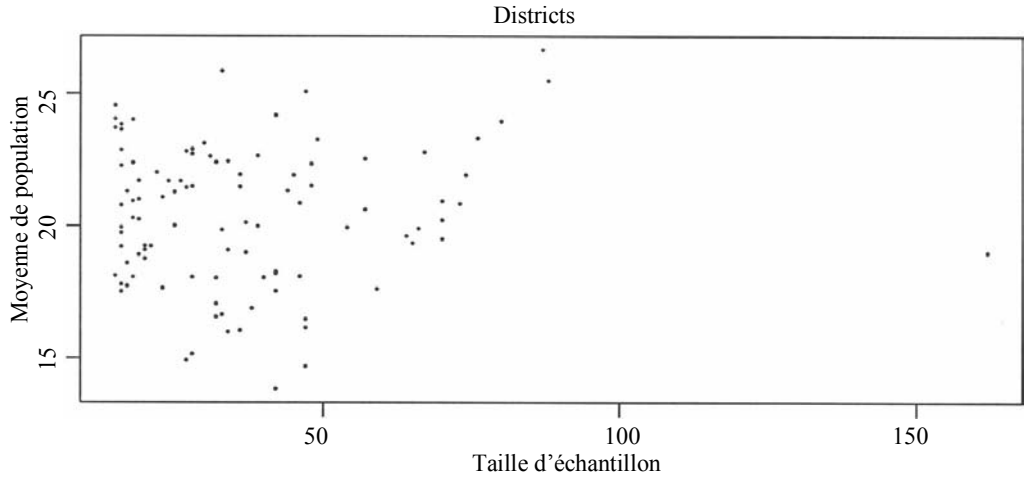


Figure 1 Tailles d'échantillon au niveau du district et moyennes de population de *Y*. Valeurs générées artificiellement

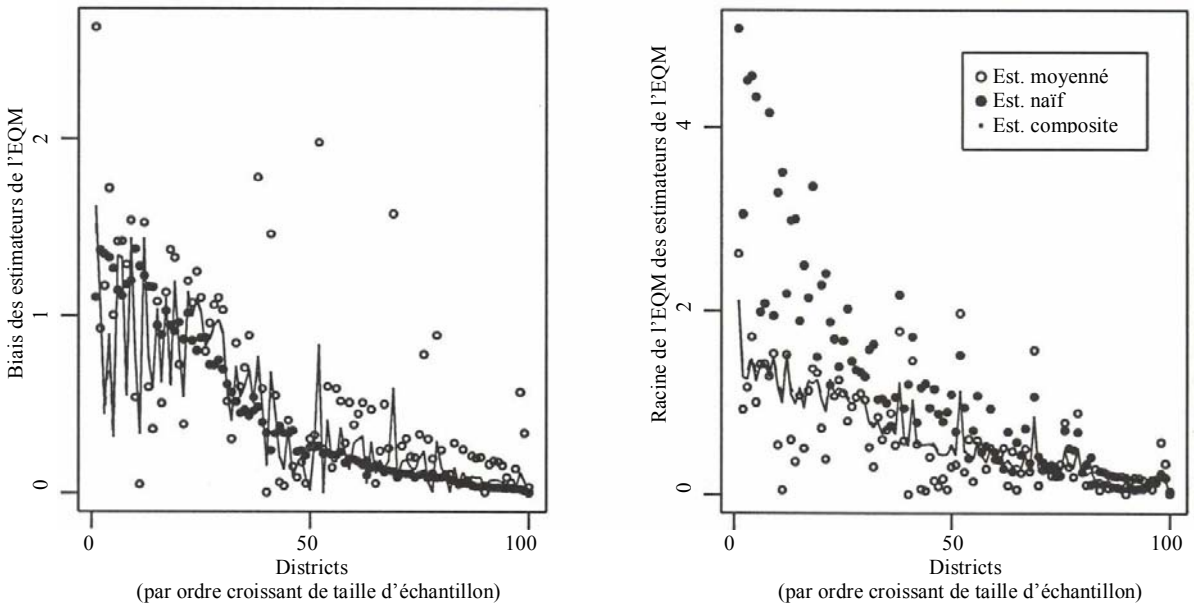


Figure 2 Biais et racine de l'EQM des estimateurs de l'EQM des estimateurs pour petits domaines empiriques bayésiens. Fondés sur des simulations dans des conditions artificielles. Les valeurs du biais et de la racine de l'EQM de l'estimateur composite sont reliées par des traits pleins

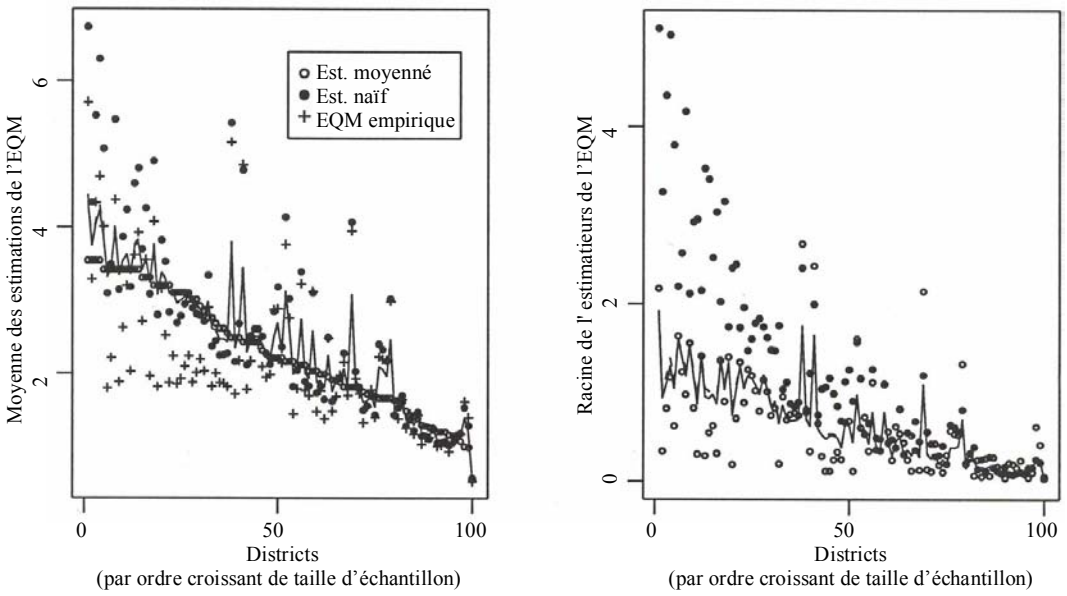


Figure 3 Moyenne et racine de l'EQM des estimateurs de l'EQM des estimateurs pour petits domaines empiriques bayésiens. Les paramètres globaux μ, σ_W^2 et σ_B^2 sont estimés

Le panneau de droite représente la racine de l'EQM des estimateurs de l'EQM. L'estimateur naïf est inefficace, tandis que l'estimateur moyenné est très efficace pour certains districts et assez inefficace pour d'autres. L'estimateur composite de l'EQM est plus efficace que l'estimateur naïf ou l'estimateur moyenné pour 36 districts; il est plus efficace que l'estimateur moyenné pour exactement la moitié des districts, mais il ne présente aucune faiblesse manifeste. Comme dans les conditions favorables (figure 2), les différences dues à la correction du biais de $(\hat{\mu}_d - \hat{\mu})^2$ dans l'estimation composite de l'EQM (en utilisant les coefficients c_d^* ou $c_d^{*\prime}$) sont négligeables.

Puis, nous comparons les estimateurs de l'EQM pour les moyennes de $Y^2/100$ au niveau du district que nous dénotons ν_d. Les hypothèses de normalité intra et inter-districts ne sont plus appropriées. Nous appliquons les méthodes qui s'appuient sur les hypothèses de normalité pour évaluer la robustesse des estimateurs composites, mais aussi pour mettre en contraste les différences dues au moyennage et les conséquences de l'utilisation de modèles « incorrects ». Nous choisissons la transformation quadratique, parce que les espérances intra-district sont connues, égales à $(\mu_d^2 + \sigma_W^2)/100$, et pourraient être estimées par

$$\tilde{\nu}_d^* = \frac{\tilde{\mu}_d^2 - \widehat{\mathrm{EQM}}(\tilde{\mu}_d) + \hat{\sigma}_W^2}{100}. \qquad (8)$$

Nous dénotons par $\tilde{\nu}_d$ les estimateurs empiriques bayésiens appliqués à $y_{id}^2/100$.

Les résultats des simulations fondées sur les valeurs de $y_{id}^2/100$ sont présentés à la figure 4, en utilisant la même disposition et les mêmes symboles qu'à la figure 3. Nous arrivons aux mêmes conclusions qu'auparavant au sujet des biais et des racines de l'EQM, excepté que l'estimateur naïf est encore plus inefficace et que les propriétés de l'estimateur moyenné sont encore plus erratiques, en ce sens qu'il est à la fois très efficace et très inefficace pour un plus grand nombre de districts que dans les conditions plus favorables de la figure 3. L'estimateur naïf est prudent, mais pour certains districts dont la taille d'échantillon n_d est faible, il l'est nettement trop, et son EQM est très grande.

Nous contrastons ces conclusions à l'aide d'une comparaison de l'estimation des moyennes de $Y^2/100$ au niveau du district par $\tilde{\nu}_d^*$, en transformant les estimations $\tilde{\mu}_d$ conformément à (8). L'estimateur $\tilde{\nu}_d^*$ est plus efficace que $\tilde{\nu}_d$ pour la plupart des districts (90, en fait), et quand il est moins efficace, la différence relative entre leurs EQM est inférieure à 4 %. Pour quelques districts, la différence d'efficacité est perceptible, dépassant 20 % pour dix districts. Toutefois, les écarts entre les EQM sont faibles comparativement aux biais dans l'estimation de ces EQM, comme le montre la figure 5. Les biais et les EQM de $\tilde{\nu}_d$ sont représentés par des points noirs reliés entre eux pour $\tilde{\nu}_d^*$.

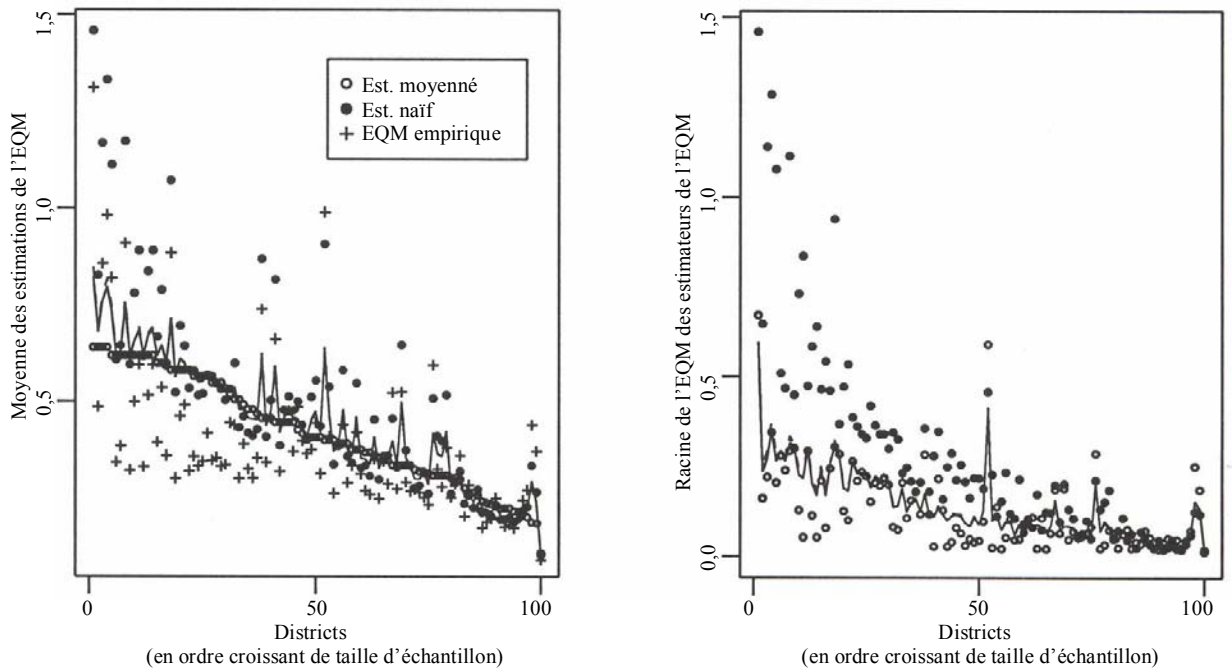


Figure 4 Moyenne et racine de l'EQM des estimateurs de l'EQM des estimateurs pour petits domaines empiriques bayésiens; estimation des moyennes de Y^2 /100

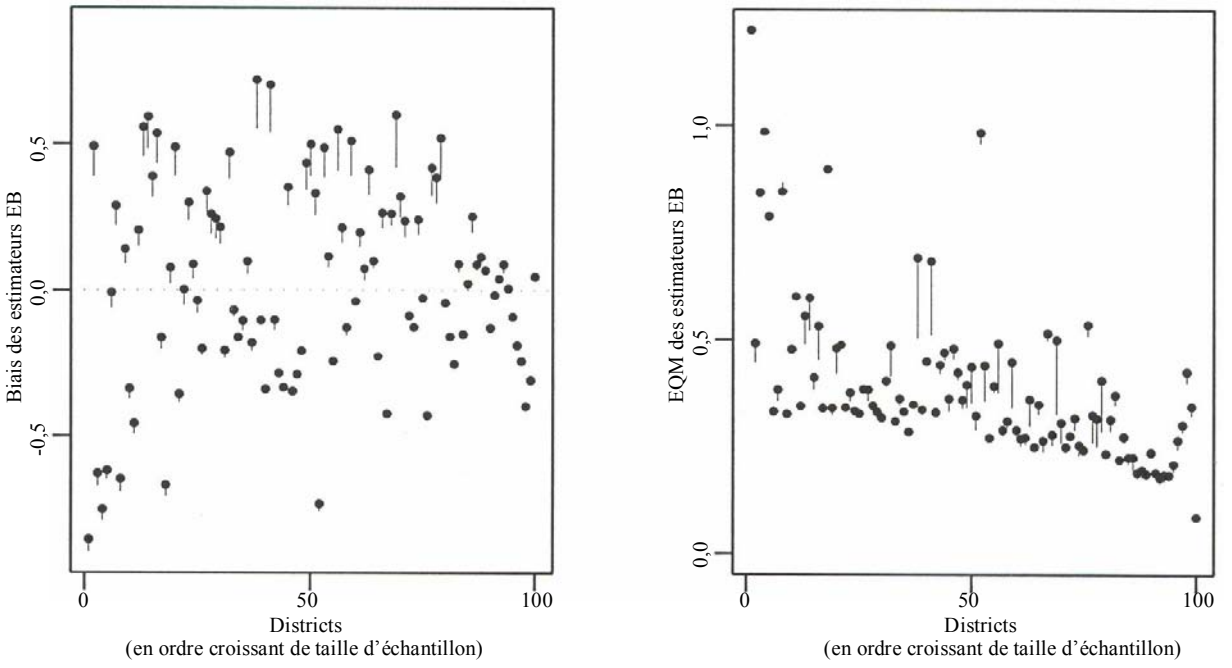


Figure 5 Biais et EQM des estimateurs de ν_d. Les segments verticaux relient les valeurs associées à $\tilde{\nu}_d^*$ et à $\tilde{\nu}_d$. Les valeurs associées à $\tilde{\nu}_d$ sont représentées par des points noirs

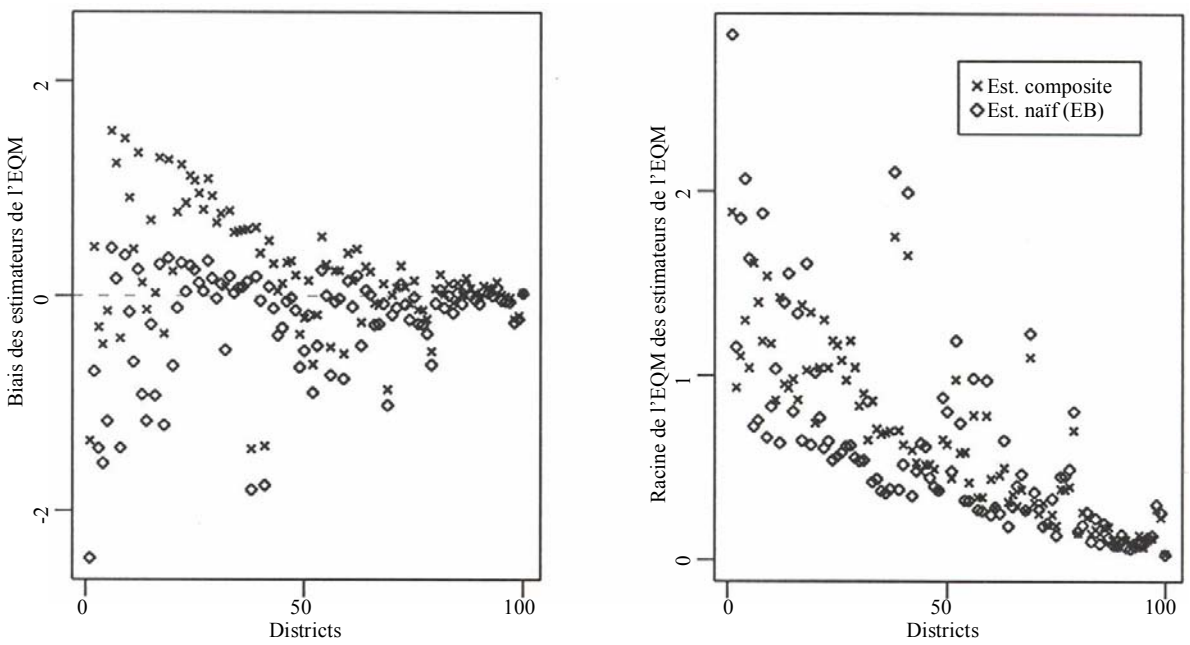


Figure 6 Biais et racine de l'EQM des estimateurs composite et naïf empirique bayésien de l'EQM de $\tilde{\mu}_d$

Le manque d'efficacité de $\tilde{\nu}_d$ est dû, en partie, à son biais; ce dernier est supérieur à celui de $\tilde{\nu}_d^*$ pour tous les districts, sauf deux, mais la différence n'est non négligeable que si les deux estimateurs sont biaisés positivement. Donc, le gain d'efficacité est faible si l'on effectue l'analyse de façon à ce que les hypothèses concernant les lois soient satisfaites et faibles. Les gains sont modestes comparativement à l'accroissement de la difficulté à estimer l'efficacité, telle qu'elle est exprimée par $\mathrm{EQM}(\tilde{\nu}_d^*; \nu_d)$. Bien que la variance d'échantillonnage de $\hat{\sigma}_W^2$ soit négligeable dans les enquêtes à grande échelle, la contribution de $\widehat{\mathrm{EQM}}(\tilde{\mu}_d; \mu_d)$ à $\mathrm{EQM}(\tilde{\nu}_d^*; \nu_d)$ ne peut pas être ignorée.

La figure 6 compare l'estimateur composite de l'EQM à l'estimateur naïf de l'EQM de $\tilde{\mu}_d$ basé sur l'estimateur empirique bayésien de μ_d; il est dérivé par substitution de $\tilde{\mu}_d$ à μ_d dans (2). Par souci de concision, nous l'appelons estimateur naïf EB. Comme prévu à la section 3, il a tendance à sous-estimer sa cible. Il est plus efficace que l'estimateur composite de l'EQM pour la moitié des districts (52 sur 100), mais ses propriétés sont moins uniformes. En principe, l'estimateur naïf EB pourrait être amélioré par combinaison à l'estimateur moyenné; cependant, cette combinaison ne produit qu'une amélioration faible, même dans les conditions favorables (μ, σ_W^2 et σ_B^2 connue), et elle est nuisible pour plusieurs districts dans les conditions moins favorables. Nous omettons les détails.

À titre de simulation finale, nous considérons une variable de résultat binaire qui indique si $Y < 5$, de façon que les pourcentages au niveau du district soient dans la fourchette de 1,5 à 18,8 et que la dépendance du pourcentage à l'égard de la variance intra-district soit importante. La moyenne des pourcentages au niveau du district est de 6,85; l'asymétrie importante de ces pourcentages (coefficient d'asymétrie égal à 1,01 et aplatissement égal à 3,78) représente un test sévère de la méthode.

Dans la simulation, les pourcentages au niveau du district sont estimés par la version univariée de la méthode de rétrécissement décrite dans Longford (1999 et 2005, chapitre 8). Les résultats sont résumés à la figure 7. L'EQM est surestimée par les trois estimateurs pour la plupart des districts, sauf une minorité pour lesquels l'EQM empirique est plusieurs fois plus élevée que pour les autres. L'estimateur naïf présente un biais important pour la plupart des districts. L'estimateur moyenné est contrôlé moins strictement que pour les résultats qui suivent une loi normale, car le coefficient de rétrécissement dépend aussi de la proportion estimée, qui est tronquée par le bas à 2 % pour éviter une variance estimée nulle $\hat{p}_d(1-\hat{p}_d)/n_d$. Le graphique des estimations composites de l'EQM présente des pics pour les districts appropriés, mais ces pics sont beaucoup trop courts pour réduire considérablement le biais.

Dans le cas de l'estimateur moyenné, les EQM sont pour la plupart satisfaisantes, mais elles sont très grandes pour plusieurs districts. Pour ces districts, l'estimateur naïf de l'EQM est encore moins efficace. L'estimateur composite de l'EQM est moins efficace que l'estimateur moyenné pour un grand nombre de districts, mais les écarts sont assez faibles, compensés par les gains d'efficacité pour les districts pour lesquels l'estimateur moyenné est moins efficace. L'estimateur naïf EB de l'EQM ressemble à de nombreux égards à l'estimateur naïf de l'EQM; il n'est pas représenté dans le diagramme.

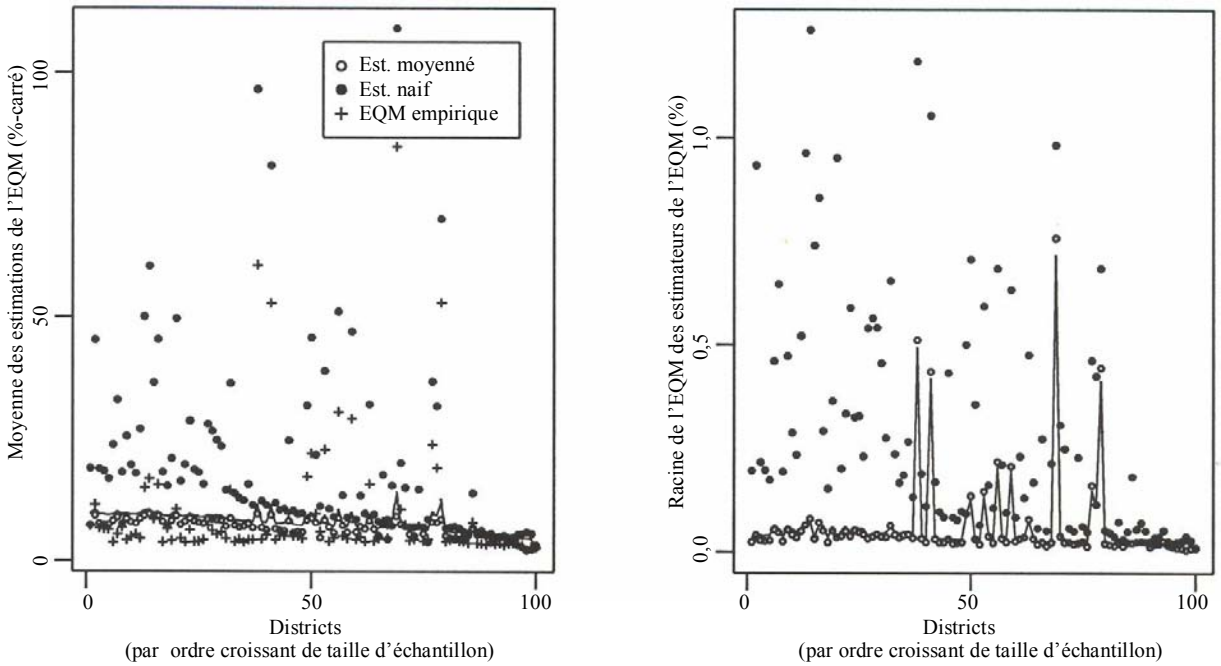


Figure 7 Moyenne et racine de l'EQM des estimateurs composites naïf et moyenné des EQM des pourcentages au niveau du district

En conclusion, cette simulation montre que, quand l'un des estimateurs de l'EQM, ici l'estimateur naïf, est très inefficace, il contribue néanmoins, ne fût-ce que très modestement, à l'efficacité de l'estimateur composite de l'EQM. L'estimateur composite tire parti du meilleur des estimateurs moyenné et naïf, même dans des conditions non favorables. Une difficulté qui persiste est d'arriver à combiner les estimateurs naïf et moyenné de façon à satisfaire un critère particulier représentant un compromis entre le niveau de précision correspondant aux districts pour lesquels l'estimation est de haute précision et un plus haut niveau de précision pour les districts où la précision des estimations est faible. Par exemple, nous pourrions nous préoccuper moins de l'estimation de l'EQM pour les districts dont la représentation dans l'échantillon est importante et davantage de celle de l'EQM pour les districts représentés parcimonieusement. En outre, certains districts (par exemple, ceux situés dans une région particulière) pourraient présenter un intérêt spécial, non relié à leur représentation. Naturellement, dans ces circonstances, la première étape est la définition d'un critère ou d'une classe de critères qui reflètent les priorités inférentielles, et cette définition sera obligatoirement particulière à chaque enquête et à chaque client. Voir Longford (2006) pour certaines propositions.

4.1 Perfectionnements et extensions

Plusieurs éléments de réalisme peuvent être intégrés dans la dérivation de l'estimateur composite de l'EQM. Premièrement, l'incertitude au sujet de $\hat{\mu}_d$ peut être reflétée en reconnaissant que $\hat{\mu}_d$ et $\hat{\mu}$ sont corrélées. Donc, $\text{var}(\hat{\mu}_d - \hat{\mu}) = \sigma^2_W\,(1/n_d - 1/n)$ et l'approximation (5) devient une égalité quand les deux occurrences de σ^2_W/n_d sont remplacées par $\sigma^2_W(1/n_d - 1/n)$. Cela n'entraîne qu'un léger changement quand $n_d \ll n$, ce qui est le cas pour la plupart des districts. Si le pays possède un district dominant, dont la taille d'échantillon représente une grande fraction de la taille d'échantillon globale, cet ajustement pourrait être pertinent, mais il a un effet négligeable sur l'estimation de l'EQM, car même l'estimation directe de la moyenne pour le district est quasi efficace.

Un perfectionnement similaire peut être appliqué à l'estimateur empirique bayésien de μ_d. Il revient remplacer n_d par $1/(n_d^{-1} - n^{-1}) = n_d n/(n - n_d)$ dans le coefficient $b_d = 1/(1 + n_d\omega)$. Le changement ne devient non négligeable que dans le cas d'un district dominant, mais pour un tel district, le rétrécissement ne produit qu'une très petite amélioration par rapport à l'estimation directe avec ou sans cet ajustement.

L'adaptation à des plans d'échantillonnage qui diffèrent de l'échantillonnage aléatoire stratifié et qui associent des sujets à des poids d'échantillonnage ne génère pas plus de

problèmes dans l'estimation composite que dans l'estimation directe basée sur ce genre de plans et de pondérations, parce que l'estimation composite requiert uniquement les variances d'échantillonnage de $\hat{\mu}_d, \hat{\mu}$ et de leurs fonctions. De même, l'exploitation de l'information auxiliaire par recours à la régression (empirique bayésienne)

$$y_{jd} = \mathbf{x}_{jd}\boldsymbol{\beta} + \delta_d + \varepsilon_{jd},$$

avec des échantillons aléatoires indépendants, $\delta_d \sim N(0, \sigma_B^2)$ et $\varepsilon_{jd} \sim N(0, \sigma_W^2)$, équivaut à remplacer $\hat{\mu}$ dans (1) par la prédiction $\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}$, où $\hat{\mathbf{x}}_d$ est le vecteur des moyennes pour le district d et $\hat{\boldsymbol{\beta}}$ est le vecteur des estimations des paramètres de régression. Pour le voir, nous exprimons l'ajustement empirique bayésien pour le district d sous la forme

$$\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}} + \frac{n_d\omega}{1+n_d\omega}(\hat{\mu}_d - \hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}) = \frac{n_d\omega}{1+n_d\omega}\hat{\mu}_d + \frac{1}{1+n_d\omega}\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}.$$

Pfeffermann et coll. (1998) discutent des problèmes associés à l'ajustement des modèles empiriques bayésiens aux observations avec poids d'échantillonnage. L'estimation composite s'appuie sur des estimateurs directs $\hat{\boldsymbol{\mu}}_d$ et $\hat{\boldsymbol{\mu}}$ pour les vecteurs de toutes les variables concernées et leurs matrices de variances d'échantillonnage estimées; leur évaluation est une tâche standard en théorie de l'échantillonnage. Un problème demeure non résolu dans le cas des estimateurs empiriques bayésiens quand $\hat{\mathbf{x}}_d$ est basé sur un très petit nombre d'observations, parce que l'incertitude au sujet de μ_d est alors grossie, même si l'ajustement du modèle est très bon; si le vecteur des moyennes $\mathbf{x}_d$ était connu (d'après des sources externes à l'enquête), μ_d pourrait être estimée de manière beaucoup plus efficace en utilisant $\mathbf{x}_d\hat{\boldsymbol{\beta}}$. L'estimation composite permet de contourner ce problème en recherchant la combinaison de moyennes de variables auxiliaires, qu'elles soient connues ou estimées d'après l'enquête ou d'autres sources, visant directement à minimiser l'EQM de la combinaison (Longford 1999).

L'approche élaborée à la section 3 peut être adaptée facilement à d'autres distributions que la loi normale, à condition que les aplatissements nécessaires pour évaluer la variance au niveau du district de $(\mu_d - \mu)^2$ et la variance d'échantillonnage de $(\hat{\mu}_d - \mu)^2$ soient connus. En pratique, l'aplatissement dépend de la moyenne μ_d, ce qui crée des difficultés qui ne peuvent être surmontées que par des approximations ou des calculs de valeur moyenne. L'estimation de proportions p_d d'après des données dichotomiques est un exemple typique. Nous avons

$$\begin{aligned}\text{var}\{(\hat{p}_d - p)^2\} &= \frac{v_d}{n_d^3}(1 - 3p_d + 3p_d^2) \\ &+ \frac{4v_d}{n_d^2}(1-2p_d)(p_d - p) + \frac{6v_d}{n_d}(p_d - p)^2 - \frac{v_d^2}{n_d^2},\end{aligned}$$

où $v_d = p_d(1-p_d)/n_d$ et p est la proportion nationale. La dépendance complexe à l'égard des proportions p_d mal estimées présente un défi analytique qui n'a pas de solution universelle.

Tout au long de l'exposé, nous avons supposé que la valeur du ratio des variances ω est connue. En pratique, ω est estimé. Il est difficile de tenir compte de l'incertitude au sujet de ω analytiquement, mais son effet sur l'estimation de μ_d et de $\text{EQM}(\tilde{\mu}_d; \mu_d)$ peut être évalué par analyse de sensibilité en répétant les simulations décrites à la section 4 pour une gamme de valeurs plausibles de ω. Puisqu'un ensemble de simulations requiert environ une minute de temps de processeur, il s'agit d'une tâche de calcul faisable. Une difficulté de ce genre d'évaluation tient au fait qu'avec une valeur hypothétique altérée de ω, l'estimateur $\tilde{\mu}_d$ est changé, et, donc, la cible de l'estimateur composite de l'EQM est changée également. Une approche de rechange informelle consiste à considérer les conséquences d'une sous-estimation et d'une surestimation de la valeur de ω. Lors de l'estimation de μ_d, il est conseillé d'errer dans la direction d'une valeur de ω plus grande, donnant plus de poids à l'estimateur direct $\hat{\mu}_d$ (Longford 2005, chapitre 8). Pour estimer l'EQM de $\tilde{\mu}_d$, nous pourrions préférer errer dans la direction de l'estimateur moyenné, plus stable. Cela revient à accroître la valeur du coefficient c_d^* et, comme c_d^* est une fonction décroissante de ω, à réduire la valeur de ω utilisée pour fixer c_d^*. Naturellement, la modération est de rigueur, afin de ne pas écarter entièrement la contribution de l'estimateur naïf de l'EQM.

5. Conclusion

L'approche élaborée dans le présent article applique la notion générale de rétrécissement à l'estimation de l'EQM des estimateurs pour petits domaines et réduit l'effet du moyennage, considéré comme indésirable sous l'angle de l'approche fondée sur le plan de sondage, dans laquelle les quantités de population μ_d des districts du pays sont fixes. Nous nous sommes concentrés sur l'amélioration individuelle de l'estimation de l'EQM pour chaque district. En pratique, l'amélioration de l'estimation est plus importante pour certains districts que pour d'autres. De nombreuses enquêtes sont conçues pour faire d'autres inférences que l'estimation pour petits domaines ou ne tiennent compte que périphériquement des petits domaines dans la planification, si bien qu'elles peuvent produire des estimateurs plus que satisfaisants pour certains districts, habituellement les plus peuplés, mais moins satisfaisants pour d'autres, souvent ceux qui sont peu peuplés. Dans de telles conditions, on devrait accorder une priorité inférentielle relativement plus grande à ces derniers districts. Les estimateurs de rétrécissement des moyennes et des proportions de petits domaines

ont cette propriété, et les simulations décrites à la section 4 indiquent que l'estimation composite de l'EQM a une propriété semblable, du moins en ce qui a trait à l'estimateur moyenné.

Pour une taille donnée du biais dans l'estimation d'une EQM, nous préférons un biais positif, parce que nous considérons la sous-estimation de la précision comme statistiquement « malhonnête », tandis que la surestimation ne présente simplement pas l'estimation sous la lumière qu'elle mérite, autrement dit, nous mettons mal en valeur le résultat de notre effort analytique. Dans cette perspective, le coefficient optimal c_d dans (7) ne devrait pas être recherché par minimisation de l'EQM de la combinaison, mais à l'aide d'un critère qui considère la sous-estimation de l'EQM comme une erreur plus grave que sa surestimation dans la même proportion. Trouver un critère approprié pour lequel l'optimisation est soluble est un problème ouvert. L'estimateur composite de l'EQM dérivé à la section 3 a tendance à surestimer l'EQM, mais il ne s'agit pas d'un acte délibéré.

Nous avons expérimenté avec l'estimation du maximum de vraisemblance (ML) et du maximum de vraisemblance restreint (REML); dans les conditions utilisées pour les simulations, les différences entre les deux approches sont très faibles. L'avantage de l'estimation sans biais de la variance σ_B^2 est perdu quand $\hat{\sigma}_B^2$ est soumis à une transformation non linéaire, et l'efficacité n'est maintenue qu'asymptotiquement par les transformations. Cependant, l'estimation pour petits domaines est un problème typiquement de petit échantillon.

L'approche décrite dans le présent article illustre l'universalité de l'idée générale consistant à combiner deux estimateurs possibles. L'estimateur composite exploite les avantages et réduit les inconvénients des estimateurs composants. Son application n'est pas préjudiciable lorsque l'un des estimateurs est de loin inférieur à l'autre. Une forme de moyennage intervient même dans l'estimateur composite de l'EQM, de sorte qu'elle contribue à sa robustesse en améliorant les écarts par rapport aux hypothèses faites dans le développement théorique, telles que la présence d'hétéroscédasticité et de lois asymétriques (non normales) dans le district.

L'intégration des priorités inférentielles, en fait la rerépartition de la précision dans l'estimation des EQM pour les petits domaines, est un problème ouvert. Un problème similaire, c'est-à-dire concevoir des enquêtes pour l'estimation pour petits domaines de telle façon que la précision soit suffisante dans l'approche fondée sur un modèle (avec moyennage) est abordé par Longford (2006).

Remerciements

Les travaux décrits dans ce manuscrit ont été financés en partie par les subventions SEC2003-04476 et SAB2004-0190 du ministère espagnol de l'Éducation et des Sciences. L'auteur remercie deux examinateurs et un rédacteur associé de leurs commentaires éclairés et constructifs.

Bibliographie

EURAREA Consortium. (2004). EURAREA Project Final Reference Volume. Enhancing Small-Area Estimation Techniques to Meet European Needs. Office for National Statistics, London. Disponible à http://www.statistics.gov.uk/eurarea.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society*, Séries A, 162, 227-245.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. New York : Springer-Verlag.

Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, 97-106.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1988). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Séries B, 60, 23-40.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Shen, W., et Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society*, Séries B, 60, 455-471.

Techniques d'enquête, juin 2007
Volume 33, N° 1, pp. 93-97
Statistique Canada, N° 12-001 au catalogue

Traitement de la non-réponse dans les sondages en grappes

Jun Shao [1]

Résumé

Dans les sondages en grappes, la non-réponse concernant une variable dépend souvent d'un effet aléatoire au niveau de la grappe et n'est donc pas ignorable. Les estimateurs de la moyenne de population obtenus par imputation par la moyenne ou par repondération sous l'hypothèse de non-réponse ignorable sont alors biaisés. Nous proposons un estimateur sans biais de la moyenne de population obtenu par imputation ou par repondération dans chaque grappe échantillonnée ou dans un groupe de grappes échantillonnées ayant une caractéristique commune. Nous présentons certains résultats obtenus par simulation en vue d'étudier les propriétés de l'estimateur proposé.

Mots clés : Non-réponse non ignorable; non-réponse basée sur un effet aléatoire; imputation; regroupement de grappes.

1. Introduction

La non-réponse existe dans la plupart des problèmes de sondage. La probabilité d'avoir un non-répondant à un item (variable) y dépend habituellement de la valeur inobservée de y, ce qui rend fort difficile le traitement des non-réponses. Les méthodes appliquées habituellement (comme la repondération et l'imputation) sont toutes fondées sur l'hypothèse que la non-réponse est ignorable étant donné une variable auxiliaire. Plus précisément,

$$P(y \text{ est un répondant} | y, z) = P(y \text{ est un répondant} | z), \quad (1)$$

où z est une variable auxiliaire dont les valeurs sont observées pour toutes les unités échantillonnées. Autrement dit, sachant z, la valeur de y et sa situation de réponse sont statistiquement indépendantes. L'hypothèse (1) est appelée mécanisme de réponse non confondu par Lee, Rancourt et Särndal (1994). Selon la terminologie de Rubin (1976), la non-réponse sous (1) est ignorable sachant z.

Dans certaines situations, il est difficile de trouver une variable z qui satisfait (1). L'objectif du présent article est d'étudier une méthode de traitement de la non-réponse dans le cas d'un sondage en grappes, en supposant qu'une variable z satisfaisant (1) n'est pas disponible. Le sondage en grappes comporte un échantillonnage à deux degrés; les unités sélectionnées au premier degré sont des grappes contenant des unités qui, à leur tour, sont échantillonnées au deuxième degré. L'échantillonnage en grappes est utilisé pour des raisons économiques. Il est nécessaire lorsque l'on ne dispose d'aucune liste fiable des unités de population de deuxième degré (par exemple, lorsqu'il n'existe aucune liste complète des personnes, mais que l'on dispose d'une liste de ménages). Dans le sondage en grappes, la variable d'intérêt y peut être décomposée sous la forme $y = \mu + b + e$, où μ est une moyenne globale inconnue de y, b est un effet aléatoire au niveau de la grappe (toutes les unités d'une grappe donnée ont en commun le même effet aléatoire b) et e est un effet aléatoire intragrappe. Dans de nombreux cas, l'instrument de la corrélation de la valeur de la variable y et de la situation de réponse est l'effet aléatoire non observé au niveau de la grappe b :

$$P(y \text{ est un répondant} | y, b) = P(y \text{ est un répondant} | b), \quad (2)$$

c'est-à-dire que, si b était observé, alors nous obtiendrions l'hypothèse (1) avec $z = b$. Par exemple, supposons que les grappes soient les ménages et que dans chacun ceux-ci, une seule personne réponde au questionnaire pour tous les membres du ménage échantillonné. Il est vraisemblable que la probabilité des réponses dépende de la variable b au niveau du ménage mais non de la variable e intraménage.

L'hypothèse (2) a été la première utilisée par Wu et Carroll (1988) dans la résolution d'un problème relevant du domaine de la santé où les grappes ont une structure longitudinale (mesure répétée). Ils ont donné à (2) le nom de censure informative (données manquantes) et proposé une méthode sous certaines hypothèses paramétriques concernant la probabilité $P(y \text{ est un répondant} | b)$ et la loi de y. Plus tard, Little (1995) a donné à ce genre de mécanisme de création de données manquantes le nom de mécanisme non ignorable de création de données manquantes basée sur un coefficient aléatoire. Donc, nous donnerons à l'hypothèse (2) le nom de mécanisme de réponse non ignorable basée sur un effet aléatoire. Puisque b n'est pas observé, le mécanisme de réponse (2) est effectivement non ignorable.

Dans le cas de données d'enquête, il est difficile de postuler un modèle paramétrique pour la loi de y. Il est, de surcroît, difficile d'ajuster un modèle paramétrique au mécanisme de non-réponse sous (2), car b n'est pas observé. Après la présentation de certains détails sur le plan de sondage et de nos hypothèses, nous proposons à la section 2

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706, États-Unis.

une méthode d'estimation de la moyenne de population de y sous le mécanisme de réponse (2) qui ne nécessite pas de modèle paramétrique pour le mécanisme de réponse. Nous supposons que y suit un modèle à effet aléatoire (de grappe), mais nous ne formulons aucune hypothèse paramétrique concernant la loi de y. À la section 3, nous exposons les résultats d'une étude par simulation réalisée en vue d'étudier les propriétés de l'estimateur proposé. Enfin, nous présentons une discussion à la dernière section.

2. Principaux résultats

Soit S un échantillon de grappes de taille n tiré d'une population P. Dans la $i^{\text{ième}}$ grappe échantillonnée, soit S_i l'échantillon de deuxième degré de taille $m_i \geq 2$ provenant d'une population P_i. Pour l'unité échantillonnée $j \in S_i$, nous construisons un poids de sondage w_{ij} (d'après la spécification du plan d'échantillonnage) tel que, en l'absence de non-réponse, $\hat{Y} = \sum_{i\in S}\sum_{j\in S_i} w_{ij} y_{ij}$ est un estimateur sans biais du total de population Y de toute variable y, c'est-à-dire $E_s(\hat{Y} - Y) = 0$, où y_{ij} est la valeur de y de l'unité j dans la grappe i, $Y = \sum_{i\in P}\sum_{j\in P_i} y_{ij}$, et E_s est l'espérance sous échantillonnage répété.

Soit y la variable d'intérêt. Nous adoptons une approche avec modèle d'imputation, autrement dit, nous supposons que chaque y_{ij} dans la population est une variable aléatoire telle que

$$y_{ij} = \mu_i + b_i + e_{ij}, \tag{3}$$

où μ_i est un paramètre inconnu, b_i est un effet aléatoire au niveau de la grappe non observé de moyenne 0 et de variance finie, e_{ij} est un effet aléatoire intragrappe non observé de moyenne 0 et de variance finie, et les b_i et les e_{ij} sont indépendants. Notons que la loi de y_{ij} peut varier en fonction de (i, j).

Soit δ_{ij} l'indicateur de réponse pour y_{ij} ($\delta_{ij} = 1$ si y_{ij} est un répondant et $\delta_{ij} = 0$ si y_{ij} est un non-répondant). Nous adoptons l'approche décrite dans Shao et Steel (1999), c'est-à-dire que δ_{ij} est défini pour chaque unité de la population et que le mécanisme de non-réponse fait partie du modèle. Soit $\boldsymbol{\delta}_i$ le vecteur contenant δ_{ij}, $j \in S_i$, et $\boldsymbol{y}_i$ le vecteur contenant y_{ij}, $j \in S_i$. Nous émettons l'hypothèse que le mécanisme de réponse non ignorable basée sur un effet aléatoire est le suivant : pour chaque échantillon,

$$P_m(\boldsymbol{\delta}_i \mid b_i, \boldsymbol{y}_i) = P_m(\boldsymbol{\delta}_i \mid b_i), \quad i \in S, \tag{4}$$

où P_m est la probabilité sous le modèle et $P_m(\xi \mid \eta)$ dénote la loi conditionnelle de ξ sachant η. Autrement dit, sachant b_i, $\boldsymbol{y}_i$ et $\boldsymbol{\delta}_i$ sont indépendants. (Inconditionnellement, ils pourraient être dépendants.) Nous supposons que le mécanisme stochastique sous le modèle est indépendant du mécanisme d'échantillonnage de sorte que $E_s E_m(X) = E_m E_s(X)$ à condition que X soit intégrable, où E_m est l'espérance sous P_m.

En outre, nous supposons que

$$\text{pour tout } i \in S, \text{ au moins un } \delta_{ij} \text{ est égal à 1.} \tag{5}$$

Autrement dit, chaque grappe contient au moins un répondant. Sans cette hypothèse (ou une autre hypothèse), il pourrait être impossible d'estimer le total de population Y. Nous présentons une discussion plus approfondie à la section 4.

Si nous supposons que la non-réponse est ignorable, c'est-à-dire que $P_m(\delta_{ij} = 1 \mid y_{ij}) = P_m(\delta_{ij} = 1)$, alors une méthode utilisée fréquemment consiste à imputer la valeur manquante pour chaque non-répondant par la moyenne $\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij} y_{ij} / \sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}$, ce qui mène à l'estimateur de Y suivant :

$$\begin{aligned}\hat{Y}_r &= \sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij} y_{ij},\ \tilde{w}_{ij} \\ &= w_{ij}\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\Big/\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\right).\end{aligned} \tag{6}$$

Sous les hypothèses (3) à (5),

$$\begin{aligned}&E_s E_m(\hat{Y}_r) \\ &= E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij}(\mu_i + b_i + e_{ij})\right) \\ &= E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij}\mu_i\right) + E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\tilde{w}_{ij} b_i\right),\end{aligned} \tag{7}$$

où la dernière égalité découle de

$$\begin{aligned}E_m(\delta_{ij}\tilde{w}_{ij} e_{ij}) &= E_m[E_m(\delta_{ij}\tilde{w}_{ij} e_{ij} \mid b_i)] \\ &= E_m[E_m(\delta_{ij}\tilde{w}_{ij} \mid b_i) E_m(e_{ij} \mid b_i)] = 0\end{aligned} \tag{8}$$

sous (4). Le premier terme de (7) est égal à

$$E_s E_m\left[\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\mu_i\right)\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\right)\Big/\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\right)\right]$$

qui est approximativement égal à (quand n est grand)

$$\begin{aligned}&\frac{E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\mu_i\right) E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\right)}{E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij} w_{ij}\right)} \\ &= \frac{E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\mu_i E_m(\delta_{ij})\right) E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\right)}{E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij} E_m(\delta_{ij})\right)}.\end{aligned}$$

Notons que

$$E_s E_m(Y) = E_m(Y) = \sum_{i\in P}\sum_{j\in P_i} \mu_i = E_s\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\mu_i\right).$$

Donc, le fait que $\mu_i = \mu$ pour tout i ou que $E_m(\delta_{ij})$ ne dépende pas de (i, j) implique que l'espérance du premier terme de (7) est approximativement égale à l'espérance de Y. Cependant, en général $E_m(\delta_{ij}\tilde{w}_{ij}b_i) \neq 0$, parce que δ_{ij} et b_i sont dépendants. Donc, le deuxième terme de (7) n'est pas égal à 0 et, par conséquent, $\hat{Y}_r$ défini par (6) est biaisé sous le mécanisme de non-réponse non ignorable basée sur un effet aléatoire. Ce biais ne disparaît pas asymptotiquement quand $n \to \infty$ et (ou) que $m_i \to \infty$ pour tout i.

Étant donné que le biais de $\hat{Y}_r$ est dû à ce que l'imputation est effectuée sur l'échantillon complet, alors que la non-réponse dépend d'un effet aléatoire au niveau de la grappe, nous pouvons trouver un estimateur sans biais en exécutant l'imputation dans chaque grappe. Cela serait une méthode d'imputation naturelle si l'effet aléatoire de grappe b_i était observé. Si nous corrigeons une non-réponse y_{ij} dans la grappe i par imputation de la moyenne de grappe $\sum_{j\in S_i}\delta_{ij}w_{ij}y_{ij}/\sum_{j\in S_i}\delta_{ij}w_{ij}$, alors l'estimateur résultant est

$$\hat{Y}_c = \sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\overline{w}_{ij}y_{ij},$$

avec

$$\overline{w}_{ij} = w_{ij}\left(\sum_{j\in S_i} w_{ij} \Big/ \sum_{j\in S_i}\delta_{ij}w_{ij}\right). \qquad (9)$$

L'hypothèse (5) nous assure que $\overline{w}_{ij}$ est bien défini. Notons que

$$\begin{aligned} E_s E_m(\hat{Y}_c) &= E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\overline{w}_{ij}\mu_i\right) + E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\overline{w}_{ij}b_i\right) \\ &= E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\,\mu_i\right) + E_s E_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\,b_i\right) \\ &= E_m(Y), \end{aligned}$$

où la première égalité découle de l'hypothèse (3) et du fait que, sous l'hypothèse (4), le résultat (8) est encore vérifié si l'on remplace $\tilde{w}_{ij}$ par $\overline{w}_{ij}$, la deuxième égalité découle de la définition de $\overline{w}_{ij}$ et du fait que μ_i et b_i ne dépendent pas de j, et la dernière égalité découle de $E_m(b_i) = 0$. Donc, $\hat{Y}_c$ est un estimateur sans biais de Y.

Puisque l'imputation est faite dans chaque grappe, l'estimateur défini par (9) paraît inefficace lorsque certaines tailles d'échantillon de grappes m_i sont très faibles. Cependant, le problème ne se pose pas dans le cas où $w_{ij} = w_i$ pour tout j (par exemple, si les probabilités de sélection sont égales lors de l'échantillonnage de deuxième degré). Quand $w_{ij} = w_i$ pour tout j, l'imputation menant à $\hat{Y}_c$ dans (9) est, en fait, effectuée dans une beaucoup plus grande classe, c'est-à-dire un groupe de grappes ayant une caractéristique en commun. Soit $\overline{\delta}_i = m_i^{-1}\sum_{j\in S_i}\delta_{ij}$ le taux de réponse dans la grappe i et soit

$$G_l = \left\{i \in S : m_i = m, \overline{\delta}_i = k/m\right\}, \quad l = (k, m), k \le m. \qquad (10)$$

Pour chaque $l = (k, m)$, G_l donné par (10) est le groupe de grappes échantillonnées ayant les mêmes $m_i = m$ et $\overline{\delta}_i = k$. Si $w_{ij} = w_i$ pour tout j, alors, pour $i \in G_l$ avec $l = (k, m)$,

$$\begin{aligned} \overline{w}_{ij} &= w_{ij}\left(\sum_{j\in S_i} w_{ij} \Big/ \sum_{j\in S_i}\delta_{ij}w_{ij}\right) \\ &= w_i\left(\sum_{j\in S_i} w_i \Big/ \sum_{j\in S_i}\delta_{ij}w_i\right) \\ &= w_i/\overline{\delta}_i \\ &= w_i/(k/m) \\ &= w_i\left(\sum_{i\in G_l} m_i w_i\right) \Big/ \left(\sum_{i\in G_l}\frac{k}{m}m_i w_i\right) \\ &= w_i\left(\sum_{i\in G_l} m_i w_i\right) \Big/ \left(\sum_{i\in G_l}\overline{\delta}_i m_i w_i\right) \\ &= w_i\left(\sum_{i\in G_l}\sum_{j\in S_i} w_i\right) \Big/ \left(\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_i\right) \\ &= w_{ij}\left(\sum_{i\in G_l}\sum_{j\in S_i} w_{ij}\right) \Big/ \left(\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}\right). \end{aligned}$$

Par conséquent, l'imputation donnant $\hat{Y}_c$ dans (9) est, en fait, effectuée dans chaque groupe G_l quand $w_{ij} = w_i$ pour tout j, c'est-à-dire que la valeur pour un non-répondant dans S_i est imputée par la moyenne d'échantillon pour les répondants dans G_l, $\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}y_{ij}/\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}$.

Si w_{ij} varie en fonction de j pour un i donné, certaines conditions supplémentaires sont nécessaires afin de combiner les grappes. Ce point est discuté à la section 4.

Nous terminons la présente section par une discussion de l'estimation de la variance, puisque pour la plupart des sondages, un estimateur de variance est requis pour chaque estimateur ponctuel. Nous pouvons dériver une formule de la variance ou son approximation (quand $n \to \infty$) pour $\hat{Y}_c$ qui pourrait nécessiter plus de renseignements sur le plan d'échantillonnage. Quand la taille de l'échantillon de premier degré n est grande, que $m_i \le m$ pour tout i et un nombre entier fixé m, et que n/N est faible, où N est la taille de P, nous pouvons appliquer la méthode corrigée du jackknife décrite dans Rao et Shao (1992). Plus précisément, nous pouvons procéder aux étapes suivantes.

1. Créer n répliques jackknife, où la $i^{\text{ième}}$ réplique est obtenue en supprimant la $i^{\text{ième}}$ grappe et en rajustant les poids à $w_{kj}^{(i)}, k \neq i, i = 1, \ldots, n,$ conformément au plan d'échantillonnage. Par exemple, si l'échantillonnage de premier degré est stratifié, alors $w_{kj}^{(i)} = w_{kj}$ si k et i ne sont pas dans la même strate et $w_{kj}^{(i)} = n_h w_{kj}/(n_h - 1)$ si k et i

sont dans la même strate h, où n_h est la taille de la strate.

2. Réimputer les valeurs pour les non-répondants dans la $i^{\text{ième}}$ réplique jackknife en utilisant les valeurs des répondants dans la $i^{\text{ième}}$ réplique jackknife, $i = 1, \ldots, n$.

3. Calculer $\hat{Y}_{c,i}$ de la même façon que $\hat{Y}_c$ mais en se fondant sur la $i^{\text{ième}}$ réplique jackknife réimputée, $i = 1, \ldots, n$.

4. Calculer l'estimateur par le jackknife de la variance pour $\hat{Y}_c$ en utilisant une formule du jackknife standard (par exemple, Shao et Tu 1995, chapitre 6). Par exemple, si l'échantillonnage de premier degré est stratifié et comporte H strates, alors l'estimateur par le jackknife de la variance est

$$v = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i \in S_h} \left(\hat{Y}_{c,i} - \frac{1}{n} \sum_{k \in S} \hat{Y}_{c,k} \right)^2,$$

où S_h est l'échantillon provenant de la $h^{\text{ième}}$ strate et n_h est la taille de S_h.

3. Résultats des simulations

Nous présentons maintenant les résultats d'une étude par simulation effectuée en vue d'examiner les propriétés des estimateurs $\hat{Y}_r$ et $\hat{Y}_c$.

Nous créons une population finie semblable à la population d'instituteurs et institutrices du comté de Maricopa, en Arizona (Lohr 1999, pages 446-447). La population finie contient 311 grappes (écoles). Dans chaque grappe, les unités de deuxième degré sont les instituteurs et institutrices. La taille de la grappe (le nombre d'instituteurs et institutrices) varie de 6 à 59, si bien que l'échantillonnage de premier degré est un échantillonnage avec probabilités inégales proportionnelles à la taille de la grappe. L'échantillonnage de premier degré est fait avec remise et la taille de l'échantillon est 31. L'échantillonnage de deuxième degré est un échantillonnage aléatoire simple sans remise de taille 6 (pour toute grappe).

Pour chaque instituteur ou institutrice, la variable d'intérêt est le nombre de minutes par semaine consacrées aux travaux de préparation à l'école. Les valeurs de y_{ij} pour cette variable dans la simulation sont générées conformément au modèle (3), où μ_i est le nombre moyen de minutes consacrées par semaine aux travaux de préparation à l'école pour la $i^{\text{ième}}$ école, b_i est un effet aléatoire de la $i^{\text{ième}}$ école, et e_{ij} est un effet aléatoire du ou de la $j^{\text{ième}}$ instituteur ou institutrice dans la $i^{\text{ième}}$ école. Les valeurs de μ_i sont les moyennes d'échantillon de l'ensemble de données de Lohr (1999, pages 446-447), qui varient de 25,52 à 42,18 avec une moyenne de 33,76 et une médiane de 33,47. La valeur de b_i est générée conformément à $b_i = 8,31(X_i - 2)$, où X_i suit la loi gamma dont le paramètre de forme est 2 et le paramètre d'échelle est 1. La valeur de e_{ij} est générée à partir de la loi normale de moyenne 0 et d'écart-type 2,27. Les b_i et les e_{ij} sont générés indépendamment les uns des autres. Les valeurs de $y_{ij} = \mu_i + b_i + e_{ij}$ sont générées dans chaque exécution de la simulation afin de pouvoir évaluer le biais et les erreurs-types des estimateurs en utilisant la probabilité conjointe sous le plan d'échantillonnage et les modèles (3) à (5).

Pour les unités échantillonnées, les non-répondants sont générés conformément à (4) et (5). Autrement dit, chaque grappe échantillonnée contient un répondant et la situation de réponse des autres unités échantillonnées dans chaque grappe est déterminée indépendamment par $P(y_{ij}$ manque $\mid b_i) = e^{b_i - 1}/(1 + e^{b_i - 1})$. La probabilité moyenne de non-réponse est de 33,76 %.

Pour l'estimation de la moyenne de population finie, une simulation comptant 1 000 passages machine montre que, si l'on utilise $\hat{Y}_r$ le biais, l'erreur-type et la racine de l'erreur quadratique moyenne valent -2,89, 1,32 et 3,17, respectivement, et que le biais relatif $E(\hat{Y}_r - Y)/E(Y)$ est de -8,5 %; si l'on utilise $\hat{Y}_c$, le biais, l'erreur-type et la racine de l'erreur quadratique moyenne valent 0,12, 1,81 et 1,82, respectivement et le biais relatif $E(\hat{Y}_c - Y)/E(Y)$ est de 0,3 %. Ce résultat de simulation corrobore notre théorie, c'est-à-dire que $\hat{Y}_c$ est approximativement sans biais, mais que $\hat{Y}_r$ est biaisé. Dans ce cas, l'erreur-type de $\hat{Y}_c$ est plus grande que celle de $\hat{Y}_r$, mais la racine de l'erreur quadratique moyenne de $\hat{Y}_r$ est nettement plus grande que celle de $\hat{Y}_c$ à cause de son biais important.

4. Discussion

Si nous n'émettons pas l'hypothèse que chaque grappe échantillonnée contient au moins un répondant, il pourrait être impossible d'estimer le total de population, à moins d'ajouter une autre hypothèse. Sous le mécanisme de non-réponse (4), quand toutes les observations dans une grappe sont des non-réponses, aucune information dans cette grappe ne peut être recouvrée d'après les données observées dans d'autres grappes, à moins que soit émise une hypothèse supplémentaire. Par exemple, on pourrait supposer que la population de grappes ne contenant aucun répondant est semblable à celle des grappes contenant un répondant, auquel cas on pourrait regrouper les grappes en répartissant les poids de celles ne contenant aucun répondant de la même façon que les poids de celles contenant un répondant. Une autre approche consiste à utiliser un modèle hypothétique permettant d'extrapoler les résultats aux grappes ne contenant pas de répondant.

Les résultats de la section 2 sont présentés pour l'imputation par la moyenne. Les extensions à certaines autres méthodes d'imputation sont simples. Par exemple, si l'on considère l'imputation hot deck aléatoire, notre résultat mène à l'imputation dans les grappes (ou les G_l). S'il existe une covariable x dont les valeurs sont toutes observées, notre résultat peut être étendu à l'imputation par la régression en utilisant le modèle (3) modifié pour donner $y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij}$. Dans le cas de la non-éponse partielle, notre résultat peut également être appliqué à la repondération, c'est-à-dire à l'ajustement des poids dans les grappes (ou les G_l).

Notre méthode est basée sur un modèle d'imputation. Nous utilisons le modèle à effet aléatoire hypothétique (3) et le mécanisme de réponse basé sur les effets aléatoires hypothétiques (4). Si le modèle (4) n'est pas vérifié, alors $E_m(\delta_{ij}\tilde{w}_{ij}e_{ij}) \neq 0$ et notre estimateur $\hat{Y}_c$ a un biais dont la grandeur dépend de la taille de $|E_m(\delta_{ij}\tilde{w}_{ij}e_{ij})|$. De même, $\hat{Y}_c$ n'est pas valide si le modèle (3) ne tient pas.

Nous avons montré à la section 2 qu'en imposant la condition $w_{ij} = w_i$ pour tout j, nous nous assurons que l'imputation est faite dans chaque G_l qui est le groupe de grappes ayant les mêmes taille et taux de réponse. Dans le cas de l'échantillonnage à deux degrés, cette condition est satisfaite quand l'échantillonnage de deuxième degré est réalisé avec probabilités égales (par exemple, l'échantillonnage aléatoire simple sans remise). Dans le cas de l'échantillonnage à trois degrés, le modèle (3) doit être remplacé par $y_{ijk} = \mu_{ij} + b_{ij} + e_{ijk}$ et b_i dans (4), par b_{ij}. Le poids de sondage w_{ijk} satisfait $w_{ijk} = w_{ij}$ à condition que l'échantillonnage de dernier degré soit exécuté avec probabilités égales et que notre résultat soit encore vérifié. En cas d'échantillonnage à deux degrés avec w_{ij} variant en fonction de j, nous pouvons procéder à l'imputation dans un groupe de grappes qui ont la même $E_m(\boldsymbol{y}_i \mid \boldsymbol{\delta}_i)$. Par exemple, supposons qu'en plus des conditions (3) à (5), les $\mu_i = \mu$, les b_i sont indépendants et identiquement distribués (iid) et, sachant b_i, les composantes de $\boldsymbol{\delta}_i$ sont iid. Alors $E_m(b_i \mid \boldsymbol{\delta}_i) = E_m(b_i \mid \bar{\delta}_i)$ dépend uniquement de la taille de la grappe m_i et de $\bar{\delta}_i$. Donc, nous pouvons procéder à l'imputation dans chaque G_l défini par (10).

Remerciements

Ce travail a été financé partiellement par la subvention CA53786 du NCI et par la subvention DMS-0404535 de la NSF. L'auteur remercie M. Lei Xu de la programmation de l'étude par simulation et deux examinateurs de leurs commentaires constructifs.

Bibliographie

Lee, H., Rancourt, E., et Särndal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Duxbury Press, New York.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Shao, J., et Steel, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Wu, M.C., et Carroll, R.J. (1988). Estimation and comparisons of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175-188.

Techniques d'enquête, juin 2007
Vol. 33, N° 1, pp. 99-107
Statistique Canada, N° 12-001 au catalogue

Plan d'échantillonnage proportionnel à la taille le plus proche contrôlé optimal

Neeraj Tiwari, Arun Kumar Nigam et Ila Pant [1]

Résumé

Le concept de « plan d'échantillonnage proportionnel à la taille le plus proche » proposé par Gabler (1987) est utilisé en vue d'obtenir un plan d'échantillonnage contrôlé optimal assurant que les probabilités de sélection des échantillons non privilégiés soient nulles. L'estimation de la variance pour un plan d'échantillonnage contrôlé optimal à l'aide de la forme de Yates-Grundy de l'estimateur d'Horvitz-Thompson est discutée. La variance d'échantillonnage réelle de la méthode proposée est comparée à celle des méthodes existantes de sélection contrôlée et non contrôlée sous grande entropie. L'utilité de la méthode proposée est démontrée au moyen d'exemples.

Mots clés : Échantillonnage contrôlé; échantillons non privilégiés; programmation quadratique; variance sous grande entropie.

1. Introduction

Dans de nombreuses situations, certains échantillons peuvent être indésirables, à cause de complications administratives, de l'éloignement, de la similarité des unités ou de questions de coût. Les échantillons de ce genre sont qualifiés de non privilégiés et la méthode pour les éviter est appelée « sélection contrôlée » ou « échantillonnage contrôlé ». Cette méthode, proposée pour la première fois par Goodman et Kish (1950), a suscité beaucoup d'intérêt ces dernières années à cause de son importance pratique.

La méthode d'échantillonnage contrôlé est la plus appropriée lorsque des considérations financières ou autres obligent à sélectionner un petit nombre de grandes unités primaires d'échantillonnage, comme des hôpitaux, des entreprises ou des écoles, en vue de leur inclusion dans l'étude. L'objectif principal de l'échantillonnage contrôlé est d'accroître la probabilité de sélectionner une combinaison privilégiée de sorte qu'elle soit supérieure à celle possible par échantillonnage stratifié, tout en maintenant les probabilités de sélection initiales des unités de la population, donc en préservant la propriété d'échantillon probabiliste. Cette situation se présente généralement dans le cas d'enquêtes sur le terrain, où la sélection de certaines unités est indésirable pour des raisons pratiques, mais un échantillonnage probabiliste est nécessaire. Des contraintes peuvent être imposées afin d'assurer que la répartition géographique ou autre des unités soit appropriée et que la taille de l'échantillon soit adéquate pour certains sous-groupes de la population. Goodman et Kish (1950) ont considéré la réduction de la variance d'échantillonnage des estimations clés comme étant l'objectif principal de la sélection contrôlée, mais ont aussi mis en garde contre le fait que cela n'est peut-être pas toujours réalisable. Ils ont aussi discuté d'un problème réel destiné à mettre l'accent sur la nécessité d'utiliser des contraintes au-delà de la stratification (Goodman et Kish 1950, page 354) dans le but de sélectionner 21 unités primaires d'échantillonnage pour représenter les États du Centre-Nord. Hess et Srikantan (1966) ont utilisé les données sur l'univers de 1961 des hôpitaux généraux de soins de courte durée non fédéraux des États-Unis pour illustrer les applications des formules d'estimation et de calcul de la variance au cas de la sélection contrôlée. Waterton (1983) a utilisé les données provenant d'une enquête par la poste sur les sortants des écoles écossaises réalisée en 1977 pour décrire les avantages de la sélection contrôlée et comparer l'efficacité de cette dernière à celle de l'échantillonnage aléatoire stratifié proportionnel multiple (c'est-à-dire le plan d'échantillonnage dans lequel, au lieu d'une variable de stratification unique, ou utilise de nombreuses variables individuellement associées à la variable d'intérêt y par classification croisée de la population en fonction de ces variables) et a constaté que la sélection contrôlée donnait des résultats favorables.

Trois approches distinctes ont été proposées dans la littérature récente en vue de mettre en œuvre l'échantillonnage contrôlé, à savoir i) l'utilisation de configurations typiques de plan d'expérience, ii) la méthode du vidage de boîtes et iii) le recours à la programmation linéaire. Bien que certains chercheurs soient partis de plans d'échantillonnage aléatoire simple pour construire des plans d'échantillonnage contrôlé, l'une des stratégies les plus répandues consiste à conjuguer l'échantillonnage avec probabilité d'inclusion proportionnelle à la taille (PIPT) et l'estimateur d'Horvitz-Thompson (1952). Pour construire des plans d'échantillonnage aléatoire simple contrôlé, Chakrabarti (1963), ainsi qu'Avadhani et Sukhatme (1973) ont proposé d'utiliser des plans en blocs incomplets équilibrés (BIE) ayant pour

1. Neeraj Tiwari, Ila Pant, Département de statistiques, Université Kumaon, campus S.S.J., Almora-263601, Inde. Courriel : kumarn_amo@yahoo.com; Arun Kumar Nigam, Institut de statistique appliquée et d'études sur le développement, Lucknow-226017, Inde. Courriel : dr_aknigam@yahoo.com.

paramètres $v = N$, $k = n$ et λ, où N est la taille de population et n, la taille d'échantillon. Wynn (1977) ainsi que Foody et Hedayat (1977) ont utilisé les plans BIE avec blocs répétés dans des situations où des plans BIE non triviaux n'existent pas. Gupta, Nigam et Kumar (1982) ont étudié des plans d'échantillonnage contrôlé avec probabilité d'inclusion proportionnelle à la taille et utilisé des plans BIE conjugués à l'estimateur d'Horvitz-Thompson du total de population $Y(=\sum_{i=1}^{N} y_i$, où y_i est la valeur de la $i^{\text{ième}}$ unité de population, U). Nigam, Kumar et Gupta (1984) ont utilisé certaines configurations de divers types de plans expérimentaux, y compris des plans BIE, pour obtenir des plans d'échantillonnage PIPT contrôlé ayant la propriété supplémentaire que $c\pi_i\pi_j \le \pi_{ij} \le \pi_i\pi_j$ pour tout $i \ne j = 1, ..., N$ et une constante positive donnée c telle que $0 < c < 1$, où π_i et π_{ij} représentent les probabilités d'inclusion de premier et de deuxième ordres, respectivement. Hedayat et Lin (1980), ainsi que Hedayat, Lin et Stufken (1989) ont utilisé la méthode du « vidage des boîtes » pour construire des plans d'échantillonnage PIPT contrôlé ayant la propriété supplémentaire que $0 < \pi_{ij} \le \pi_i\pi_j$, $i < j = 1, ..., N$. Srivastava et Saleh (1985), ainsi que Mukhopadhyay et Vijayan (1996) ont proposé de remplacer l'échantillonnage aléatoire simple sans remise (EASSR) par des « t-designs » pour construire des plans d'échantillonnage contrôlé.

Toutes les méthodes d'échantillonnage contrôlé mentionnée dans le paragraphe qui précède peuvent être appliquées manuellement avec divers degrés de difficulté, mais aucune n'exploite les avantages de l'informatique moderne. Recourant à la méthode du simplex en programmation linéaire, Rao et Nigam (1990, 1992) ont proposé des plans d'échantillonnage contrôlé optimal qui réduisent au minimum la probabilité de sélectionner les échantillons non privilégiés, tout en retenant certaines propriétés d'un plan non contrôlé connexe. En suivant l'approche de Rao et Nigam (1990, 1992), Sitter et Skinner (1994), ainsi que Tiwari et Nigam (1998) ont utilisé la méthode du simplex en programmation linéaire pour résoudre des problèmes de stratification multidimensionnelle avec des « contraintes allant au-delà de la stratification ».

Dans le présent article, nous utilisons la programmation quadratique pour proposer un plan d'échantillonnage contrôlé optimal qui assure que la probabilité de sélectionner les échantillons non privilégiés soit exactement égale à zéro, au lieu de la minimiser, sans sacrifier l'efficacité de l'estimateur d'Horvitz-Thompson fondé sur un plan d'échantillonnage PIPT non contrôlé connexe. Nous utilisons la notion de « plan d'échantillonnage proportionnel à la taille le plus proche » introduite par Gabler (1987) pour construire le plan proposé. Nous nous servons du Solveur Microsoft Excel du progiciel Microsoft Office 2000 pour résoudre le problème de programmation quadratique. Nous discutons de l'applicabilité de l'estimateur d'Horvitz-Thompson au plan proposé. Nous comparons empiriquement la variance d'échantillonnage réelle de l'estimation pour le plan proposé aux variances pour les plans contrôlés optimaux avancés par Rao et Nigam (1990, 1992) et pour les méthodes de sélection non contrôlée sous grande entropie proposées par Goodman et Kish (1950), ainsi que par Brewer et Donadio (2003). À la section 3, nous examinons certains exemples en vue de démontrer l'utilité de la méthode que nous proposons par comparaison des probabilités de sélection des échantillons non privilégiés et des variances d'échantillonnage des estimations. Enfin, à la section 4, nous résumons les résultats présentés dans l'article.

2. Le plan d'échantillonnage contrôlé optimal

À la présente section, nous nous fondons sur le concept de « plan d'échantillonnage avec probabilité de sélection proportionnelle à la taille le plus proche » pour proposer un plan d'échantillonnage PIPT contrôlé qui produit des probabilités concordant avec les valeurs originales de π_i, satisfait la condition suffisante $\pi_{ij} \le \pi_i\pi_j$ pour que la forme de Yates-Grundy (1953) de l'estimateur d'Horvitz-Thompson (HT) (1952) de la variance soit non négative et assure en outre que la probabilité de sélection des échantillons non privilégiés soit exactement égale à zéro. Avant de discuter du plan proposé, nous décrivons brièvement les plans PIPT de Misdzuno-Sen et de Sampford que nous utiliserons dans le plan proposé pour obtenir le plan PIPT initial $p(s)$.

2.1 Les plans PIPT de Midzuno-Sen et de Sampford

Afin d'introduire le concept des plans PIPT, nous supposons qu'une quantité positive connue, x_i, est associée à la valeur de la $i^{\text{ième}}$ unité de la population et qu'il existe une raison de croire que les valeurs de y_i sont approximativement proportionnelles aux x_i. Ici, nous supposons que la valeur de x_i est connue pour toutes les unités de la population et que les valeurs de y_i doivent être recueillies pour toutes les unités échantillonnées. Dans le cas des plans d'échantillonnage PIPT, π_i, la probabilité d'inclusion de la $i^{\text{ième}}$ unité dans un échantillon de taille n, est égale à np_i, où p_i est la probabilité de sélection en un seul tirage de la $i^{\text{ième}}$ unité de la population (également appelée mesure de taille normale de l'unité i) donnée par

$$p_i = \frac{x_i}{\sum_{j=1}^{N} x_j},\ i = 1, 2, ..., N.$$

Nous commençons par décrire le plan PIPT de Midzuno-Sen, puis nous discutons du plan de Sampford.

Le plan de Midzuno-Sen (MS) (1952, 1953) a pour contrainte que les probabilités de sélection de la $i^{\text{ième}}$ unité de la population (p_i) doivent satisfaire la condition

$$\frac{1}{n}.\frac{n-1}{N-1} \le p_i \le \frac{1}{n}, \qquad i=1,\ 2,\ ...,\ N. \quad (1)$$

Si (1) est satisfaite pour les valeurs de p_i étudiées, nous appliquons le plan de MS pour obtenir un plan PIPT dont les probabilités de sélection sont révisées, p_i^*, [également appelées mesures révisées de la taille normale] données par

$$p_i^* = n\,p_i\,.\,\frac{N-1}{N-n} - \frac{n-1}{N-n}, \qquad i=1,\ 2,\ ...,\ N. \quad (2)$$

Maintenant, en supposant que le $s^{\text{ième}}$ échantillon est constitué des unités $i_1,\ i_2,\ ...,\ i_n$, la probabilité d'inclure ces unités dans le $s^{\text{ième}}$ échantillon sous le plan de MS est donnée par

$$p(s) = \pi_{i_1,\,i_2,\,...,\,i_n} = \frac{1}{\binom{N-1}{n-1}}\ (p_{i_1}^* + p_{i_2}^* + \ ...\ + p_{i_n}^*). \quad (3)$$

Toutefois, à cause de la contrainte (1), le plan de MS limite l'applicabilité de la méthode à des unités de taille relativement semblable. Par conséquent, lorsque les probabilités initiales ne satisfont pas la condition du plan de MS, nous proposons d'utiliser le plan de Sampford (1967) pour obtenir le plan PIPT initial $p(s)$.

En utilisant le plan de Sampford, la probabilité d'inclure n unités $i_1,\ i_2,\ ...,\ i_n$ dans le $s^{\text{ième}}$ échantillon est donnée par

$$p(s) = \pi_{i_1,\,i_2,\,...,\,i_n} = n\,K_n\,\lambda_{i_1},\,\lambda_{i_2},\,...,\,\lambda_{i_n}\,(1-\sum_{u=1}^{n} p_{i_u}), \quad (4)$$

où $K_n = (\sum_{t=1}^{n} t L_{n-t}/n^t)^{-1}$, $\lambda_i = p_i/(1-p_i)$ pour un ensemble $S(m)$ de $m \le N$ unités différentes, $i_1,\ i_2,\ ...,\ i_m$, et L_m est définie comme étant

$$L_0 = 1,\ L_m = \sum_{S(m)} \lambda_{i_1},\,\lambda_{i_2},\,...,\,\lambda_{i_m} \quad (1 \le m \le N).$$

2.2 Le plan proposé

Considérons une population de N unités. Supposons que l'on doive sélectionner un échantillon de taille n à partir de cette population. Les probabilités de sélection par tirage unique de ces N unités de la population (valeurs de p_i) sont connues. Soit S et S_1 l'ensemble de tous les échantillons possibles et l'ensemble des échantillons non privilégiés, respectivement.

Sachant les probabilités de sélection pour les N unités de la population, nous obtenons d'abord un plan PIPT non contrôlé approprié $p(s)$, tel que le plan de Midzuno-Sen (1952, 1953) ou celui de Sampford (1967) qui sont décrits à la section 2.1. Après avoir obtenu le plan PIPT initial $p(s)$, l'idée qui sous-tend le plan proposé est de se débarrasser des échantillons non privilégiés S_1 en se limitant à l'ensemble $S - S_1$ grâce à l'introduction d'un nouveau plan $p_0(s)$ qui attribue une probabilité nulle de sélection à chaque échantillon non privilégié appartenant à S_1. Ce plan est donné par

$$p_0(s) = \begin{cases} \dfrac{p(s)}{1-\sum_{s\in S_1} p(s)} & \text{pour } s \in S - S_1 \\ 0 & \text{autrement,} \end{cases} \quad (5)$$

où $p(s)$ est le plan d'échantillonnage PIPT non contrôlé initial.

Conséquemment, $p_0(s)$ n'est plus un plan PIPT. Donc, en appliquant l'idée de Gabler (1987), nous nous intéressons au « plan d'échantillonnage proportionnel à la taille le plus proche » $p_1(s)$ en ce sens que $p_1(s)$ minimise la distance directe D entre le plan d'échantillonnage $p_0(s)$ et le plan d'échantillonnage $p_1(s)$ définie comme étant

$$D(p_0,\ p_1) = E_{p_0}\left[\frac{p_1}{p_0} - 1\right]^2 = \sum_{s\in S-S_1} \frac{p_1^{\,2}(s)}{p_0(s)} - 1 \quad (6)$$

sous les contraintes suivantes :

$$\begin{aligned} &\text{(i)} && p_1(s) \ge 0, \\ &\text{(ii)} && \sum_{s\in S-S_1} p_1(s) = 1, \\ &\text{(iii)} && \sum_{s\ni i} p_1(s) = \pi_i, \\ &\text{(iv)} && \sum_{s\ni i,\,j} p_1(s) > 0 \quad \text{et} \\ &\text{(v)} && \sum_{s\ni i,\,j} p_1(s) \le \pi_i\,\pi_j. \end{aligned} \quad (7)$$

Le classement de ces cinq contraintes est établi en fonction de leur degré de nécessité et de désirabilité. Les contraintes (i) et (ii) sont nécessaires pour tout plan d'échantillonnage probabiliste. La contrainte (iii), qui impose que les probabilités de sélection soient les mêmes dans l'ancien et le nouveau plan d'échantillonnage, assure que le plan résultant sera PIPT. Cette contrainte est très forte et influe dans une large mesure sur les propriétés de convergence du plan proposé. La contrainte (iv) est hautement souhaitable, parce qu'elle assure l'estimation sans biais de la variance. La contrainte (v) est souhaitable, car elle assure que soit respectée la condition suffisante pour la non-négativité de l'estimateur de la variance de Yates-Grundy.

La solution du problème de programmation quadratique susmentionné, c'est-à-dire minimiser la fonction objectif (6)

sous les contraintes (7), nous donne le plan d'échantillonnage PIPT contrôlé optimal assurant que la probabilité de sélection soit nulle pour les échantillons non privilégiés. Le plan proposé est aussi proche que possible du plan contrôlé $p_0(s)$ défini en (5) tout en produisant le même ensemble de probabilités d'inclusion de premier ordre π_i que le plan d'échantillonnage PIPT non contrôlé original $p(s)$. Étant donné les contraintes (iv) et (v) dans (7), le plan proposé garantit également que soient satisfaites les conditions $\pi_{ij} > 0$ et $\pi_{ij} \leq \pi_i \pi_j$ assurant que l'estimateur de Yates-Grundy de la variance soit stable et non négatif.

La mesure de distance $D(p_0, p_1)$ définie en (6) est semblable à la statistique χ^2 souvent utilisée dans des problèmes apparentés, et a également été utilisée par Cassel et S≅rndal (1972) et par Gabler (1987). D'autres mesures de distance sont discutées par Takeuchi, Yanai et Mukherjee (1983). Dans le contexte de la présente discussion, nous pourrions aussi définir une autre mesure de distance de la forme

$$D(p_0, p_1) = \sum_s \frac{(p_0 - p_1)^2}{(p_0 + p_1)}. \qquad (8)$$

Lorsqu'elle est appliquée aux divers problèmes numériques que nous considérons, nous constatons que l'équation (8) donne des résultats comparables à (6) en ce qui a trait à la convergence et à l'efficacité, si bien que nous présenterons les résultats obtenus en utilisant (6) comme mesure de distance.

Alors que tous les plans d'échantillonnage contrôlé discutés par les auteurs antérieurs avaient pour objectif de minimiser les probabilités de sélection des échantillons non privilégiés, celui que nous proposons exclut entièrement la possibilité de sélectionner ces échantillons en garantissant que leurs probabilités de sélection soient nulles tout en assurant la non-négativité de l'estimateur de la variance de Yates-Grundy. Cependant, dans certaines situations, il se pourrait qu'aucune solution faisable du problème de programmation quadratique, satisfaisant toutes les contraintes énoncées en (7), n'existe. Le cas échéant, la contrainte (v) peut être relâchée. La non-négativité de la forme de Yates-Grundy de l'estimateur de la variance risque alors de ne plus être garantie. Cependant, puisque la condition $\pi_{ij} \leq \pi_i \pi_j$ est suffisante pour que cet estimateur soit non négatif, mais qu'elle n'est pas nécessaire pour $n > 2$, comme l'a souligné Singh (1954), il existera encore une possibilité d'obtenir un estimateur non négatif de la variance. Après relâchement de la contrainte (v) en (7), si l'estimateur de la variance de Yates-Grundy est négatif, un autre estimateur de la variance peut être utilisé. Nous le démontrons à l'aide de l'exemple 5 à la section 3. Si, même après le relâchement de la contrainte (v), une solution faisable du problème de programmation quadratique est introuvable, on peut aussi relâcher la contrainte (iv) et, par conséquent, utiliser un autre estimateur de la variance à la place de la forme de Yates-Grundy de l'estimateur HT de la variance. L'effet du relâchement de ces contraintes sur l'efficacité du plan proposé est difficile à étudier, car après le relâchement de la contrainte de non-négativité (v), l'estimateur de la variance de Yates-Grundy ne fournit pas de résultats exacts. Lorsqu'on utilise cet estimateur, pour certains problèmes, l'estimation de la variance est plus faible après le relâchement de la contrainte (v) [comme dans le cas des exemples 2(a), 2(b) et 3(a) à la section 3], tandis que pour d'autres, elle est plus grande [comme dans le cas des exemples 1(a), 1(b), 3(b), 4(a) et 4(b) à la section 3]. Ce relâchement d'une contrainte donnant lieu à un accroissement de l'estimation de la variance pourrait être dû à l'incapacité de la forme de Yates-Grundy de l'estimateur de la variance d'estimer correctement la variance d'échantillonnage réelle, lorsque la condition de non-négativité n'est pas satisfaite.

La méthode proposée peut aussi être considérée comme supérieure aux méthodes plus anciennes de sélection contrôlée optimale, car elle consiste à imposer que la probabilité de sélection de certains échantillons soit nulle, au lieu d'associer un coût à chaque échantillon, puis à essayer de minimiser le coût, comme cela a été fait lors des approches antérieures de sélection contrôlée. La technique de sélection contrôlée appliquée par les auteurs antérieurs était une approche grossière consistant à attribuer un coût très élevé à certains échantillons et un coût très faible à d'autres.

L'une des limites du plan proposé est qu'il devient impossible à appliquer lorsque $\binom{N}{n}$ est très grand, car l'énumération de tous les échantillons possibles et la formation de la fonction objective et des contraintes deviennent assez fastidieuses. Cette limite existe aussi pour l'approche optimale de Rao et Nigam (1990, 1992) et d'autres approches d'échantillonnage contrôlé discutées à la section 1. Cependant, grâce aux systèmes informatiques plus rapides et aux progiciels statistiques modernes, l'utilisation de la méthode proposée dans le cas de populations moyennement grandes ne devrait pas être trop difficile. Sur la base des tailles de population que nous avons considérées pour l'évaluation empirique, nous constatons que la méthode proposée permet de traiter facilement les problèmes de sélection contrôlée pour une population allant jusqu'à 12 unités et un échantillon allant jusqu'à 5 unités. La méthode proposée peut être utilisée pour sélectionner un petit nombre d'unités de premier degré dans chacune d'un grand nombre de strates. Cela comprend la résolution d'une série de problèmes de programmation quadratique, ayant chacun une taille raisonnable, à condition que l'ensemble d'échantillons non privilégiés soit spécifié séparément dans chaque strate.

Comme dans le cas de la programmation linéaire, il n'existe aucune garantie de convergence d'un problème de programmation quadratique. Kuhn et Tucker (1951) ont établi certaines conditions nécessaires pour obtenir la solution optimale d'un algorithme de programmation quadratique, mais il n'existe aucune condition suffisante pour la convergence. Par conséquent, à moins que les conditions de Kuhn-Tucker soient satisfaites d'avance, il n'existe aucun moyen de vérifier si un algorithme de programmation quadratique converge vers un optimum absolu (global) ou relatif (local). En outre, il n'existe aucun moyen de prédire si la solution d'un problème de programmation quadratique existe ou non.

2.3 Comparaison de la variance d'échantillonnage de l'estimation

Pour estimer la moyenne de population $\bar{Y}(=N^{-1}\sum_{i=1}^{N} y_i)$ fondée sur un échantillon s de taille n, nous utilisons l'estimateur HT de $\bar{Y}$ défini comme étant

$$\hat{\bar{Y}}_{\mathrm{HT}} = \sum_{i\in s} \frac{Y_i}{N\pi_i}. \qquad (9)$$

Sen (1953), ainsi que Yates et Grundy (1953) ont montré indépendamment que, pour des plans d'échantillonnage à taille fixe, la variance de $\hat{\bar{Y}}_{\mathrm{HT}}$ est donnée par

$$V(\hat{\bar{Y}}_{\mathrm{HT}}) = \frac{1}{N^2}\sum_{i<j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2, \qquad (10)$$

et un estimateur sans biais de $V(\hat{\bar{Y}}_{\mathrm{HT}})$ est donné par

$$\hat{V}(\hat{\bar{Y}}_{\mathrm{HT}}) = \frac{1}{N^2}\sum_{i<j=1}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2. \qquad (11)$$

La contrainte (v), quand elle est utilisée dans le plan proposé, assure la non-négativité de l'estimateur de la variance (11).

Pour démontrer l'utilité de la méthode proposée, nous utilisons des exemples empiriques donnés à la section 3 pour comparer la variance d'échantillonnage réelle de l'estimateur HT pour la méthode proposée obtenue grâce à (10) aux variances de l'estimateur HT lors de l'utilisation du plan contrôlé optimal de Rao et Nigam (1990, 1992) et à celles des deux méthodes non contrôlées sous grande entropie (c'est-à-dire en l'absence de toute régularité décelable ou de tout ordonnancement dans les unités échantillonnées) de Goodman et Kish (1950) et de Brewer et Donadio (2003). Nous reproduisons ci-dessous les expressions des variances pour ces deux méthodes à grande entropie.

L'expression de la variance de $\hat{\bar{Y}}_{\mathrm{HT}}$ correcte jusqu'à l'ordre $O(N^{-2})$ en utilisant la méthode de Goodman et Kish (1950) est donnée par

$$V(\hat{\bar{Y}}_{\mathrm{HT}})_{\mathrm{GK}} = \frac{1}{nN^2}\left[\sum_{i\in U} p_i A_i^2 - (n-1)\sum_{i\in U} p_i^2 A_i^2\right] - \frac{n-1}{nN^2} \times\left[2\sum_{i\in U} p_i^3 A_i^2 - \sum_{i\in U} p_i^2 \sum_{i\in U} p_i^2 A_i^2 - 2\left(\sum_{i\in U} p_i^2 A_i\right)^2\right], \qquad (12)$$

où $A_i = Y_i / p_i - Y$, $Y = \sum_{i=1}^{N} Y_i$ et U représente la population finie de N unités.

Récemment, Brewer et Donadio (2003) ont dérivé la formule ne contenant pas π_{ij} pour la variance sous grande entropie de l'estimateur HT. Ils ont montré que les propriétés de cet estimateur de la variance, sous les conditions de grande entropie, étaient raisonnablement bonnes pour toutes les populations. Leur expression de la variance de l'estimateur HT est donnée par

$$V(\hat{\bar{Y}}_{\mathrm{HT}})_{\mathrm{BD}} = \frac{1}{N^2}\sum_{i\in U}\pi_i(1-c_i\pi_i)\left(\frac{Y_i}{\pi_i^{-1}} - \frac{Y}{n^{-1}}\right)^2, \qquad (13)$$

où $c_i = (n-1)/\{n-(2n-1)(n-1)^{-1}\pi_i + (n-1)^{-1}\sum_{k\in U}\pi_k^2\}$ pour tout $i\in U$, ce qui semble donner de meilleurs résultats que les autres valeurs de c_i qu'ils ont proposées.

3. Exemples

À la présente section, nous examinons certains exemples empiriques en vue de démontrer l'utilité de la méthode proposée et comparons cette dernière aux méthodes existantes d'échantillonnage contrôlé optimal. Nous commençons par discuter du plan PIPT de Midzuno-Sen (1952, 1953) pour démontrer notre méthode, car il est relativement facile de calculer la probabilité de tirer chaque échantillon possible sous ce plan. Cependant, si les conditions du plan de Midzuno-Sen ne sont pas satisfaites, nous démontrons que d'autres méthodes d'échantillonnage PIPT sans remise, comme celle de Sampford (1967), peuvent être utilisées pour obtenir le plan PIPT initial $p(s)$. Nous comparons aussi la variance d'échantillonnage réelle de l'estimateur HT sous le plan proposé à celle obtenue pour les méthodes existantes de sélection contrôlée optimale et de sélection non contrôlée sous grande entropie données par (12) et (13).

Exemple 1 : Considérons une population constituée de six villages, empruntée à Hedayat et Lin (1980). L'ensemble S de tous les échantillons possibles comprend 20 échantillons chacun de taille $n = 3$. Compte tenu des contraintes de déplacement, d'organisations du travail sur le terrain et de coût, Rao et Nigam (1990) ont défini les sept échantillons qui suivent comme étant des échantillons non privilégiés :

123; 126; 136; 146; 234; 236; 246

a). Les valeurs de Y_i et p_i associées aux six villages de la population sont :

Y_i :	12	15	17	24	17	19
p_i :	0,14	0,14	0,15	0,16	0,22	0,19

Puisque les valeurs de p_i satisfont la condition (1), nous appliquons le plan de MS (3) pour obtenir un plan PIPT pour lequel les mesures de taille normale révisées (les valeurs p_i^*) sont données par (2).

En appliquant la méthode décrite à la section 2 et en résolvant les problèmes de programmation quadratique résultants à l'aide du Solveur Microsoft Excel du progiciel Microsoft Office 2000, nous obtenons le plan PIPT contrôlé donné au tableau 1.

Tableau 1 Plan PIPT contrôlé optimal correspondant aux plans de Midzuono-Sen (MS) et de Sampford (SAMP) pour l'exemple 1

s	$p_1(s)$ [MS]	$p_1(s)$ [SAMP]	s	$p_1(s)$ [MS]	$p_1(s)$ [SAMP]
124	0,14	0,09	245	0,03	0,12
125	0,03	0,05	256	0,13	0,14
134	0,00	0,00	345	0,02	0,06
135	0,09	0,03	346	0,20	0,10
145	0,03	0,06	356	0,06	0,06
156	0,13	0,07	456	0,06	0,16
235	0,09	0,05			

Ce plan reproduit les valeurs originales de π_i, satisfait la condition $\pi_{ij} \le \pi_i \pi_j$ et assure que les probabilités de sélection des échantillons non privilégiés soient exactement égales à zéro. Évidemment, puisque la condition $\pi_{ij} \le \pi_i \pi_j$ est satisfaite, nous pouvons appliquer la forme de Yates-Grundy de l'estimateur de la variance de HT pour estimer la variance du plan proposé.

Nous avons également résolu l'exemple susmentionné en utilisant le plan (3) de Rao et Nigam (1990, page 809) avec les probabilités π_{ij} spécifiées d'après le plan de Sampford [que nous dénoterons RN3] et d'après leur plan (4) [que nous dénoterons RN4]. Si l'on utilise le plan RN3, la probabilité de sélection des échantillons non privilégiés (ϕ) est égale à 0,155253 et si l'on utilise le plan RN4 avec $c = 0,005$, elle est égale à zéro, alors que le plan proposé assure systématiquement que la probabilité de sélection des échantillons non privilégiés soit nulle.

La valeur de la variance d'échantillonnage réelle de l'estimateur HT $[V(\hat{\bar{Y}}_{\mathrm{HT}})]$ pour le plan proposé, le plan RN3, le plan RN4, le plan d'échantillonnage PIPT randomisé systématique de Goodman et Kish (1950) [que nous dénoterons GK] et le plan d'échantillonnage non contrôlé sous grande entropie de Brewer et Donadio (2003) [que nous dénoterons BD] sont présentés à la première ligne du tableau 2. L'examen de ce tableau montre clairement que le plan proposé donne presque la même valeur de la variance de l'estimateur HT que le plan RN4. La valeur de $V(\hat{\bar{Y}}_{\mathrm{HT}})$ pour le plan proposé est légèrement plus élevée que celle obtenue pour les plans RN3, GK et BD. Cette augmentation de la variance pourrait être acceptable, étant donné que le plan proposé élimine les échantillons indésirables.

Tableau 2 Valeurs de la variance d'échantillonnage réelle de l'estimateur HT $[V(\hat{\bar{Y}}_{\mathrm{HT}})]$ pour les plans proposés, RN3, RN4, GK et BD

$V(\hat{\bar{Y}}_{\mathrm{HT}})$	RN3	RN4	GK	BD	PLAN PROPOSÉ
Ex1(a)					
$N = 6, n = 3$	2,93	4,02	3,03	2,92	4,06
Ex 1(b)					
$N = 6, n = 3$	4,76	5,07	4,89	4,15	4,78
Ex 2(a)					
$N = 7, n = 3$	4,48	5,01	4,61	4,45	3,56
Ex 2(b)					
$N = 7, n = 3$	11,97	14,52	12,25	11,44	9,49
Ex 3(a)					
$N = 8, n = 3$	4,85	4,29	4,96	4,86	3,90
Ex 3(b)					
$N = 8, n = 3$	7,29	8,43	7,74	7,37	8,17
Ex 4(a)					
$N = 8, n = 4$	3,19	3,46	3,23	3,15	3,75
Ex 4(b)					
$N = 8, n = 4$	2,41	2,53	2,54	2,38	2,25
Ex 5					
$N = 7, n = 4$	3,08	3,93	3,12	3,07	5,10

b). Supposons maintenant que les valeurs p_i pour la population susmentionnée de six unités sont les suivantes :

p_i :	0,10	0,15	0,10	0,20	0,27	0,18

Puisque ces valeurs de p_i ne satisfont pas la condition (1) du plan de MS, nous appliquons le plan de Sampford (1967) pour obtenir le plan PIPT initial $p(s)$ en utilisant (4).

En appliquant la méthode décrite à la section 2 et en résolvant le problème de programmation quadratique résultant, nous obtenons le plan PIPT contrôlé donné au tableau 1. De nouveau, ce plan assure que la probabilité de sélection des échantillons non privilégiés soit nulle et satisfait la condition de non-négativité de la forme de

Yates-Grundy de l'estimateur de la variance de HT. Nous avons également résolu cet exemple à l'aide des plans RN3 et RN4. La valeur de ϕ est égale à 0,064135 pour la plan RN3 et nulle pour le plan RN4 avec $c = 0,005$. Le plan proposé assure systématiquement que la probabilité de sélection des échantillons non privilégiés soit nulle.

Les valeurs de $V(\hat{\bar{Y}}_{\mathrm{HT}})$ pour le plan proposé, le plan RN3, le plan RN4, le plan GK et le plan BD sont présentées à la deuxième ligne du tableau 2. Le plan proposé semble donner de meilleurs résultats que les plans RN4 et GK, et assez proches de ceux produits par les autres plans étudiés ici.

D'autres exemples ont été construits pour analyser les propriétés du plan proposé. Les populations, avec les valeurs de Y_i et p_i et l'ensemble d'échantillons non privilégiés pour chaque population, sont résumées en annexe. Les valeurs de p_i pour les exemples 2(a), 3(a) et 4(a) satisfont la condition (1) du plan de Midzuno-Sen et, donc, pour ces exemples, nous utilisons le plan PIPT de Midzuno-Sen pour obtenir le plan PIPT initial $p(s)$. Toutefois, pour les exemples 2(b), 3(b) et 4(b), les valeurs de p_i ne satisfont pas cette condition et, par conséquent, nous appliquons le plan PIPT de Sampford pour obtenir le plan PIPT initial. Les probabilités des échantillons non privilégiés (ϕ) pour ces exemples en utilisant le plan RN3, le plan RN4 et la méthode proposée sont présentées au tableau 3. Ce dernier montre que, alors que les plans RN3 et RN4 visent uniquement à minimiser la probabilité de sélection des échantillons non privilégiés, le plan proposé assure systématiquement que cette probabilité de sélection soit nulle.

Tableau 3 Probabilités de sélection des échantillons non privilégiés en utilisant les plans RN3, RN4 et proposés

Probabilité des échantillons non privilégiés (ϕ)	PLAN RN3	PLAN RN4	Plan proposé
Exemple 2(a) $N = 7, n = 3$	0,06	0 (c = 0,5)	0
Exemple 2(b) $N = 7, n = 3$	0,05	0 (c = 0,5)	0
Exemple 3(a) $N = 8, n = 3$	0,12	0 (c = 0,005)	0
Exemple 3(b) $N = 8, n = 3$	0,17	0 (c = 0,005)	0
Exemple 4(a) $N = 8, n = 4$	0,05	0 (c = 0,005)	0
Exemple 4(b) $N = 8, n = 4$	0,13	0 (c = 0,005)	0
Exemple 5 $N = 7, n = 4$	0,30	0,1008 (c = 0,5)	0

Les valeurs de $V(\hat{\bar{Y}}_{\mathrm{HT}})$ pour le plan proposé, le plan RN3, le plan RN4, le plan GK et le plan BD pour la population résumée en annexe sont présentées au tableau 2. L'examen de celui-ci nous permet de conclure que, dans tous les problèmes empiriques considérés, le plan proposé semble donner de meilleurs résultats que les plans RN3, RN4, GK et BD, ou des résultats assez approchants. L'accroissement de la variance de l'estimation observé dans certains cas pour le plan proposé pourrait être acceptable, étant donné l'élimination des échantillons indésirables.

Exemple 5 : Nous considérons maintenant un autre exemple pour démontrer la situation où le plan proposé ne fournit pas de solution faisable satisfaisant toutes les contraintes énoncées en (7). Le cas échéant, nous devons laisser tomber une contrainte en (7) pour obtenir une solution faisable du problème de programmation quadratique connexe.

Considérons une population de sept villages. Supposons qu'un échantillon de taille $n = 4$ est tiré à partir de cette population. Il existe 35 échantillons possibles, parmi lesquels les 14 qui suivent sont considérés comme étant non privilégiés :

1234; 1236; 1246; 1346; 1357; 1456; 1567;
2345; 2346; 2456; 2567; 3456; 3567; 4567.

Supposons que les valeurs de p_i qui suivent sont associées aux sept villages :

p_i : 0,14 0,13 0,15 0,13 0,16 0,15 0,14 .

Puisque les valeurs de p_i satisfont la condition (1), nous appliquons le plan MS (3) pour obtenir le plan PIPT initial $p(s)$ et nous résolvons le problème de programmation quadratique par la méthode exposée à la section 2. Toutefois, aucune solution faisable du problème de programmation quadratique connexe n'existe dans ce cas. Par conséquent, nous laissons tomber la contrainte (v) dans (7) pour ce problème particulier afin d'obtenir une solution faisable. Les probabilités de sélection des échantillons non privilégiés lorsqu'on utilise le plan RN3, le plan RN4 et le plan proposé pour ce problème empirique sont présentées à la dernière ligne du tableau 3. De nouveau, le plan proposé produit les valeurs de π_i originales et assure que la probabilité de sélection des échantillons non privilégiés soit exactement égale à zéro. Cependant, comme la contrainte $\pi_{ij} \le \pi_i \pi_j$ n'est pas satisfaite pour cet exemple, la non-négativité de l'estimateur de la variance de Yates-Grundy n'est pas assurée. Les valeurs de variance réelle, $V(\hat{\bar{Y}}_{\mathrm{HT}})$, pour le plan proposé, le plan RN3, le plan RN4, le plan GK et le plan BD sont présentées à la dernière ligne du tableau 2. La valeur de $V(\hat{\bar{Y}}_{\mathrm{HT}})$ pour cet exemple empirique en utilisant le plan proposé ne paraît pas être satisfaisante. Pour ce genre de problème, où la contrainte (v) n'est pas satisfaite, nous proposons d'utiliser d'autres estimateurs de la variance que celui de Yates-Grundy.

Nous avons également résolu un dernier exemple en prenant $N = 9$ et $n = 4$ et en utilisant les méthodes de

Midzuno-Sen ainsi que de Sampford pour obtenir le plan PIPT initial $p(s)$. Les solutions détaillées de ces problèmes, que nous omettons ici pour être brefs, peuvent être obtenues auprès des auteurs.

4. Conclusion

Nous avons proposé une approche de programmation quadratique pour résoudre les problèmes d'échantillonnage contrôlé en assurant que la probabilité de sélection des échantillons non privilégiés soit nulle. Le concept de « plan d'échantillonnage avec probabilité proportionnelle à la taille le plus proche » de Gabler (1987) est utilisé pour obtenir le plan proposé. Conceptuellement simple et très souple, l'approche permet d'utiliser une gamme de fonctions objectif et diverses contraintes. La seule limite de la méthode tient au fait qu'elle ne peut pas être appliquée à de grandes populations, car le processus de calcul devient assez fastidieux dans ces conditions. L'utilité de la méthode proposée est démontrée à l'aide d'exemples et sa variance d'échantillonnage réelle est comparée empiriquement à celle des plans d'échantillonnage contrôlé et des méthodes d'échantillonnage non contrôlé sous grande entropie existants. Le plan proposé donne des résultats satisfaisants.

Remerciements

Les auteurs remercient un rédacteur adjoint et deux examinateurs de leurs suggestions judicieuses et de leurs commentaires constructifs au sujet d'une version antérieure du présent article qui leur ont permis d'améliorer considérablement la présentation de ces travaux.

Annexe

Populations pour les exemples 2 à 4 avec les valeurs de Y_i et p_i et l'ensemble d'échantillons non privilégiés

Exemple 2. $N = 7$, $n = 3$.

Échantillons non privilégiés : 123; 126; 136; 146; 234; 236; 246; 137; 147; 167; 237; 247; 347; 467.

Y_i :	12	15	17	24	17	19	25
(a). p_i :	0,12	0,12	0,13	0,14	0,20	0,15	0,14
(b). p_i :	0,08	0,08	0,16	0,11	0,24	0,20	0,13

Exemple 3. $N = 8$, $n = 3$.

Échantillons non privilégiés : 123; 126; 136; 146; 234; 236; 246; 137; 147; 167; 237; 247; 347; 467; 128; 178; 248; 458; 468; 478; 578.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0,10	0,10	0,11	0,12	0,18	0,13	0,12	0,14
(b). p_i :	0,05	0,09	0,20	0,15	0,10	0,11	0,12	0,18

Exemple 4. $N = 8$, $n = 4$.

Échantillons non privilégiés : 1234; 1236; 1238; 1246; 1248; 1268; 1346; 1348; 1357; 1456; 1468; 1567; 1568; 1678; 2345; 2346; 2456; 2468; 2567; 2568; 2678; 3456; 3468; 3567; 3678; 4567; 4678; 5678.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0,11	0,11	0,12	0,13	0,17	0,12	0,11	0,13
(b). p_i :	0,09	0,09	0,18	0,11	0,12	0,14	0,17	0,10

Bibliographie

Avadhani, M.S., et Sukhatme, B.V. (1973). Controlled sampling with equal probabilities and without replacement. *Revue Internationale de Statistique*, 41, 175-182.

Brewer, K.R.W., et Donadio, M.E. (2003). La variance sous grande entropie de l'estimateur de Horvitz-Thompson. *Techniques d'enquête*, 29, 213-220.

Cassel, C.M., et Särndal, C.-E. (1972). A model for studying robustness of estimators and informativeness of labels in sampling with varying probabilities. *Journal of Royal Statistical Society*, Séries B, 34, 279-289.

Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *Journal of Indian Statistical Association*, 1, 78-85.

Foody, W., et Hedayat, A. (1977). On theory and applications of BIB designs and repeated blocks. *Annals of Statistics*, 5, 932-945.

Gabler, S. (1987). The nearest proportional to size sampling design. *Communications in Statistics-Theory & Methods*, 16(4), 1117-1131.

Goodman, R., et Kish, L. (1950). Controlled selection-a technique in probability sampling. *Journal of American Statistical Association*, 45, 350-372.

Gupta, V.K., Nigam, A.K. et Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, 69, 191-196.

Hedayat, A., et Lin, B.Y. (1980). Controlled probability proportional to size sampling designs. Rapport technique, *University of Illinois at Chicago*.

Hedayat, A., Lin, B.Y. et Stufken, J. (1989). The construction of IPPS sampling designs through a method of emptying boxes. *Annals of Statistics*, 17, 1886-1905.

Hess, I., et Srikantan, K.S. (1966). Some aspects of probability sampling technique of controlled selection. *Health Serv. Res. Summer 1966*, 8-52.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. *Journal of American Statistical Association*, 47, 663-85.

Kuhn, H.W., et Tucker A.W. (1951). Non-linear programming. *Proceedings of Second Berkely Symposium on Mathematical Statistics and Probability*, 481-492.

Midzuno, H. (1952). On the sampling system with probability proportional to sums of sizes. *Annals of Institute of Statistics & Mathematics*, 3, 99-107.

Mukhopadhyay, P., et Vijayan, K. (1996). On controlled sampling designs. *Journal of Statistical Planning & Inference*, 52, 375-378.

Nigam, A.K., Kumar, P. et Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *Journal of Royal Statistical Society*, B, 46, 564-571.

Rao, J.N.K., et Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.

Rao, J.N.K., et Nigam, A.K. (1992). 'Optimal' controlled sampling: A unified approach. *Revue Internationale de Statistique*, 60, 89-98.

Sampford, M.R. (1967). On sampling with replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, 5, 119-127.

Singh, D. (1954). On efficiency of sampling with varying probabilities without replacement. *Journal of Indian Society of Agricultural Statistics*, 6, 48-57.

Sitter, R.R., et Skinner, C.J. (1994). Stratification multidimensionnelle par programmation linéaire. *Techniques d'enquête*, 20, 69-78.

Srivastava, J., et Saleh, F. (1985). Need of *t*-designs in sampling theory. *Utilitas Mathematica*, 28, 5-17.

Takeuchi, K., Yanai, H. et Mukherjee, B.N. (1983). The Foundations of Multivariate Analysis. 1re Éd. New Delhi : Wiley Eastern Ltd.

Tiwari, N., et Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning & Inference*, 69, 89-100.

Waterton, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

Wynn, H.P. (1977). Convex sets of finite population plans. *Annals of Statistics*, 5, 414-418.

Yates, F., et Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society*, B, 15, 253-261.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 22, No. 3, 2006

Volume 22, No. 4, 2006

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 34, No. 4, December/décembre 2006

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) *etc*.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; l, 1).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, *etc.* à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.