Nº 12-001-X au catalogue

Techniques d'enquête

Juin 2009

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 toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

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

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

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

Comment accéder à ce produit ou le commander

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

Ce produit nº 12-001-X 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 les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, 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 que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

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.
∞

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2009

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2009

Tous droits réservés. Le contenu de la présente publication électronique peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada, sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire le contenu de la présente publication, ou de l'emmagasiner dans un système d'extraction, ou de le transmettre sous quelque forme ou par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2009

N° 12-001-XIF au catalogue
ISSN 1712-5685

N° 12-001-XPB au catalogue
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-X)

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

Volume 35, numéro 1, juin 2009

Table des matières

Dans ce numéro 1

Articles Réguliers

Victor M. Estevao et Carl-Erik Särndal
Un nouveau visage pour l'échantillonnage à deux phases avec estimateurs par calage 3

Jianzhu Li et Richard Valliant
Matrice chapeau et effets de levier pondérés par les poids de sondage 17

Jean-François Beaumont et Cynthia Bocci
Une méthode bootstrap pratique pour les tests d'hypothèses à partir des données d'enquête 29

Bo-Seung Choi, Jai Won Choi et Yousung Park
Méthodes bayésiennes pour un tableau de contingence à double entrée incomplet avec application aux sondages électoraux de l'État de l'Ohio 41

Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff et Van L. Parsons
Estimations pour petits domaines bayésiennes hiérarchiques et empiriques de la proportion de personnes sans assurance-maladie dans les groupes de population minoritaires 57

Tucker McElroy et Scott Holan
Un test non paramétrique pour la saisonnalité résiduelle 71

Siegfried Gabler et Partha Lahiri
De la définition et de l'interprétation de la variabilité d'intervieweur pour un plan d'échantillonnage complexe 91

Barry Schouten, Fannie Cobben et Jelke Bethlehem
Indicateurs de la représentativité de la réponse aux enquêtes 107

Guillaume Chauvet
Échantillonnage équilibré stratifié 123

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 1-2
Statistique Canada, N° 12-001-X au catalogue

Dans ce numéro

Dans le premier article de ce numéro de *Techniques d'enquête*, Estevao et Särndal examinent le problème de l'estimation par calage dans le contexte de l'échantillonnage à deux phases. Les apports de l'article comprennent le choix des poids initiaux dans la procédure de calage, ainsi que l'important problème de l'estimation de la variance. Les auteurs proposent de nouveaux estimateurs de la variance et procèdent à une étude par simulation, dont les résultats montrent que ces nouveaux estimateurs s'avèrent plus efficaces que les estimateurs classiques.

Ensuite, Li et Valliant étudient le problème de la détection des unités influentes dans l'analyse des données d'enquête par régression linéaire. Ils commencent par donner une expression de la matrice chapeau et des effets de levier connexes (éléments diagonaux de la matrice chapeau) quand les paramètres du modèle sont estimés par les moindres carrés pondérés. Puis, ils proposent une décomposition des effets de levier et soulignent que l'effet de levier d'une unité donnée peut être important si le poids de sondage de cette dernière est grand ou que son vecteur de variables explicatives est éloigné du centre. Ils illustrent l'effet des unités influentes sur les moindres carrés ordinaires ainsi que pondérés au moyen d'un exemple numérique.

Beaumont et Bocci proposent une méthode bootstrap pour vérifier les hypothèses au sujet d'un vecteur de paramètres inconnus du modèle quand l'échantillon a été tiré d'une population finie. La méthode s'appuie sur des statistiques de test fondées sur un modèle dans lesquelles sont intégrés les poids de sondage et qui peuvent habituellement être obtenues facilement en se servant de progiciels standard. Au moyen d'une étude par simulation, les auteurs montrent que la méthode proposée donne des résultats comparables à ceux de la procédure de Rao-Scott, et de meilleurs résultats que celles de Wald et de Bonferroni lorsque l'on vérifie les hypothèses au sujet d'un vecteur de paramètres d'un modèle de régression linéaire.

Dans leur article, Park, Choi et Choi présentent une approche intéressante de traitement de la non-réponse. Des études ont montré que le comportement de vote des électeurs indécis peut avoir une incidence importante sur le résultat final d'une élection et que l'on peut accroître l'exactitude de la prédiction des résultats en tenant compte de ces électeurs. Les auteurs présentent deux modèles bayésiens dont les distributions de probabilité a priori dépendent de l'information donnée par les répondants et les indécis. Ils analysent un tableau de contingence à double entrée incomplet en se servant de quatre jeux de données provenant des sondages électoraux réalisés en 1998 dans l'État de l'Ohio pour illustrer comment il convient d'utiliser et d'interpréter les résultats des estimations pour les élections.

Ghosh, Kim, Sinha, Maiti, Katzoff et Parsons élaborent des méthodes bayésiennes hiérarchiques et empiriques pour l'estimation de proportions dans de petits domaines en utilisant des modèles au niveau de l'unité. Ils proposent un modèle hiérarchique bayésien analogue au modèle linéaire mixte généralisé pour obtenir les moyennes et les erreurs-types a posteriori des proportions pour des petits domaines de population. Adoptant une approche fondée sur la théorie des fractions d'estimation optimales, ils obtiennent aussi les estimateurs bayésiens empiriques et les estimateurs asymptotiques de l'erreur quadratique moyenne correspondants. Ils illustrent les méthodes en se servant de données provenant de la National Health Interview Survey (NHIS) pour obtenir des estimations sur petits domaines des proportions d'Asiatiques ne possédant pas d'assurance-maladie.

L'article de McElroy et Holan porte sur les tests de détection de la saisonnalité résiduelle dans les données désaisonnalisées. Les auteurs proposent un test de signification statistique permettant de détecter dans la densité spectrale de la série chronologique étudiée les pics indicateurs de saisonnalité. Ils élaborent la théorie qui sous-tend la méthode proposée et illustrent et comparent cette dernière aux méthodes existantes au moyen d'études par simulation et d'études empiriques.

Gabler et Lahiri fournissent une justification assistée par modèle de la formule classique de la variance due à l'intervieweur pour l'échantillonnage avec probabilités égales sans grappes spatiales. Puis, ils obtiennent, dans le contexte d'un plan d'échantillonnage complexe, une définition de la variabilité due à l'intervieweur qui tient compte comme il convient des probabilités inégales de sélection et des grappes spatiales. Ils proposent aussi une décomposition des effets totaux en effets dus à la pondération, aux grappes spatiales et aux intervieweurs. Leurs résultats peuvent aider à mieux comprendre et contrôler les sources de variabilité.

Dans leur article, Schouten, Cobben et Bethlehem se penchent sur l'évaluation de la similarité entre la réponse à une enquête et l'échantillon ou la population à l'étude. Ils proposent un indicateur de représentativité pour remplacer les taux de réponse comme indicateur de la qualité lorsque l'on évalue l'effet du biais de non-réponse. Ils montrent que cet indicateur, appelé indicateur R, est apparenté jusqu'à un certain point au V de Cramer qui mesure l'association entre la variable de réponse et les variables auxiliaires. En fait, il est préférable de considérer l'indicateur R comme une mesure de l'absence d'association, puisqu'une association faible implique qu'il n'existe aucune preuve que la non-réponse a eu une incidence sur la composition des données observées. Les auteurs établissent les propriétés théoriques de l'indicateur proposé et illustrent ce dernier au moyen d'études empiriques.

Finalement, dans son article, Chauvet s'attaque au problème de l'échantillonnage équilibré lorsque les tailles dans chaque strate sont trop petites pour permettre un équilibrage exact. L'auteur propose un algorithme adapté de la méthode du Cube garantissant l'équilibrage au niveau de la population. Une étude par simulation confirme la bonne performance de la méthode proposée.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 3-16
Statistique Canada, N° 12-001-X au catalogue

Un nouveau visage pour l'échantillonnage à deux phases avec estimateurs par calage

Victor M. Estevao et Carl-Erik Särndal [1]

Résumé

Le présent article décrit un cadre pour l'estimation par calage sous les plans d'échantillonnage à deux phases. Les travaux présentés découlent de la poursuite du développement de logiciels généralisés d'estimation à Statistique Canada. Un objectif important de ce développement est d'offrir une grande gamme d'options en vue d'utiliser efficacement l'information auxiliaire dans différents plans d'échantillonnage. Cet objectif est reflété dans la méthodologie générale pour les plans d'échantillonnage à deux phases exposée dans le présent article.

Nous considérons le plan d'échantillonnage à deux phases classique. Un échantillon de première phase est tiré à partir d'une population finie, puis un échantillon de deuxième phase est tiré en tant que sous-échantillon du premier. La variable étudiée, dont le total de population inconnu doit être estimé, est observée uniquement pour les unités contenues dans l'échantillon de deuxième phase. Des plans d'échantillonnage arbitraires sont permis à chaque phase de l'échantillonnage. Divers types d'information auxiliaire sont identifiés pour le calcul des poids de calage à chaque phase. Les variables auxiliaires et les variables étudiées peuvent être continues ou catégoriques.

L'article apporte une contribution à quatre domaines importants dans le contexte général du calage pour les plans d'échantillonnage à deux phases :

1) nous dégageons trois grands types d'information auxiliaire pour les plans à deux phases et les utilisons dans l'estimation. L'information est intégrée dans les poids en deux étapes : un calage de première phase et un calage de deuxième phase. Nous discutons de la composition des vecteurs auxiliaires appropriés pour chaque étape et utilisons une méthode de linéarisation pour arriver aux résidus qui déterminent la variance asymptotique de l'estimateur par calage ;
2) nous examinons l'effet de divers choix de poids de départ pour le calage. Les deux choix « naturels » produisent généralement des estimateurs légèrement différents. Cependant, sous certaines conditions, ces deux estimateurs ont la même variance asymptotique ;
3) nous réexaminons l'estimation de la variance pour l'estimateur par calage à deux phases. Nous proposons une nouvelle méthode qui peut représenter une amélioration considérable par rapport à la technique habituelle de conditionnement sur l'échantillon de première phase. Une simulation décrite à la section 10 sert à valider les avantages de cette nouvelle méthode ;
4) nous comparons l'approche par calage à la méthode de régression assistée par modèle classique qui comporte l'ajustement d'un modèle de régression linéaire à deux niveaux. Nous montrons que l'estimateur assisté par modèle a des propriétés semblables à celles d'un estimateur par calage à deux phases.

Mots clés : Information auxiliaire ; estimateur par la régression à deux phases ; poids de départ ; estimateur de variance à résidus distincts ; estimateur de la variance à résidu combiné.

1. Introduction

L'expression *échantillonnage double* fait référence aux plans d'échantillonnage dont la caractéristique commune est le tirage de deux échantillons probabilistes, désignés par s_1 et s, qui sont tous deux des sous-ensembles de la population finie d'intérêt donnée par $U = \{1, ..., k, ..., N\}$. L'échantillon s_1 est réalisé et observé avant l'échantillon s. Une variable étudiée type est désignée par y ; sa valeur y_k est obtenue uniquement pour les unités $k \in s$. L'objectif est d'estimer le total y de population $Y = \sum_U y_k$ (si A est un ensemble d'unités, $A \subseteq U$, nous utilisons la notation $\sum_A$ comme forme abrégée de $\sum_{k \in A}$ s'il n'y a aucune ambiguïté).

Hidiroglou (2001) discute de deux types d'échantillonnage double, *emboîté* et *non emboîté*. Le présent article porte sur le type emboîté, habituellement appelé échantillonnage à deux phases : l'échantillon de deuxième phase s est un sous-échantillon de l'échantillon de première phase s_1 tiré à partir de U, de sorte que $s \subseteq s_1 \subseteq U$.

L'estimation sous échantillonnage à deux phases a été examinée par plusieurs auteurs dans le contexte où deux sources d'information auxiliaire sont reconnues et traitées en fonction de leur niveau. Au niveau de la population, le total $\sum_U \mathbf{x}_{1k}$ est connu, où $\mathbf{x}_{1k}$ est un vecteur connu pour chaque $k \in s_1$; par conséquent, il est également connu pour chaque $k \in s$. Au niveau du premier échantillon, la valeur du vecteur $\mathbf{x}_{2k}$ est observée pour chaque $k \in s_1$ et est par conséquent connue pour chaque $k \in s$; le total $\sum_U \mathbf{x}_{2k}$ est inconnu, mais peut être estimé sans biais par rapport au plan au niveau de s_1. On trouve dans la littérature deux arguments en faveur de l'intégration de ces deux

1. Victor M. Estevao, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), Canada, K1A OT6. Courriel : victor.estevao@statcan.gc.ca ; Carl-Erik Särndal, professeur. Courriel : carl.sarndal@rogers.com.

types d'information auxiliaire dans l'estimation $Y = \sum_U y_k$: celui de l'ajustement d'une régression et celui du calage. Sous certaines conditions, ils peuvent aboutir à des estimateurs identiques, mais il n'en est généralement pas ainsi.

L'ajustement d'une régression l'emporte dans Särndal et Swensson (1987), Särndal, Swensson et Wretman (1992), Sitter (1997), Hidiroglou et Särndal (1998), Axelson (1998), ainsi qu'Hidiroglou, Rao et Haziza (2006). L'approche du calage décrite dans Deville et Särndal (1992) a été appliquée à l'échantillonnage à deux phases par Dupont (1995). Celle-ci compare les estimateurs par calage résultant à ceux obtenus selon l'approche par la régression. Pour la même information auxiliaire, les deux approches ne donnent pas nécessairement des estimateurs identiques, mais il est probable qu'en pratique l'écart ait peu de conséquence. Kott et Stukel (1997) ont étudié la méthode du rééchantillonnage pour l'estimation de la variance dans le cas de l'échantillonnage en deux phases. Estevao et Särndal (2002) se concentrent sur l'argument du calage et distinguent dix moyens différents d'utiliser totalement ou partiellement l'information disponible aux deux niveaux. Le présent article est également axé sur l'approche par calage. Il étoffe les travaux antérieurs en reconnaissant trois (plutôt que deux) types d'information auxiliaire, possédant chacun des caractéristiques différentes.

Dans l'approche par régression, il est naturel d'ajuster deux régressions par les moindres carrés linéaires. Un ensemble de valeurs de y prédites par régression sont produites pour $k \in s_1$ en utilisant à la fois $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$ comme prédicteurs ; un autre est produit pour $k \in s_1$ en utilisant seulement le vecteur $\mathbf{x}_{1k}$ comme prédicteur. Les deux ensembles de valeurs prédites de y, ainsi que le total connu $\sum_U \mathbf{x}_{1k}$ sont utilisés pour construire l'estimateur de type régression de Y, de la façon décrite à la section 9.

L'approche par calage est motivée par deux facteurs : créer un ensemble de poids concordant avec les totaux connus ou estimés pour les valeurs auxiliaires et réduire la variance des estimations faites pour la ou les variables étudiées. Nous voulons qu'il y ait une cohérence entre les poids w_k dans $\hat{Y}_{2P} = \sum_s w_k y_k$ et le total $\sum_U \mathbf{x}_{1k}$ connu au niveau de la population et (ou) une estimation (approximativement) sans biais, faite au niveau de l'échantillon de première phase, du total $\sum_U \mathbf{x}_{2k}$ inconnu. Puisque y est observé uniquement au dernier niveau (échantillon de deuxième phase), la cohérence « aux niveaux plus élevés » sur les variables auxiliaires importantes réduira souvent de manière significative la variance de $\hat{Y}_{2P} = \sum_s w_k y_k$. Nous pouvons distinguer deux étapes dans le processus aboutissant aux poids w_k, un calage de première phase et un calage de deuxième phase.

Le plan d'échantillonnage à deux phases est le suivant : de la population finie d'unités $U = \{1, 2, \ldots k, \ldots N\}$, nous tirons un échantillon de première phase s_1. La probabilité d'inclusion positive connue de l'unité k est $\pi_{1k} = \Pr(k \in s_1)$ et le poids de sondage de première phase est $a_{1k} = 1/\pi_{1k}$. Certaines variables sont peut-être observées pour les unités $k \in s_1$. Alors, sachant s_1, nous sélectionnons un échantillon de deuxième phase s à partir de s_1. La probabilité d'inclusion conditionnelle connue et positive de k est $\pi_{2k} = \Pr(k \in s \mid s_1)$ pour $k \in s_1$, et le poids de sondage conditionnel de deuxième phase est $a_{2k} = 1/\pi_{2k}$. (Pour que la notation reste simple, nous utilisons π_{2k} et a_{2k} plutôt que les formes plus suggestives $\pi_{2k|s_1}$ et $a_{2k|s_1}$; il convient de se rappeler que π_{2k} et a_{2k} sont calculés conditionnellement à l'échantillon de première phase s_1). Le poids de sondage combiné ou à double facteur d'extension est $a_k = a_{1k} a_{2k}$ pour $k \in s$. Dans le présent article, l'analyse des estimateurs est fondée sur le plan et l'expression « (approximativement) sans biais » signifie « (approximativement) sans biais par rapport au plan ». Nous émettons l'hypothèse de conditions faibles sur la population et les deux plans d'échantillonnage, ce qui nous permet d'écarter les termes d'ordre inférieur dans l'analyse de nos estimateurs quand les tailles attendues des échantillons de première et de deuxième phases sont suffisamment grandes.

L'estimateur à double facteur d'extension $\sum_s a_k y_k$ est sans biais pour $Y = \sum_U y_k$. Nous pouvons produire des estimateurs plus efficaces en tenant compte de la disponibilité de l'information auxiliaire. Pour les plans d'échantillonnage à deux phases, nous pouvons distinguer trois types ou ensembles de variables auxiliaires (appelées variables x). Ils sont désignés par $\mathfrak{X}^{\oplus}$, $\mathfrak{X}^{\dagger}$ et $\mathfrak{X}^{\circ}$. Le tableau qui suit donne leurs caractéristiques en ce qui concerne l'information.

Tableau 1.1
Ensembles de variables auxiliaires pour le calage dans l'échantillonnage à deux phases

Ensemble de variables auxiliaires	Total de la variable auxiliaire sur U	Valeurs unitaires de la variable pour $k \in s_1$	Valeurs unitaires de la variable pour $k \in s$
$\mathfrak{X}^{\oplus}$	connu	connues	connues
$\mathfrak{X}^{\dagger}$	connu	inconnues	connues
$\mathfrak{X}^{\circ}$	inconnu	connues	connues

Chaque ensemble peut contenir n'importe quel nombre de variables x. Les trois ensembles sont mutuellement exclusifs. Les propriétés indiquées dans les trois dernières colonnes s'appliquent à chaque variable x dans l'ensemble correspondant. Toutes les variables x utilisées pour le calage appartiennent à l'un de ces trois ensembles.

2. Calage de première phase

Pour le calage de première phase, nous utilisons un vecteur $\mathbf{x}_{1k}$ de variables auxiliaires tirées de l'ensemble

$\mathfrak{X}^{\oplus}$. Bien qu'il soit naturel de permettre que $\mathbf{x}_{1k}$ soit constitué de toutes les variables contenues dans $\mathfrak{X}^{\oplus}$, ici, l'exposé général nous permet de définir $\mathbf{x}_{1k}$ de manière à ce qu'il ne comprenne que certaines variables de $\mathfrak{X}^{\oplus}$, voire aucune. Les poids de calage de première phase w_{1k} sont calculés en modifiant les poids de départ de première phase a_{1k} sous la contrainte de calage $\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Dans notre formulation, les poids de calage sont donnés pour $k \in s_1$ sous la forme

$$w_{1k} = a_{1k}\left\{1 + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\left(\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1}\mathbf{z}_{1k}\right\} \quad (2.1)$$

où $\mathbf{X}_1 = \sum_U \mathbf{x}_{1k}$, $\hat{\mathbf{X}}_1 = \sum_{s_1} a_{1k}\mathbf{x}_{1k}$ et $\mathbf{z}_{1k}$ est un vecteur instrumental de même dimension que $\mathbf{x}_{1k}$. Il remplace $\mathbf{x}_{1k}/\sigma^2_{1k}$ dans la formule de l'estimateur assisté par modèle décrit par Särndal, Swensson et Wretman (1992), et permet une spécification plus générale des poids de calage. L'utilisation d'un vecteur instrumental est discutée dans Estevao et Särndal (2000) et dans Deville (2002). Ici et dans la suite, nous supposons systématiquement que les matrices sont inversibles, comme celle sur s_1 dans (2.1) et celles (sur s et U) qui apparaîtront plus tard.

3. Calage de deuxième phase

Nous utilisons un vecteur $\mathbf{x}_k$ de variables auxiliaires pour produire un ensemble de poids de calage de deuxième phase (ou finaux) w_k. Nous les utilisons pour calculer $\hat{Y}_{2P} = \sum_s w_k y_k$ à titre d'estimateur de $Y = \sum_U y_k$. Le vecteur $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$ comprend trois composantes, décrites plus loin. Aucune variable auxiliaire ne peut figurer dans plus d'une des trois composantes vectorielles. Ces trois composantes jouent des rôles différents dans l'établissement de l'équation de calage de deuxième phase $\sum_s w_k\mathbf{x}_k = \mathbf{X}$ et dans la détermination des poids de calage de deuxième phase.

Les variables figurant dans le vecteur $\mathbf{x}_{k(t)}$ sont sélectionnées parmi celles de l'ensemble $\mathfrak{X}^{\oplus} \cup \mathfrak{X}^{\dagger}$. Cela signifie que le total $\sum_U \mathbf{x}_{k(t)}$ est connu et peut être inclus dans $\mathbf{X}$. Les variables contenues dans $\mathbf{x}_{1k}$ peuvent réapparaître dans $\mathbf{x}_{k(t)}$, ce qui est habituellement préférable afin de réduire la variance de l'estimateur. Nous pouvons spécifier $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$, mais notre cadre permet que $\mathbf{x}_{k(t)}$ comprenne des variables provenant de $\mathfrak{X}^{\dagger}$. Par conséquent, nous pouvons utiliser des variables dont les totaux de population sont connus dans des situations où la collecte de données sur les variables est très coûteuse pour un grand échantillon de première phase s_1, mais sont observables pour l'échantillon de deuxième phase plus petit s. Ces variables sont exclues du calage de première phase parce que leurs valeurs ne sont pas disponibles pour $k \in s_1$.

Les variables comprises dans $\mathbf{x}_{k(w)}$ et $\mathbf{x}_{k(a)}$ sont sélectionnées parmi celles de l'ensemble $\mathfrak{X}^{\oplus} \cup \mathfrak{X}^{\dagger} \cup \mathfrak{X}^{\circ}$, à condition qu'elles ne soient pas déjà incluses dans $\mathbf{x}_{k(t)}$. Les variables comprises dans $\mathbf{x}_{k(w)}$ sont celles pour lesquelles nous voulons satisfaire l'équation de calage de deuxième phase $\sum_s w_k\mathbf{x}_{k(w)} = \sum_{s_1} w_{1k}\mathbf{x}_{k(w)}$, où le deuxième membre est approximativement sans biais pour $\sum_U \mathbf{x}_{k(w)}$. Les variables comprises dans $\mathbf{x}_{k(a)}$ sont celles pour lesquelles nous voulons satisfaire l'équation de calage de deuxième phase $\sum_s w_k\mathbf{x}_{k(a)} = \sum_{s_1} a_{1k}\mathbf{x}_{k(a)}$. Ici, le deuxième membre est sans biais pour $\sum_U \mathbf{x}_{k(a)}$. Le fait d'inclure à la fois $\mathbf{x}_{k(w)}$ et $\mathbf{x}_{k(a)}$ dans la définition de $\mathbf{x}_k$ nous permet d'effectuer le calage sur l'un de ces vecteurs ou sur les deux et offre un cadre général pour la production de divers estimateurs à partir du calage de deuxième phase.

L'équation de calage de deuxième phase est $\sum_s w_k\mathbf{x}_k = \mathbf{X}$, où $\mathbf{X}$ est le vecteur auxiliaire empilé

$$\mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_{k(t)} \\ \sum_{s_1} w_{1k}\mathbf{x}_{k(w)} \\ \sum_{s_1} a_{1k}\mathbf{x}_{k(a)} \end{pmatrix}. \quad (3.1)$$

Une variable particulière ne peut survenir qu'une seule fois dans $\mathbf{x}_k$. Sinon, l'équation de calage pourrait être incohérente et n'admettre aucune solution.

Nous désignons les poids de départ pour le calage de deuxième phase par a^*_k pour $k \in s$. Il existe plus d'un choix raisonnable pour les a^*_k. Nous considérons deux alternatives, qui semblent l'une et l'autre naturelles : 1) $a^*_k = a_k = a_{1k}\,a_{2k}$ et 2) $a^*_k = w_{1k}\,a_{2k}$, où w_{1k} est le poids de calage de première phase donné par (2.1).

Sachant les poids de départ a^*_k, nous déterminons les poids finaux w_k sous la contrainte de l'équation de calage $\sum_s w_k\mathbf{x}_k = \mathbf{X}$. Ces poids finaux sont donnés pour $k \in s$ par

$$w_k = a^*_k\left\{1 + (\mathbf{X} - \hat{\mathbf{X}})'\left(\sum_s a^*_k\mathbf{z}_k\mathbf{x}'_k\right)^{-1}\mathbf{z}_k\right\} \quad (3.2)$$

où $\hat{\mathbf{X}} = \sum_s a^*_k\mathbf{x}_k$ est un estimateur sans biais ou approximativement sans biais de $\mathbf{X}$, selon la composition de $\mathbf{x}_k$. La variable instrumentale $\mathbf{z}_k$ a la même dimension que $\mathbf{x}_k$. Nous supposons que les vecteurs $\mathbf{z}_{1k}$ et $\mathbf{z}_k$ sont des fonctions fixes de $\mathbf{x}_{1k}$ et $\mathbf{x}_k$. La question de savoir comment choisir $\mathbf{z}_{1k}$ et $\mathbf{z}_k$ est un sujet que nous laisserons à d'autres le soin d'aborder.

4. Comparaison de deux options pour les poids de départ

L'objectif ici est d'analyser comment les poids finaux w_k dans $\hat{Y}_{2P} = \sum_s w_k y_k$ dépendent de la spécification des poids de départ a^*_k dans (3.2). Nous considérons deux cas distincts dépendant du fait que les variables auxiliaires $\mathbf{x}_k$

sont utilisées ou non pour le calage de deuxième phase. Quand nous effectuons ce calage de deuxième phase, les deux choix distincts pour les poids de départ mènent généralement à des estimateurs différents. Nous montrons que ces estimateurs sont asymptotiquement équivalents sous certaines conditions fréquemment observées en pratique. Si nous n'effectuons pas de calage de deuxième phase, les deux choix pour les poids de départ produisent deux autres estimateurs qui sont habituellement moins efficaces que ceux obtenus en effectuant ce calage.

4.1 Estimateurs avec calage de deuxième phase ($\mathbf{x}_k \neq \phi$)

Comme nous l'avons mentionné plus haut, il existe deux options pour les poids de départ a_k^* figurant dans (3.2) : 1) $a_k^* = a_k = a_{1k}\, a_{2k}$ et 2) $a_k^* = w_{1k}\, a_{2k}$, où w_{1k} est le poids de calage de première phase donné par (2.1). Nous présentons maintenant une analyse détaillée de la forme de l'estimateur sous ces deux choix. À la présente sous-section, nous examinons le cas le plus intéressant où nous procédons au calage de deuxième phase ($\mathbf{x}_k \neq \phi$). À la sous-section suivante, nous examinons ce qui se passe quand nous n'effectuons pas le calage de deuxième phase ($\mathbf{x}_k = \phi$).

Notre méthode est la suivante. Premièrement, nous établissons la forme linéarisée (asymptotique) de $\hat{Y}_{2P}$ en nous fondant sur les poids de départ généraux a_k^*. Puis, dans l'expression obtenue, nous substituons à ces poids généraux les deux choix pour a_k^*. Nous déterminons $\hat{Y}_{2P}$ en nous fondant sur les poids de départ $a_k^* = a_k = a_{1k}\, a_{2k}$. Nous désignons cet estimateur par $\hat{Y}_{2Pa}$ et calculons sa forme linéarisée, $\hat{Y}_{2Pa\,\mathrm{lin}}$. De même, nous obtenons $\hat{Y}_{2P}$ en nous fondant sur les poids de départ $a_k^* = w_{1k} a_{2k}$. Nous désignons cet estimateur par $\hat{Y}_{2Pw}$ et calculons sa forme linéarisée, $\hat{Y}_{2Pw\,\mathrm{lin}}$. Ces deux formes sont légèrement différentes, mais nous prouvons dans le résultat 4.2 que $\hat{Y}_{2Pa\,\mathrm{lin}} = \hat{Y}_{2Pw\,\mathrm{lin}}$ sous certaines conditions.

Nous commençons par insérer les poids w_k dans $\hat{Y}_{2P} = \sum_s w_k\, y_k$ et écrivons l'estimateur sous la forme

$$\hat{Y}_{2P} = \sum_U \mathbf{x}'_{k(t)}\, \mathbf{B}_{(y\,;\,\mathbf{x})(t)} + \sum_{s_1} w_{1k}\, \mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\,\mathbf{x})(w)} + \sum_{s_1} a_{1k}\, \mathbf{x}'_{k(a)} \mathbf{B}_{(y\,;\,\mathbf{x})(a)} + \sum_s a_k^*\, e_{(y\,;\,\mathbf{x})k} + (\mathbf{X} - \hat{\mathbf{X}})'\, (\hat{\mathbf{B}}^*_{(y\,;\,\mathbf{x})} - \mathbf{B}_{(y\,;\,\mathbf{x})}) \qquad (4.1)$$

où $\hat{\mathbf{B}}^*_{(y\,;\,\mathbf{x})} = (\sum_s a_k^*\, \mathbf{z}_k\, \mathbf{x}'_k)^{-1} \sum_s a_k^*\, \mathbf{z}_k\, y_k$, $\mathbf{B}_{(y\,;\,\mathbf{x})} = (\sum_U \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_U \mathbf{z}_k\, y_k$ et $\mathbf{B}_{(y\,;\,\mathbf{x})} = (\mathbf{B}'_{(y\,;\,\mathbf{x})(t)}, \mathbf{B}'_{(y\,;\,\mathbf{x})(w)}, \mathbf{B}'_{(y\,;\,\mathbf{x})(a)})'$ est le partitionnement correspondant à $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$. Notre notation en indice de la forme $(\mathbf{v}_1\,;\,\mathbf{v}_2)$ identifie les variables dans la régression. Le terme $\mathbf{v}_2$ renvoie aux variables indépendantes et $\mathbf{v}_1$ identifie la ou les variables dépendantes. Pour simplifier, nous n'incluons pas dans la notation les vecteurs de variables instrumentales $\mathbf{z}_{1k}$ et $\mathbf{z}_k$.

Le terme $e_{(y\,;\,\mathbf{x})k} = y_k - \mathbf{x}'_k \mathbf{B}_{(y\,;\,\mathbf{x})}$ est défini pour $k \in U$. Notons que, même si $e_{(y\,;\,\mathbf{x})k}$ a l'air d'un résidu de régression, il ne résulte pas de l'ajustement d'un modèle de régression approprié. Ensuite, nous développons le terme $\sum_{s_1} w_{1k}\, \mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\,\mathbf{x})(w)}$ dans (4.1) en insérant l'expression (2.1) pour w_{1k} et en nous servant de l'équation de calage de première phase $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Nous obtenons

$$\sum_{s_1} w_{1k} \mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\,\mathbf{x})(w)} = \mathbf{X}'_1\, \mathbf{B}_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)} + \sum_{s_1} a_{1k} e_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)k} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\, (\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)}) \qquad (4.2)$$

où $\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)}\,;\mathbf{x}_1)} = (\sum_{s_1} a_{1k}\, \mathbf{z}_{1k}\, \mathbf{x}'_{1k})^{-1} \sum_{s_1} a_{1k}\, \mathbf{z}_{1k}\, \mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\mathbf{x})(w)}$ converge en probabilité vers (et est approximativement sans biais pour) $\mathbf{B}_{(\mathbf{xB}_{(w)}\,;\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\, \mathbf{x}'_{1k})^{-1} \sum_U \mathbf{z}_{1k}\, \mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\,\mathbf{x})(w)}$, et $e_{(\mathbf{xB}_{(w)}\,;\mathbf{x}_1)k} = \mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\mathbf{x})(w)} - \mathbf{x}'_{1k} \mathbf{B}_{(\mathbf{xB}_{(w)}\,;\mathbf{x}_1)}$ est défini pour $k \in U$.

Nous pouvons interpréter $e_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)k}$ comme étant un résidu résultant d'un ajustement au niveau de la population fondé sur une régression généralisée de $\mathbf{x}'_{k(w)} \mathbf{B}_{(y\,;\,\mathbf{x})(w)}$ comme variable dépendante et $\mathbf{x}_{1k}$ comme vecteur de variables indépendantes. L'introduction de l'expression (4.2) par substitution dans l'expression (4.1) pour $\hat{Y}_{2P}$ donne

$$\hat{Y}_{2P} = \sum_U (\mathbf{x}'_{k(t)} \mathbf{B}_{(y\,;\,\mathbf{x})(t)} + \mathbf{x}'_{1k}\, \mathbf{B}_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)}) + \sum_{s_1} a_{1k} (\mathbf{x}'_{k(a)} \mathbf{B}_{(y\,;\,\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)k}) + \sum_s a_k^*\, e_{(y\,;\,\mathbf{x})k} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\, (\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB}_{(w)}\,;\,\mathbf{x}_1)}) + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^*_{(y\,;\,\mathbf{x})} - \mathbf{B}_{(y\,;\,\mathbf{x})}). \qquad (4.3)$$

Le résultat qui suit établit la relation entre les estimateurs obtenus pour les deux choix de poids de départ.

Résultat 4.1 : Les formes linéarisées de $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$ sont reliées par l'équation $\hat{Y}_{2Pw\,\mathrm{lin}} = \hat{Y}_{2Pa\,\mathrm{lin}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y\,;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\,;\mathbf{x}_1)} \mathbf{B}_{(y\,;\,\mathbf{x})})$.

Preuve

Nous considérons l'expression (4.3) sous les deux choix possibles pour a_k^*. Premièrement, avec $a_k^* = a_k = a_{1k}\, a_{2k}$, nous obtenons $\hat{Y}_{2Pa}$ donné par

$$\hat{Y}_{2Pa} = \sum_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}) + \sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\mathbf{x}_1)k}) + \sum_s a_k e_{(y;\mathbf{x})k} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}) + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})}). \quad (4.4)$$

Le terme $\hat{\mathbf{B}}_{(y;\mathbf{x})} = (\sum_s a_k \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_s a_k \mathbf{z}_k y_k$ converge en probabilité vers (et est approximativement sans biais pour) $\mathbf{B}_{(y;\mathbf{x})} = (\sum_U \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_U \mathbf{z}_k y_k$. Le premier terme est constant et ne contribuent pas à la variance de $\hat{Y}_{2Pa}$. Les deux prochains termes du milieu sont des quantités aléatoires, définies comme étant des sommes sur s_1 et s, respectivement. Les deux derniers termes sont des produits de différences dont l'espérance est nulle ou presque nulle. Pour ce qui est du produit $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)})$, les différences sont toutes deux des fonctions de l'échantillon de première phase s_1. Nous savons que $\hat{\mathbf{X}}_1$ est sans biais pour $\mathbf{X}_1$ et que $\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}$ est approximativement sans biais pour $\mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}$. Sous des conditions assez générales, $N^{-1}(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}) = O_P(n_1^{-1})$, où n_1 est la taille attendue de s_1, supposée suffisamment grande. En suivant un raisonnement semblable, $N^{-1}(\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})}) = O_P(n^{-1})$, où n est la taille attendue de s, également supposée suffisamment grande. Conséquemment, nous pouvons laisser tomber les deux derniers termes de (4.4), parce qu'ils sont d'ordre plus faible que les termes précédents : $N^{-1}\sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\,\mathbf{B}_{(y;\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\mathbf{x}_1)k})$ est $O_P(n_1^{-1/2})$ et $N^{-1}\sum_s a_k e_{(y;\mathbf{x})k}$ est $O_P(n^{-1/2})$. Les trois premiers termes définissent la forme linéarisée de $\hat{Y}_{2Pa}$,

$$\hat{Y}_{2Pa\,\mathrm{lin}} = \sum_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}) + \sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\mathbf{x}_1)k}) + \sum_s a_k e_{(y;\mathbf{x})k}. \quad (4.5)$$

Maintenant, considérons l'expression (4.3) sous le deuxième choix, $a_k^* = w_{1k}\,a_{2k}$. Cela nous mène à $\hat{Y}_{2Pw}$ donné par

$$\hat{Y}_{2Pw} = \sum_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}) + \sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\mathbf{x}_1)k}) + \sum_s a_k e_{(y;\mathbf{x})k} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\Big(\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\Big)^{-1}\sum_s a_k \mathbf{z}_{1k} e_{(y;\mathbf{x})k} + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^w_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})}) \quad (4.6)$$

où $\hat{\mathbf{B}}^w_{(y;\mathbf{x})} = (\sum_s w_{1k} a_{2k} \mathbf{z}_k \mathbf{x}'_k)^{-1} \sum_s w_{1k} a_{2k} \mathbf{z}_k y_k$ et $\hat{\mathbf{X}} = \sum_s w_{1k} a_{2k}\,\mathbf{x}_k$. Les trois premiers termes de $\hat{Y}_{2Pw}$ sont les mêmes que ceux figurant dans l'expression (4.4) pour $\hat{Y}_{2Pa}$. Les quatrième et cinquième termes diffèrent de leurs homologues dans (4.4). Bien que $\hat{\mathbf{B}}^w_{(y;\mathbf{x})}$ et $\hat{\mathbf{X}}$ soient des fonctions des poids de calage de première phase w_{1k}, il n'est pas nécessaire que nous les remplacions dans $\hat{\mathbf{B}}^w_{(y;\mathbf{x})}$ et $\hat{\mathbf{X}}$ figurant dans le cinquième terme ; cela subdiviserait simplement le terme d'ordre inférieur $(\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^w_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})})$ en d'autres termes d'ordre inférieur. Par conséquent, nous pouvons laisser tomber le cinquième terme de (4.6) quand les tailles d'échantillon sont suffisamment grandes. Le quatrième terme peut s'écrire comme il suit.

$$(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\Big(\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\Big)^{-1}\sum_s a_k \mathbf{z}_{1k} e_{(y;\mathbf{x})k} = (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})}) + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(y;\mathbf{x}_1)s_1} - \mathbf{B}_{(y;\mathbf{x}_1)}) - (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{x};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)})\mathbf{B}_{(y;\mathbf{x})} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\Big(\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\Big)^{-1} \Big(\sum_s a_k \mathbf{z}_{1k} e_{(y;\mathbf{x})k} - \sum_{s_1} a_{1k}\mathbf{z}_{1k} e_{(y;\mathbf{x})k}\Big). \quad (4.7)$$

Dans cette expression, les quantités sont définies de la façon suivante : $\hat{\mathbf{B}}_{(\mathbf{x};\mathbf{x}_1)} = (\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_k$ et $\hat{\mathbf{B}}_{(y;\mathbf{x}_1)s_1} = (\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\mathbf{z}_{1k} y_k$. La statistique $\hat{\mathbf{B}}_{(y;\mathbf{x}_1)s_1}$ ne peut pas être calculée d'après l'échantillon de première phase, parce que les valeurs de y_k ne sont connues que pour $k \in s$. Elle est définie implicitement dans le but de déterminer la forme linéarisée. Nous pouvons définir ce genre de construction de la même manière que $\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}$ est une fonction de la quantité inconnue $\mathbf{B}_{(y;\mathbf{x})(w)}$. Or, $\hat{\mathbf{B}}_{(y;\mathbf{x}_1)s_1}$ est approximativement sans biais pour sa quantité de population correspondante $\mathbf{B}_{(y;\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\,y_k$. De même, $\hat{\mathbf{B}}_{(\mathbf{x};\mathbf{x}_1)}$ est approximativement sans biais pour $\mathbf{B}_{(\mathbf{x};\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\mathbf{x}'_k$. Comme auparavant, nous pouvons soutenir que les trois derniers termes de (4.7) sont d'ordre plus faible que le premier terme $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})})$, ce qui fournit l'approximation linéaire. L'introduction de ce terme par substitution dans (4.6) mène à la forme linéarisée de $\hat{Y}_{2Pw}$,

$$\hat{Y}_{2Pw\,\mathrm{lin}} = \sum_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{xB}_{(w)};\mathbf{x}_1)}) + \sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\mathbf{x}_1)k}) + \sum_s a_k e_{(y;\mathbf{x})k} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})}). \quad (4.8)$$

Si nous comparons (4.5) et (4.8), nous constatons que $\hat{Y}_{2Pw\,\mathrm{lin}} = \hat{Y}_{2Pa\,\mathrm{lin}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})})$, comme il est énoncé dans le résultat. Cela complète la preuve du résultat 4.1.

Le résultat 4.1 montre qu'en général, les formes linéarisées de $\hat{Y}_{2Pw}$ et $\hat{Y}_{2Pa}$ ne sont pas les mêmes. Cependant, elles le sont sous certaines conditions. Considérons le cas du calage emboîté (à ne pas confondre avec l'échantillonnage emboîté), ce qui signifie que $\mathbf{x}_k$ inclut $\mathbf{x}_{1k}$. Alors, $\mathbf{x}_k$ est de la forme $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$, où le vecteur $\mathbf{x}_{+k}$ est composé des variables restantes. Nous énonçons et prouvons maintenant le résultat suivant.

Résultat 4.2 : Si $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ et $\mathbf{z}_k = (\mathbf{z}'_{1k}, \mathbf{z}'_{+k})'$, alors $\hat{Y}_{2Pw\,\mathrm{lin}} = \hat{Y}_{2Pa\,\mathrm{lin}}$, et $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$ sont asymptotiquement équivalents.

Preuve

La preuve découle du résultat 4.1 en montrant que $\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})} = \mathbf{0}$ sous les conditions spécifiées. Nous avons

$$\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})} = \left(\sum_U \mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1}\left(\sum_U \mathbf{z}_{1k}h_k\right)$$

où $h_k = y_k - \mathbf{x}'_k(\sum_U \mathbf{z}_k\mathbf{x}'_k)^{-1}(\sum_U \mathbf{z}_k y_k)$. Puisque $\sum_U \mathbf{z}_{1k}h_k = \mathbf{0}$ et que nous supposons que $\mathbf{z}_k = (\mathbf{z}'_{1k}, \mathbf{z}'_{+k})'$, il s'ensuit que $\sum_U \mathbf{z}_{1k}h_k = \mathbf{0}$ et $\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\,\mathbf{B}_{(y;\mathbf{x})} = \mathbf{0}$. Donc, il découle du résultat 4.1 que $\hat{Y}_{2Pw\,\mathrm{lin}} = \hat{Y}_{2Pa\,\mathrm{lin}}$. Puisque leurs formes linéaires sont les mêmes, $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$ sont des estimateurs asymptotiquement équivalents.

Il se trouve que le résultat 4.2 requiert seulement que nous incluions $\mathbf{x}_{1k}$ quelque part dans $\mathbf{x}_k$. De toute évidence, il est logique d'inclure $\mathbf{x}_{1k}$ dans la composante $\mathbf{x}_{k(t)}$ de $\mathbf{x}_k$, parce que les totaux de $\mathbf{x}_1$ sont connus. Cependant, nous obtenons le même résultat asymptotique à condition que toutes les variables comprises dans $\mathbf{x}_{1k}$ soient incluses quelque part dans $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$. En pratique, nous constatons souvent que $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$ avec $\mathbf{z}_{1k} = \mathbf{x}_{1k}$ et $\mathbf{z}_k = \mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ où $\mathbf{x}_{+k}$ est le vecteur des variables restantes $\mathbf{x}_{k(w)}$ et $\mathbf{x}_{k(a)}$. Cela satisfait les exigences pour l'équivalence asymptotique de $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$.

Pour étudier les propriétés de $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$, nous nous servons des formes linéarisées données respectivement par (4.5) et (4.8). Moyennant des définitions appropriées pour les résidus e_{0k}, e_{1k} et e_{2k}, nous pouvons représenter $\hat{Y}_{2Pa\,\mathrm{lin}}$ et $\hat{Y}_{2Pw\,\mathrm{lin}}$ comme la somme de trois termes : un terme constant $\sum_U e_{0k}$, un terme d'extension de première phase $\sum_{s_1} a_{1k}e_{1k}$ et un terme d'extension double $\sum_s a_k e_{2k}$,

$$\hat{Y}_{2P\,\mathrm{lin}} = \sum_U e_{0k} + \sum_{s_1} a_{1k}\,e_{1k} + \sum_s a_k\,e_{2k}. \quad (4.9)$$

Cela fait de (4.9) un point de départ convenable pour étudier le biais et la variance asymptotique des deux estimateurs $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$.

Pour la forme linéarisée $\hat{Y}_{2Pa\,\mathrm{lin}}$ donnée par (4.5), les trois quantités résiduelles sont définies comme il suit pour $k \in U$:

$$\begin{aligned}
e_{0k} &= \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} \\
e_{1k} &= \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)} \\
&\quad - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} \\
e_{2k} &= y_k - \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\mathbf{x})(t)} - \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)} \\
&\quad - \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)}. \qquad (4.10)
\end{aligned}$$

Notons que e_{2k} est simplement $e_{(y;\mathbf{x})k}$. De même, pour $\hat{Y}_{2Pw\,\mathrm{lin}}$ donné par (4.8), les résidus ont les définitions suivantes pour $k \in U$:

$$\begin{aligned}
e_{0k} &= \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\mathbf{x})(t)} \\
&\quad + \mathbf{x}'_{1k}(\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} + \mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})}) \\
e_{1k} &= \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)} \\
&\quad - \mathbf{x}'_{1k}(\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} + \mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})}) \\
e_{2k} &= y_k - \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\mathbf{x})(t)} \\
&\quad - \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)} - \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)}. \qquad (4.11)
\end{aligned}$$

Notons que, dans les deux cas, $e_{0k} + e_{1k} + e_{2k} = y_k$ pour chaque k, d'où $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y$. Cette additivité nous permet de prouver à la section 5 que $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$ sont approximativement sans biais. Faute d'espace, dans les sections qui suivent, nous nous concentrons sur les propriétés de $\hat{Y}_{2Pa}$. Cependant, l'analyse est la même pour $\hat{Y}_{2Pw}$ et la méthode d'estimation de la variance proposée à la section 7 peut également être utilisée pour cet estimateur.

4.2 Estimateurs sans calage de deuxième phase ($\mathbf{x}_k = \phi$)

En l'absence de calage de deuxième phase ($\mathbf{x}_k = \phi$), nous avons $w_k = a_k^*$. Conséquemment, les poids finaux sont soit $w_k = a_k = a_{1k}\,a_{2k}$ soit $w_k = w_{1k}\,a_{2k}$. La première option donne l'estimateur à double facteur d'extension $\sum_s a_k\,y_k$. La deuxième produit un estimateur différent qui est habituellement moins efficace. Cependant, ces estimateurs sont généralement tous deux inefficaces

comparativement à ceux obtenus en exécutant le calage de deuxième phase. La forme linéarisée de l'estimateur à deux phases avec $w_k = w_{1k}\, a_{2k}$ s'obtient en l'écrivant sous la forme

$$\begin{aligned}\hat{Y}_{2P} &= \mathbf{X}_1' \mathbf{B}_{(y;\,\mathbf{x}_1)} - \hat{\mathbf{X}}_1' \mathbf{B}_{(y;\,\mathbf{x}_1)} + \sum_s a_k y_k \\ &\quad + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1} - \mathbf{B}_{(y;\,\mathbf{x}_1)}) \\ &\quad + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \Big(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}'\Big)^{-1} \\ &\quad \Big(\sum_s a_k \mathbf{z}_{1k} y_k - \sum_{s_1} a_{1k} \mathbf{z}_{1k} y_k\Big). \qquad (4.12)\end{aligned}$$

Les termes $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1}$ et $\mathbf{B}_{(y;\,\mathbf{x}_1)}$ sont définis à la section précédente. Quand les échantillons sont suffisamment grands, $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}')^{-1} (\sum_s a_k \mathbf{z}_{1k} y_k - \sum_{s_1} a_{1k} \mathbf{z}_{1k} y_k)$ et $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\ (\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1} - \mathbf{B}_{(y;\,\mathbf{x}_1)})$ sont d'ordre plus faible et peuvent être ignorés. Nous obtenons ainsi la forme linéarisée de l'estimateur

$$\hat{Y}_{2P\,\mathrm{lin}} = \mathbf{X}_1' \mathbf{B}_{(y;\,\mathbf{x}_1)} - \hat{\mathbf{X}}_1 \mathbf{B}_{(y;\,\mathbf{x}_1)} + \sum_s a_k\, y_k. \quad (4.13)$$

Nous pouvons également écrire cette forme linéarisée comme une somme (4.9) de trois termes résiduels, les résidus e_{0k}, e_{1k} et e_{2k} ayant les définitions suivantes pour $k \in U$.

$$\begin{aligned} e_{0k} &= \mathbf{x}_{1k}'\ \mathbf{B}_{(y;\,\mathbf{x}_1)} \\ e_{1k} &= -\mathbf{x}_{1k}' \mathbf{B}_{(y;\,\mathbf{x}_1)} \\ e_{2k} &= y_k. \qquad (4.14)\end{aligned}$$

Ces résidus ressemblent à ceux donnés par (4.10) si nous posons que $\mathbf{x}_k = \phi$ et que nous supprimons $\mathbf{B}_{(y;\,\mathbf{x})}$. Notons que $\mathbf{B}_{(y;\,\mathbf{x}_1)}$ joue le même rôle que $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}$ dans (4.10). Comme auparavant, $e_{0k} + e_{1k} + e_{2k} = y_k$ pour chaque k, d'où $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y$.

L'estimateur à double facteur d'extension est un cas particulier de cet estimateur quand nous avons aussi $\mathbf{x}_{1k} = \phi$. Cela signifie que $\mathbf{B}_{(y;\,\mathbf{x}_1)}$ n'est pas défini. Les définitions correspondantes pour e_{0k}, e_{1k} et e_{2k} sont simplement $e_{0k} = 0$, $e_{1k} = 0$ et $e_{2k} = y_k$ pour $k \in U$.

Aux sections qui suivent, nous examinons le biais et la variance de l'estimateur par calage à deux phases $\hat{Y}_{2Pa}$ et nous proposons une nouvelle méthode d'estimation de la variance. Nous pouvons obtenir les résultats correspondants en l'absence de calage de deuxième phase, parce que les résidus pour ces deux groupes d'estimateurs ont des propriétés et une forme linéarisée semblables. La seule différence tient à l'estimation de la variance. Nous utilisons le même estimateur de variance (décrit à la section 7), et nous estimons les résidus en utilisant $\hat{e}_{1k} = -\mathbf{x}_{1k}' \hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s}$ où $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s} = (\sum_s a_k\ \mathbf{z}_{1k}\ \mathbf{x}_{1k}')^{-1}\ \sum_s a_k\ \mathbf{z}_{1k}\ y_k$, et $\hat{e}_{2k} = y_k$.

5. Biais et variance de l'estimateur par calage à deux phases $\hat{Y}_{2Pa}$

L'estimateur par calage à deux phases $\hat{Y}_{2Pa} = \sum_s w_k y_k$ est approximativement sans biais pour $Y = \sum_U y_k$. Pour le montrer, nous calculons l'espérance de la forme linéarisée donnée par (4.9) par la méthode habituelle de conditionnement sur l'échantillon de première phase s_1. Nous avons $E(\sum_s a_k e_{2k}) = E_{s_1} E_{s|s_1} (\sum_s a_k e_{2k}) = E_{s_1} (\sum_{s_1} a_{1k} e_{2k}) = \sum_U e_{2k}$, $E_{s_1} (\sum_{s_1} a_{1k} e_{1k}) = \sum_U e_{1k}$, et $\sum_U e_{0k}$ est un terme constant, d'où

$$E(\hat{Y}_{2Pa\,\mathrm{lin}}) = \sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y.$$

Cela montre que $\hat{Y}_{2Pa\,\mathrm{lin}}$ est sans biais pour Y. En vertu de (4.4), $\hat{Y}_{2Pa} = \hat{Y}_{2Pa\,\mathrm{lin}} + R$, de sorte que le biais de $\hat{Y}_{2Pa}$ est égal à l'espérance de R, qui est la somme des deux termes d'ordre inférieur $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)})$ et $(\mathbf{X} - \hat{\mathbf{X}})' (\hat{\mathbf{B}}_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})})$. Comme nous l'avons mentionné à la section 4, l'espérance de chacun de ces termes est presque nulle. Il s'ensuit que $\hat{Y}_{2Pa}$ est approximativement sans biais pour Y.

La variance de $\hat{Y}_{2Pa} = \sum_s w_k y_k$ est approximée de près par la variance de la forme linéarisée $\hat{Y}_{2Pa\,\mathrm{lin}}$ donnée par (4.9) avec les résidus définis par (4.10). Son premier terme, $\sum_U e_{0k}$, est constant et ne contribue pas à la variance. Donc,

$$V(\hat{Y}_{2Pa\,\mathrm{lin}}) = V\Big(\sum_{s_1} a_{1k}\, e_{1k} + \sum_s a_k\, e_{2k}\Big). \quad (5.1)$$

Nous utilisons (5.1) comme point de départ pour déterminer un estimateur de variance pour $\hat{Y}_{2Pa\,\mathrm{lin}}$. Deux approches distinctes peuvent être suivies et leur comparaison est intéressante. Celle décrite à la section 7 est nouvelle et plus intéressante que celle décrite à la section 8, obtenue par la méthode classique de conditionnement sur l'échantillon de première phase s_1, parce qu'elle produit un estimateur de variance plus efficace. Les résidus e_{1k} et e_{2k} donnés par (4.10) jouent un rôle important dans les deux dérivations.

6. Préliminaires de l'estimation de la variance

Notre objectif est d'estimer la variance $V(\hat{Y}_{2Pa\,\mathrm{lin}})$ donnée par (5.1). Nous le faisons aux sections 7 et 8 en nous servant de deux arguments différents. Les résidus e_{1k} et e_{2k} sont définis pour tout $k \in U$, mais ils ne peuvent pas être calculés. Ils doivent être remplacés par les estimations $\hat{e}_{1k}$ et $\hat{e}_{2k}$. Ces estimations, formées à l'image de (4.10), sont

$$\hat{e}_{1k} = \mathbf{x}'_{k(a)}\hat{\mathbf{B}}_{(y;\mathbf{x})(a)} + \mathbf{x}'_{k(w)}\hat{\mathbf{B}}_{(y;\mathbf{x})(w)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\mathbf{x}_1)} \quad \text{for } k \in s_1$$

$$\hat{e}_{2k} = y_k - \mathbf{x}'_k\,\hat{\mathbf{B}}_{(y;\mathbf{x})} = y_k - \mathbf{x}'_{k(t)}\,\hat{\mathbf{B}}_{(y;\mathbf{x})(t)} - \mathbf{x}'_{k(w)}\hat{\mathbf{B}}_{(y;\mathbf{x})(w)} - \mathbf{x}'_{k(a)}\hat{\mathbf{B}}_{(y;\mathbf{x})(a)} \quad \text{for } k \in s \tag{6.1}$$

où

$$\hat{\mathbf{B}}_{(y;\mathbf{x})} = \left(\sum_s a_k\,\mathbf{z}_k\,\mathbf{x}'_k\right)^{-1}\sum_s a_k\,\mathbf{z}_k\,y_k = (\hat{\mathbf{B}}'_{(y;\mathbf{x})(t)}, \hat{\mathbf{B}}'_{(y;\mathbf{x})(w)}, \hat{\mathbf{B}}'_{(y;\mathbf{x})(a)})'$$

et (6.2)

$$\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\mathbf{x}_1)} = \left(\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_{1k}\right)^{-1}\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_{k(w)}\hat{\mathbf{B}}_{(y;\mathbf{x})(w)}.$$

Dans la définition de $\hat{e}_{1k}$, le terme $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\mathbf{x}_1)}$ est l'estimation de $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)}$ dans (4.10). Deux remplacements sont requis dans $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}$ pour arriver à $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\mathbf{x}_1)}$: pour commencer, les sommes sur U sont remplacées par les sommes pondérées appropriées sur s_1, ce qui donne $\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} = (\sum_{s_1} a_{1k}\mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)}$. Dans cette expression, $\mathbf{B}_{(y;\mathbf{x})(w)}$ est encore inconnu, de sorte que nous le remplaçons par son estimation $\hat{\mathbf{B}}_{(y;\mathbf{x})(w)}$ pour arriver à $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\mathbf{x}_1)}$.

Un point important à souligner est que les estimations $\hat{e}_{1k}$ peuvent être obtenues pour $k \in s_1$, parce que $\mathbf{x}_{k(a)}$, $\mathbf{x}_{k(w)}$ et $\mathbf{x}_{1k}$ sont toutes connues pour $k \in s_1$, mais que les estimations $\hat{e}_{2k}$ ne peuvent être calculées que pour $k \in s$, parce que y_k n'est disponible que pour $k \in s$. Le fait que les estimations $\hat{e}_{1k}$ soient disponibles pour $k \in s_1$ plutôt que pour $k \in s$ nous permet de construire (à la section 7) un estimateur plus efficace de $V(\hat{Y}_{2Pa\,\mathrm{lin}})$ que celui obtenu selon la méthode classique d'estimation de la variance (à la section 8) où tous les résidus estimés sont calculés uniquement pour $k \in s$.

Les poids de sondage $a_{1k} = 1/\pi_{1k}$, $a_{2k} = 1/\pi_{2k}$ et $a_k = a_{1k}\,a_{2k}$ sont définis à la section 1. Dans les sections qui suivent, nous avons également besoin des quantités données ci-dessous, définies comme les fonctions des probabilités d'inclusion de deuxième ordre $\pi_{1k\ell} = \Pr(k\ \&\ \ell \in s_1)$ et $\pi_{2k\ell} = \Pr(k\ \&\ \ell \in s \mid s_1)$:

$$a_{1k\ell} = 1/\pi_{1k\ell},\ a_{2k\ell} = 1/\pi_{2k\ell},\ a_{k\ell} = a_{1k\ell}a_{2k\ell}$$
$$D_{1k\ell} = a_{1k}a_{1\ell} - a_{1k\ell},\ D_{2k\ell} = a_{2k}a_{2\ell} - a_{2k\ell},$$
$$D_{k\ell} = a_k a_\ell - a_{k\ell}.$$

Ici, $\pi_{2k\ell}$ et $a_{2k\ell}$ sont conditionnés sur l'échantillon s_1. Nous supposons que toutes les probabilités d'inclusion de premier et de deuxième ordre sont positives. En utilisant cette notation et les résultats susmentionnés, nous allons maintenant développer deux estimateurs de variance différents aux deux sections suivantes.

7. Estimateur de variance avec résidus distincts

La variance de $\hat{Y}_{2Pa\,\mathrm{lin}}$ est donnée par (5.1), où e_{1k} et e_{2k} sont définis par (4.10). Elle peut être développée sous la forme

$$V(\hat{Y}_{2Pa\,\mathrm{lin}}) = V\left(\sum_{s_1} a_{1k}\,e_{1k}\right) + V\left(\sum_s a_k\,e_{2k}\right) + 2\,\mathrm{Cov}\left(\sum_{s_1} a_{1k}\,e_{1k}, \sum_s a_k\,e_{2k}\right). \tag{7.1}$$

Si nous connaissions les résidus e_{1k} et e_{2k}, des estimations sans biais de ces trois composantes seraient données, respectivement, par

$$\sum_{k\in s_1}\sum_{\ell\in s_1} D_{1k\ell}\,e_{1k}\,e_{1\ell},$$
$$\sum_{k\in s}\sum_{\ell\in s} D_{k\ell}\,e_{2k}\,e_{2\ell},$$
$$2\sum_{k\in s_1}\sum_{\ell\in s} D_{1k\ell}\,a_{2\ell}\,e_{1k}\,e_{2\ell}. \tag{7.2}$$

La preuve de l'absence de biais est semblable pour les trois composantes. Par exemple, pour la deuxième, nous avons

$$E_{s_1}E_{s|s_1}\left(\sum_{k\in s}\sum_{\ell\in s} D_{k\ell}\,e_{2k}e_{2\ell}\right) = E_{s_1}\left(\sum_{k\in s_1}\sum_{\ell\in s_1}(D_{k\ell}/a_{2k\ell})e_{2k}e_{2\ell}\right) = \sum_{k\in U}\sum_{\ell\in U}(D_{k\ell}/a_{k\ell})e_{2k}e_{2\ell} = \sum_{k\in U}\sum_{\ell\in U}(a_k a_\ell/a_{k\ell})e_{2k}e_{2\ell} - \left(\sum_U e_{2k}\right)^2 = E\left[\left(\sum_s a_k e_{2k}\right)^2\right] - \left[E\left(\sum_s a_k e_{2k}\right)\right]^2 = V\left(\sum_s a_k e_{2k}\right).$$

Nous remplaçons maintenant dans (7.2) les résidus inconnus par les estimations respectives données par (6.1), c'est-à-dire e_{1k} par $\hat{e}_{1k}$ pour $k \in s_1$ et e_{2k} par $\hat{e}_{2k}$ pour $k \in s$. Puis, nous additionnons les trois composantes résultantes pour obtenir l'estimateur de variance avec « résidus distincts »

$$\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}}) = \sum_{k\in s_1}\sum_{\ell\in s_1} D_{1k\ell}\,\hat{e}_{1k}\,\hat{e}_{1\ell} + \sum_{k\in s}\sum_{\ell\in s} D_{k\ell}\,\hat{e}_{2k}\,\hat{e}_{2\ell} + 2\sum_{k\in s_1}\sum_{\ell\in s} D_{1k\ell}a_{2\ell}\,\hat{e}_{1k}\hat{e}_{2\ell}. \qquad (7.3)$$

Le terme « résidus distincts » et l'indice inférieur correspondant *sr* pour (*separate residual*) reflètent le fait que, dans (7.3), les résidus demeurent distincts, $\hat{e}_{1k}$ étant défini sur le grand échantillon s_1 et $\hat{e}_{2k}$, sur l'échantillon plus petit s. Le fait que les résultats calculés pour le grand échantillon s_1 peuvent être avantageux pour l'estimation de la variance a été reconnu par Axelson (1998). Cependant, sa méthode de calcul diffère de notre approche par calage fondée sur $\mathbf{x}_{1k}$ et $\mathbf{x}_k$. La méthode d'estimation de la variance de l'estimateur par la régression à deux phases présentée dans Hidiroglou, Rao et Haziza (2006) possède certains traits en commun avec notre approche, mais il existe aussi d'importantes différences.

8. Estimateur de la variance à résidu combiné

Nous arrivons à l'expression (7.3) en reconnaissant que les estimations $\hat{e}_{1k}$ peuvent être obtenues pour $k \in s_1$. L'approche classique, passée en revue ici, consiste à obtenir un estimateur de variance par conditionnement sur l'échantillon de première phase s_1. On obtient un estimateur de variance où tous les résidus requis sont définis pour $k \in s$. Plus tard, nous le comparerons à l'estimateur plus efficace (7.3). Partant de (5.1), nous conditionnons sur l'échantillon de première phase s_1 pour obtenir

$$\begin{aligned} V(\hat{Y}_{2P\,a\,\mathrm{lin}}) &= V_{s_1}E_{s|s_1}\left(\sum_{s_1} a_{1k}e_{1k} + \sum_s a_k e_{2k}\right) \\ &\quad + E_{s_1}V_{s|s_1}\left(\sum_{s_1} a_{1k}e_{1k} + \sum_s a_k e_{2k}\right) \\ &= V_{s_1}\left(\sum_{s_1} a_{1k}e_{1k} + \sum_{s_1} a_{1k}e_{2k}\right) \\ &\quad + E_{s_1}V_{s|s_1}\left(\sum_s a_k e_{2k}\right) \\ &= V_{s_1}\left(\sum_{s_1} a_{1k}e_{12k}\right) + E_{s_1}V_{s|s_1}\left(\sum_s a_k e_{2k}\right) \qquad (8.1) \end{aligned}$$

où $e_{12k} = e_{1k} + e_{2k}$ est appelé le résidu combiné. À partir de (4.10), nous obtenons les expressions suivantes.

$$\begin{aligned} e_{12k} &= y_k - \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y\,;\,\mathbf{x})(t)} - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)}\,;\,\mathbf{x}_1)} \\ e_{2k} &= y_k - \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y\,;\,\mathbf{x})(t)} - \mathbf{x}'_{k(w)}\mathbf{B}_{(y\,;\,\mathbf{x})(w)} \\ &\quad - \mathbf{x}'_{k(a)}\mathbf{B}_{(y\,;\,\mathbf{x})(a)}. \qquad (8.2) \end{aligned}$$

Il est simple de définir les estimateurs des deux composantes $V_{s_1}(\sum_{s_1} a_{1k}\,e_{12k})$ et $E_{s_1}V_{s|s_1}(\sum_s a_k\,e_{2k})$. Chacune a la forme d'une double somme sur s, parce que e_{12k} et e_{2k} contiennent y_k qui n'est disponible que pour $k \in s$. La première composante utilise $\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k} = y_k - \mathbf{x}'_{k(t)}\hat{\mathbf{B}}_{(y\,;\,\mathbf{x})(t)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)}\,;\,\mathbf{x}_1)}$ pour $k \in s$. Nous obtenons alors $\sum_{k\in s}\sum_{\ell\in s} D_{1k\ell}\,a_{2k\ell}\,\hat{e}_{12k}\,\hat{e}_{12\ell}$ comme estimateur de $V_{s_1}(\sum_{s_1} a_{1k}\,e_{12k})$.

Pour la deuxième composante, nous utilisons les estimations des résidus $\hat{e}_{2k} = y_k - \mathbf{x}'_k\hat{\mathbf{B}}_{(y\,;\,\mathbf{x})}$ données par (6.1) pour $k \in s$, et nous obtenons $\sum_{k\in s}\sum_{\ell\in s} D_{2k\ell}\,a_{1k}\,a_{1\ell}\,\hat{e}_{2k}\,\hat{e}_{2\ell}$ comme estimateur de $E_{s_1}V_{s|s_1}(\sum_s a_k e_{2k})$. Par sommation des deux termes estimés, nous obtenons l'estimateur de variance qui suit, où l'indice inférieur *cr* pour (*combined residual*) signifie « résidu combiné »,

$$\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}}) = \sum_{k\in s}\sum_{\ell\in s} D_{1k\ell}\,a_{2k\ell}\,\hat{e}_{12k}\,\hat{e}_{12\ell} + \sum_{k\in s}\sum_{\ell\in s} D_{2k\ell}\,a_{1k}\,a_{1\ell}\,\hat{e}_{2k}\,\hat{e}_{2\ell}. \qquad (8.3)$$

Examinons maintenant en quoi (7.3) et (8.3) diffèrent. L'estimateur de variance à résidus distincts (7.3) a pour point de départ le développement $V(\hat{Y}_{2P\,a\,\mathrm{lin}}) = V(\sum_{s_1} a_{1k}\,e_{1k}) + V(\sum_s a_k\,e_{2k}) + 2\,\mathrm{Cov}(\sum_{s_1} a_{1k}\,e_{1k},\ \sum_s a_k\,e_{2k})$. Nous estimons ces trois composantes séparément en tant que fonctions des résidus e_{1k} et e_{2k}. L'expression résultante de la variance contient trois termes : une double somme sur s_1 en ce qui a trait à e_{1k} et $e_{1\ell}$, une double somme sur s en ce qui a trait à e_{2k} et $e_{2\ell}$, ainsi qu'une somme croisée sur s_1 et s en ce qui a trait à $e_{1k} \in s_1$ et $e_{2\ell} \in s$. Enfin, nous arrivons à (7.3) en estimant e_{1k} par $\hat{e}_{1k}$ pour $k \in s_1$ et e_{2k} par $\hat{e}_{2k}$ pour $k \in s$.

L'estimateur de variance à résidu combiné (8.3) découle du conditionnement classique sur l'échantillon de première phase s_1 sous la forme $V(\hat{Y}_{2P\,a\,\mathrm{lin}}) = V_{s_1}E_{s|s_1}(\hat{Y}_{2P\,a\,\mathrm{lin}}) + E_{s_1}V_{s|s_1}(\hat{Y}_{2P\,a\,\mathrm{lin}})$. Cela nous amène à combiner e_{1k} et e_{2k} sous la forme $e_{12k} = e_{1k} + e_{2k}$ dans le premier terme. Le deuxième terme, $E_{s_1}V_{s|s_1}(\hat{Y}_{2P\,a\,\mathrm{lin}})$, est une fonction de e_{2k}. Puisque e_{12k} et e_{2k} ne peuvent être estimés que sur s, l'estimateur de variance résultant devient une somme de deux termes, exprimés chacun sous la forme d'une double somme sur s.

L'estimateur à résidus distincts (7.3) est plus efficace que l'estimateur à résidu combiné (8.3), parce qu'il est fondé sur les résidus $\hat{e}_{1k}$ obtenus pour l'échantillon habituellement plus grand s_1. L'avantage de (7.3) sur (8.3) est illustré par la simulation décrite à la section 10. L'approche qui sous-tend l'estimateur de variance à résidus distincts (7.3) peut être étendue à l'échantillonnage à deux phases et à d'autres plans de sondage complexes. Dans ces extensions de la méthode, nous procédons de manière semblable, en commençant par l'établissement de la forme linéarisée par développement des composantes de variance et la détermination des résidus appropriés.

9. Comparaison avec l'estimateur par la régression à deux phases

Särndal, Swensson et Wretman (1992) ont élaboré un estimateur par la régression à deux phases pour $Y = \sum_U y_k$ en se basant sur un article antérieur publié par Särndal et Swensson (1989). Il est utile de voir comment cet estimateur, désigné ici par $\hat{Y}_{\text{reg}}$, se compare à l'estimateur par calage $\hat{Y}_{2P}$ que nous avons considéré dans les sections précédentes. Même s'ils sont fondés sur la même information auxiliaire, les deux estimateurs sont « proches », mais pas identiques, parce que l'estimateur $\hat{Y}_{2P}$ est obtenu par calage à chacune des deux phases d'échantillonnage, tandis que l'estimateur par la régression à deux phases $\hat{Y}_{\text{reg}}$ est obtenu par un raisonnement assisté par modèle.

Nous décrivons maintenant l'estimateur par la régression à deux phases de Särndal, Swensson et Wretman (1992). Leur méthode de calcul comprend l'ajustement de deux modèles de régression linéaire à l'aide des données auxiliaires disponibles, l'un au « niveau supérieur » et l'autre au « niveau inférieur ». Ces auteurs ont élaboré un estimateur de variance correspondant, en suivant l'argument de conditionnement classique. Nous comparons leur estimateur de variance avec l'estimateur de variance à résidu combiné (8.3), également obtenu selon l'argument de conditionnement. Les deux estimateurs de variance ne concordent pas exactement, parce que les estimateurs ponctuels sont légèrement différents, mais ils sont numériquement proches, comme nous le montrons ici.

Soit $\mathbf{x}_{1k}$ un vecteur de variables auxiliaires dont les totaux de population sont connus et soit $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, où $\mathbf{x}_{1k}$ ainsi que $\mathbf{x}_{2k}$ sont des valeurs connues du vecteur pour $k \in s_1$. Nous supposons que le total $\sum_U \mathbf{x}_{1k}$ est connu, tandis que le total $\sum_U \mathbf{x}_{2k}$ est inconnu. Les valeurs prévues produites pour $k \in s_1$ par les deux régressions ajustées au « niveau supérieur » et au « niveau inférieur » sont données respectivement par

$$\hat{y}_{1k} = \mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s}$$

avec

$$\hat{\mathbf{B}}_{1s} = \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1}\left(\sum_s a_k \mathbf{x}_{1k} y_k/\sigma^2_{1k}\right) \quad (9.1)$$

et

$$\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}_s$$

avec

$$\hat{\mathbf{B}}_s = \left(\sum_s a_k \mathbf{x}_k\mathbf{x}'_k/\sigma^2_k\right)^{-1}\sum_s a_k \mathbf{x}_k y_k/\sigma^2_k. \quad (9.2)$$

L'estimateur par la régression à deux phases résultant $\hat{Y}_{\text{reg}}$ de $Y = \sum_U y_k$ est

$$\hat{Y}_{\text{reg}} = \left(\sum_U \mathbf{x}_{1k}\right)'\hat{\mathbf{B}}_{1s} + \sum_{s_1} a_{1k}(\hat{y}_k - \hat{y}_{1k}) + \sum_s a_k(y_k - \hat{y}_k). \quad (9.3)$$

L'estimateur $\hat{Y}_{\text{reg}}$ peut-il être interprété comme étant un estimateur par calage ? Pour répondre à cette question, déterminons les poids implicites dans (9.3). Nous pouvons écrire $\hat{Y}_{\text{reg}} = \sum_s w_k y_k$, avec les poids w_k identifiés en introduisant (9.1) et (9.2) par substitution dans (9.3) et en simplifiant. Nous trouvons que $w_k = a_k g_k = a_{1k} a_{2k} g_k$, où le facteur de calage g_k est donné pour $k \in s$ par

$$g_k = 1 + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1}\mathbf{x}_{1k}/\sigma^2_{1k} + \left(\sum_{s_1} a_{1k}\mathbf{x}_k - \sum_s a_k \mathbf{x}_k\right)' \left(\sum_s a_k \mathbf{x}_k\mathbf{x}'_k/\sigma^2_k\right)^{-1}\mathbf{x}_k/\sigma^2_k. \quad (9.4)$$

Les poids w_k ne sont pas énoncés explicitement dans Särndal, Swensson et Wretman (1992). Dans quel sens, si tant est qu'il y en ait un, pouvons-nous considérer w_k comme un poids de calage ? Pour étudier cette question, nous commençons par remplacer y_k dans (9.3) par $\mathbf{x}'_{1k}$. L'utilisation de (9.1) et (9.2) avec $y_k = \mathbf{x}'_{1k}$ donne $\sum_U \mathbf{x}'_{1k}$ comme deuxième membre de (9.3). Donc, les poids $w_k = a_k g_k$ satisfont $\sum_s w_k \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Ensuite, nous remplaçons y_k dans (9.3) par $\mathbf{x}'_{2k}$, en utilisant de nouveau (9.1) et (9.2) pour obtenir

$$\sum_{s_1} a_{1k}\mathbf{x}'_{2k} + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1} \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{2k}/\sigma^2_{1k}\right). \quad (9.5)$$

Bien que (9.5) soit une estimation approximativement sans biais du total de $\mathbf{x}_{2k}$ inconnu $\sum_U \mathbf{x}'_{2k}$, elle n'a pas la forme habituelle du deuxième membre d'une équation de calage de deuxième phase, telle que $\sum_{s_1} a_{1k}\mathbf{x}'_{2k}$ ou $\sum_{s_1} w_{1k}\mathbf{x}'_{2k}$. Cependant, elle est proche. Si nous remplaçons les deux sommes sur s par les sommes pondérées appropriées sur s_1, alors (9.5) devient $\sum_{s_1} w_{1k}\mathbf{x}'_{2k}$ où w_{1k} est donné par (2.1) avec $\mathbf{z}_{1k} = \mathbf{x}_{1k}/\sigma^2_{1k}$. Donc, les poids implicites w_k dans $\hat{Y}_{\text{reg}}$ donnent le calage exact sur le total de population de $\mathbf{x}_{1k}$ connu et un calage qui est proche de celui sur le total de $\mathbf{x}_{2k}$ estimé $\sum_{s_1} w_{1k}\mathbf{x}'_{2k}$. Cela donne à penser que $\hat{Y}_{\text{reg}}$ devrait avoir des propriétés semblables à celles d'un estimateur $\hat{Y}_{2P}$ obtenu en définissant $\mathbf{x}_k$ dans $\hat{Y}_{2P}$ comme étant $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ avec $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$, $\mathbf{x}_{k(w)} = \mathbf{x}_{2k}$ et $\mathbf{x}_{k(a)} = \phi$. En outre, la forme de l'estimateur assisté par modèle implique que $\mathbf{z}_{1k} = \mathbf{x}_{1k}/\sigma^2_{1k}$ et $\mathbf{z}_k = \mathbf{x}_k/\sigma^2_k$. Puisque $\mathbf{x}_k$ inclut $\mathbf{x}_{1k}$, il est raisonnable de définir $\mathbf{z}_k = \mathbf{x}_k/\sigma^2_k$ sous la forme $\mathbf{z}_k = (\mathbf{x}'_{1k}/\sigma^2_{1k}, \mathbf{x}'_{2k}/\sigma^2_{2k})'$. Ces spécifications satisfont les exigences pour l'équivalence asymptotique de $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$, de sorte que nous n'avons

pas à nous inquiéter du choix des poids de départ dans $\hat{Y}_{2P}$. Nous pouvons simplement travailler avec $\hat{Y}_{2Pa}$ comme estimateur comparable à $\hat{Y}_{\text{reg}}$. Maintenant, examinons une estimation de variance pour $\hat{Y}_{\text{reg}}$ et l'estimateur $\hat{Y}_{2Pa}$ sous ces spécifications.

L'estimateur de variance de Särndal, Swensson et Wretman (1992) contient des facteurs de calage désignés par g_{ks} et g_{1ks_1}. Il ne faut pas les confondre avec g_k donné par (9.4). Si nous écartons g_{ks} et g_{1ks_1}, qui ont tous deux une valeur proche de un et dont l'incidence numérique est limitée, leur estimateur de variance est donné par

$$\hat{V}(\hat{Y}_{\text{reg}}) = \sum_{k\in s}\sum_{\ell\in s} D_{1k\ell}\, a_{2k\ell}\, \hat{e}_{1ks}\, \hat{e}_{1\ell s} + \sum_{k\in s}\sum_{\ell\in s} D_{2k\ell}\, a_{1k}\, a_{1\ell}\, \hat{e}_{ks}\, \hat{e}_{\ell s} \qquad (9.6)$$

où, pour $k \in s$,

$$\hat{e}_{1ks} = y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s} \text{ et } \hat{e}_{ks} = y_k - \mathbf{x}'_k\,\hat{\mathbf{B}}_s. \qquad (9.7)$$

Les deux composantes de (9.6) sont des doubles sommes sur s, qui reflètent le fait que $\hat{e}_{1ks}$ ainsi que $\hat{e}_{ks}$ ne peuvent être obtenus que pour $k \in s$. La formule (9.6) ressemble à la formule (8.3) pour l'estimateur à résidu combiné, mais à quel point les résidus sont-ils différents dans les deux formules ? Examinons les résidus pour l'estimateur ponctuel comparable. Comme nous l'avons mentionné plus haut, cet estimateur $\hat{Y}_{2P}$ possède $\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})$ avec $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$, $\mathbf{x}_{k(w)} = \mathbf{x}_{2k}$, $\mathbf{x}_{k(a)} = \phi$, $\mathbf{z}_{1k} = \mathbf{x}_{1k}/\sigma^2_{1k}$ et $\mathbf{z}_k = \mathbf{x}_k/\sigma^2_k = (\mathbf{x}'_{1k}/\sigma^2_{1k}, \mathbf{x}'_{2k}/\sigma^2_{2k})'$. Sous ces spécifications, les résidus $\hat{e}_{1k}$ et $\hat{e}_{2k}$ qui figurent dans (6.1) sont donnés par

$$\hat{e}_{1k} = \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\,\mathbf{x}_1)} \quad \text{pour } k \in s_1$$

$$\hat{e}_{2k} = y_k - \mathbf{x}'_k\,\hat{\mathbf{B}}_{(y;\,\mathbf{x})}$$

$$= y_k - \mathbf{x}'_{1k}\,\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} - \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)} \quad \text{pour } k \in s \qquad (9.8)$$

où $\hat{\mathbf{B}}_{(y;\,\mathbf{x})} = (\hat{\mathbf{B}}'_{(y;\,\mathbf{x})(1)}, \hat{\mathbf{B}}'_{(y;\,\mathbf{x})(2)})'$ correspond au partitionnement de $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, et de (6.2) il découle que

$$\hat{\mathbf{B}}_{(y;\mathbf{x})} = \left(\sum_s a_k \mathbf{z}_k \mathbf{x}'_k\right)^{-1}\left(\sum_s a_k \mathbf{z}_k y_k\right)$$

$$\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\mathbf{x}_1)} = \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\Big/\sigma^2_{1k}\right)^{-1} \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}\Big/\sigma^2_{1k}\right). \qquad (9.9)$$

Les résidus $\hat{e}_{2k}$ figurant dans (9.8) sont les mêmes que $\hat{e}_{ks}$ dans (9.7). Mais comment les résidus $\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k}$, obtenus par addition dans (9.8), sont-ils reliés à leurs homologues $\hat{e}_{1ks}$ dans (9.7) ? Pour découvrir ce lien, nous commençons par montrer que $\hat{\mathbf{B}}_{1s} = (\sum_s a_k\, \mathbf{x}_{1k}\, \mathbf{x}'_{1k}/\sigma^2_{1k})^{-1}$ $\sum_s a_k\, \mathbf{x}_{1k}\, y_k/\sigma^2_{1k}$ peut s'écrire sous la forme

$$\hat{\mathbf{B}}_{1s} = \hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} + \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1}\left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}/\sigma^2_{1k}\right). \qquad (9.10)$$

Pour le démontrer, nous partons de $\hat{\mathbf{B}}_{(y;\,\mathbf{x})}$, qui par définition satisfait $\sum_s a_k \mathbf{z}_k y_k = (\sum_s a_k \mathbf{z}_k \mathbf{x}'_k)\hat{\mathbf{B}}_{(y;\,\mathbf{x})}$. Cette égalité peut aussi s'écrire sous la forme $\sum_s a_k \mathbf{z}_k y_k = \sum_s a_k \mathbf{z}_k(\mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} + \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)})$. Puisque $\mathbf{z}_k = (\mathbf{x}'_{1k}/\sigma^2_{1k}, \mathbf{x}'_{2k}/\sigma^2_{2k})'$, la composante de cette équation correspondant à $\mathbf{x}_{1k}$ est $\sum_s a_k \mathbf{x}_{1k} y_k/\sigma^2_{1k} = \sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)}/\sigma^2_{1k} + \sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}/\sigma^2_{1k}$. En prémultipliant les deux membres par $(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k})^{-1}$, nous obtenons (9.10).

Puis, en partant de (9.8) et en utilisant la définition de $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\,\mathbf{x}_1)}$ donnée par (9.9), nous obtenons

$$\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k}$$

$$= y_k - \mathbf{x}'_{1k}\,\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\,\mathbf{x}_1)}$$

$$= y_k - \mathbf{x}'_{1k}\left\{\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} + \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\Big/\sigma^2_{1k}\right)^{-1} \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}\Big/\sigma^2_{1k}\right)\right\}.$$

Dans l'expression entre accolades, remplaçons les deux sommes pondérées par a_{1k} sur s_1 par les sommes pondérées par a_k correspondantes sur s ; le résultat est égal à $\hat{\mathbf{B}}_{1s}$ tel qu'il est donné par (9.10). Cela signifie que $\hat{e}_{12k} \cong y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s} = \hat{e}_{1ks}$. En résumé, $\hat{e}_{12k} \cong \hat{e}_{1ks}$ pour $k \in s$ et $\hat{e}_{2k} = \hat{e}_{ks}$ pour $k \in s$. Donc, l'estimateur de variance (9.6) pour l'estimateur par la régression à deux phases $\hat{Y}_{\text{reg}}$ devrait être numériquement proche de l'estimateur de variance à résidu combiné (8.3) pour l'estimateur par calage $\hat{Y}_{2P}$ défini à la présente section. Nous appuyons empiriquement ceci au moyen de la simulation décrite à la section suivante.

10. Simulation

Dans cette section, nous présentons une petite simulation pour valider l'allégation que l'estimateur de variance à résidus distincts $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$ donné par (7.3) peut être considérablement plus efficace que l'estimateur de variance à résidu combiné $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$ donné par (8.3), et que le comportement de ce dernier est semblable à celui de l'estimateur par la régression à deux phases $\hat{V}(\hat{Y}_{\text{reg}})$ donné par (9.6). Nous avons créé une population de $N = 5\,000$ unités en deux étapes de la façon suivante : en premier lieu,

nous avons produit les valeurs (u_{1k}, u_{2k}) pour $k = 1, 2, ..., 5\,000$ par 5 000 réalisations des variables aléatoires indépendantes $u_{1k} \sim 2\,\text{Gamma}(4)$ et $u_{2k} \sim 3\,\text{Gamma}(6)$, où la loi $\text{Gamma}(a)$ a une densité $f(x) = [\Gamma(a)]^{-1} x^{a-1} e^{-x}$ pour $x > 0$. En deuxième lieu, nous avons créé les valeurs de la variable d'intérêt selon $y_k = 10 + u_{1k} + 3u_{2k} + \varepsilon_k$, $k = 1, 2, ... 5\,000$, avec $\varepsilon_k \sim 5\,\text{Normal}(0)$, où $\text{Normal}(0)$ est la loi normale standard de moyenne 0 et de variance 1. La cible d'estimation dans l'expérience est le total de population de y donné par $Y = \sum_U y_k = 358\,205$. Pour le calage de première phase, nous avons utilisé le vecteur auxiliaire $\mathbf{x}_{1k} = (1, u_{1k})'$ et $\mathbf{z}_{1k} = \mathbf{x}_{1k}$. Autrement dit, les poids w_{1k} pour $k \in s_1$ ont été déterminés par calage sur le total connu $(N, \sum_U u_{1k}) = (5\,000, 39\,611{,}8)$. Pour le calage de deuxième phase, nous avons utilisé $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$ avec $\mathbf{x}_{k(t)} = (1, u_{1k})'$, $\mathbf{x}_{k(w)} = u_{2k}$, $\mathbf{x}_{k(a)} = \phi$ et $\mathbf{z}_k = \mathbf{x}_k$. Ces spécifications satisfont les conditions pour l'équivalence asymptotique de $\hat{Y}_{2Pa}$ et $\hat{Y}_{2Pw}$. Par conséquent, pour cette simulation, nous pouvons travailler avec $\hat{Y}_{2Pa}$ et sa forme linéarisée $\hat{Y}_{2Pa\,\text{lin}}$.

Pour chaque échantillon de première phase s_1, nous avons déterminé les poids finaux w_k pour l'estimateur $\hat{Y}_{2Pa} = \sum_s w_k y_k$ par calage sur les totaux connus donnés par le vecteur $(N, \sum_U u_{1k}, \sum_{s_1} w_{1k} u_{2k}) = (5\,000,\ 39\,611{,}8,\ \sum_{s_1} w_{1k} u_{2k})$. Il importe de noter qu'il n'est pas nécessaire d'avoir $\mathbf{x}_{k(a)} = \phi$ afin d'exécuter une simulation pour comparer $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$ et $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$. Cependant, nous ne pouvons pas comparer $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$ et $\hat{V}(\hat{Y}_{\text{reg}})$ à moins de définir un estimateur $\hat{Y}_{2Pa}$ comparable à $\hat{Y}_{\text{reg}}$ et, pour ce faire, nous avons besoin de $\mathbf{x}_{k(a)} = \phi$, comme il est indiqué à la section 9.

Nous avons tiré des paires d'échantillons répétés (s_1, s), où s_1 est un EAS de n_1 unités provenant de U, et s est un EAS de n unités provenant de s_1. Ici, EAS signifie un échantillonnage aléatoire simple sans remise. Nous avons travaillé avec diverses combinaisons de tailles (n_1, n): (4 000, 3 000), (4 000, 2 000), (4 000, 1 000), (3 000, 2 000), (3 000, 1 000) et (2 000, 1 000). Si $n = n_1$, l'échantillonnage à deux phases est équivalent à l'échantillonnage à une phase, et $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$ et $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$ sont identiques.

Pour chaque combinaison (n_1, n), nous avons réalisé 100 000 paires d'échantillons (s_1, s). En nous basant sur les données pour chacun de ces résultats, nous avons calculé l'estimateur de variance à résidus distincts $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$, l'estimateur de variance à résidu combiné $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$ et l'estimateur de variance $\hat{V}(\hat{Y}_{\text{reg}})$. Pour cela, nous avons utilisé les expressions respectives qui découlent de (7.3), (8.3) et (9.6) quand l'EAS est spécifié à chaque phase. Faute d'espace, les expressions ne sont pas présentées ici. Nous avons obtenu 100 000 valeurs réalisées pour chacun des trois estimateurs de variance. La figure 10.1 montre les distributions pour les 100 000 valeurs de $\hat{V}$ pour $n_1 = 4\,000$ et $n = 2\,000$.

La figure révèle des distributions qui diffèrent de façon frappante pour $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$ et $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$. La distribution de l'estimateur à résidus distincts $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$ est beaucoup plus concentrée. Donc, $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$ est plus efficace que $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$ et, en moyenne, il produit des intervalles de confiance considérablement plus courts. Nous constatons aussi que la distribution de $\hat{V}(\hat{Y}_{\text{reg}})$ est fort semblable à celle de $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$, ce qui appuie l'analyse présentée à la section 9. Nous avons obtenu des résultats semblables pour les autres tailles d'échantillon utilisées dans la simulation.

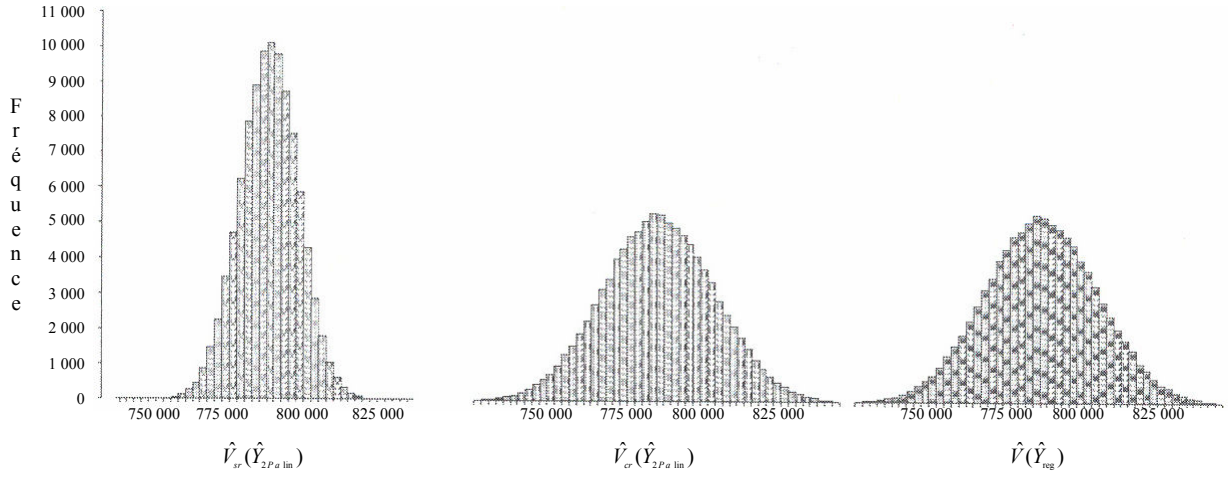


Figure 10.1 Distribution des 100 000 valeurs réalisées pour $\hat{V}_{sr}(\hat{Y}_{2Pa\,\text{lin}})$, $\hat{V}_{cr}(\hat{Y}_{2Pa\,\text{lin}})$ et $\hat{V}(\hat{Y}_{\text{reg}})$

Afin d'obtenir une mesure de l'efficacité des trois estimateurs de variance, nous avons calculé la variance de simulation des 100 000 valeurs de $\hat{V}$. Ces variances de simulation sont présentées aux tableaux 10.1, 10.2 et 10.3. Les chiffres sont spectaculairement plus faibles pour $\hat{V}_{sr}(\hat{Y}_{2Pa\,\mathrm{lin}})$ que pour les deux autres estimateurs. Le tableau 10.4 montre l'avantage relatif de $\hat{V}_{sr}(\hat{Y}_{2Pa\,\mathrm{lin}})$ par rapport à $\hat{V}_{cr}(\hat{Y}_{2Pa\,\mathrm{lin}})$. Pour cette population, la variance de simulation de $\hat{V}_{sr}(\hat{Y}_{2Pa\,\mathrm{lin}})$ est égale à moins de la moitié de la variance de simulation de $\hat{V}_{cr}(\hat{Y}_{2Pa\,\mathrm{lin}})$.

Tableau 10.1
Variance de simulation pour l'estimateur de variance à résidus distincts $\hat{V}_{sr}(\hat{Y}_{2Pa\,\mathrm{lin}})$

		n	
n_1	3 000	2 000	1 000
4 000	64,82	95,91	484,92
3 000		1 179,62	1 806,79
2 000			13 995,94

Nota : Les valeurs réelles sont égales aux valeurs présentées multipliées par 10^6.

Tableau 10.2
Variance de simulation pour l'estimateur de variance à résidu combiné $\hat{V}_{cr}(\hat{Y}_{2Pa\,\mathrm{lin}})$

		n	
n_1	3 000	2 000	1 000
4 000	153,22	364,08	1 290,41
3 000		2 449,05	6 855,69
2 000			33 220,88

Nota : Les valeurs réelles sont égales aux valeurs présentées multipliées par 10^6.

Tableau 10.3
Variance de simulation pour l'estimateur de variance $\hat{V}(\hat{Y}_{\mathrm{reg}})$

		n	
n_1	3 000	2 000	1 000
4 000	153,25	364,14	1 289,79
3 000		2 449,36	6 854,52
2 000			33 210,31

Nota : Les valeurs réelles sont égales aux valeurs présentées multipliées par 10^6.

Tableau 10.4
Ratio des entrées dans le tableau 10.1 aux entrées correspondantes dans le tableau 10.2

		n	
n_1	3 000	2 000	1 000
4 000	0,42	0,26	0,38
3 000		0,48	0,26
2 000			0,42

11. Discussion

Dans une perspective fondée sur le plan de sondage de l'estimation pour les plans d'échantillonnage à deux phases, on peut suivre une approche d'estimation par la régression ou une approche d'estimation par calage. Nous nous concentrons sur l'approche par calage pour créer des estimateurs approximativement sans biais par rapport au plan. La mesure dans laquelle on dispose d'information auxiliaire pour le calage est la clé de l'efficacité des estimations. Nous reconnaissons dans le présent article qu'il existe trois types différents de variables auxiliaires associées aux plans d'échantillonnage à deux phases. Ces trois types ont des caractéristiques d'information différentes. D'après ces caractéristiques, nous définissons quatre vecteurs auxiliaires différents, un pour le calage de première phase et les trois autres pour le calage de deuxième phase. L'approche par calage convient pour analyser les estimateurs résultant de manière systématique. Comme le montre l'article, cette approche donne aussi lieu à un estimateur de variance plus efficace que la méthode classique d'estimation de variance sous des plans d'échantillonnage à deux phases.

Bibliographie

Axelson, M. (1998). Variance estimation for the generalised regression estimator under two-phase sampling - a modified approach. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 85-89.

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

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journeés de Méthodologie*, I.N.S.E.E., Paris.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two phase sampling. *Journal of Official Statistics*, 18, 233-255.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 157-169.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

Hidiroglou, M.A., Rao, J.N.K. et Haziza, D. (2006). Variance estimation in two phase sampling. (Document accepté à paraître dans) *Australian and New Zealand Journal of Statistics*.

Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux degrés ? *Techniques d'enquête*, 23, 89-98.

Särndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.

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

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 17-27
Statistique Canada, N° 12-001-X au catalogue

Matrice chapeau et effets de levier pondérés par les poids de sondage

Jianzhu Li et Richard Valliant [1]

Résumé

Les diagnostics de régression ont pour objectif de détecter des points individuels ou des groupes de points qui exercent une influence importante sur un modèle ajusté. Lorsqu'on ajuste un modèle à l'aide de données d'enquête, les sources d'influence sont la variable réponse **Y**, les variables explicatives **X** et les poids de sondage **W**. Le présent article traite de l'utilisation de la matrice chapeau et des effets de levier pour déceler les points qui pourraient être influents dans l'ajustement des modèles linéaires parce que les valeurs des variables explicatives ou des poids sont grandes. Nous comparons aussi les résultats qu'un analyste pourrait obtenir s'il utilisait les moindres carrés ordinaires plutôt que les moindres carrés pondérés par les poids de sondage pour déterminer quels points sont influents.

Mots clés : Influence ; régression linéaire ; données d'enquête ; moindres carrés pondérés.

1. Introduction

Dans certains diagnostics de régression linéaire classiques, il est souvent utile de mesurer l'influence que chaque point de donnée peut exercer sur la détermination des valeurs des estimations des paramètres et, à leur tour, des valeurs ajustées. La matrice chapeau et ses éléments diagonaux, appelés effets de levier, sont des méthodes fréquemment utilisées pour repérer les cas possédant des valeurs aberrantes pour les variables explicatives et, par conséquent, susceptibles d'être influents dans l'ajustement du modèle s'ils sont également associés à des résidus inhabituels. Quand la régression comporte plus d'une variable explicative, les analystes peuvent calculer les effets de levier pour résumer l'influence collective des valeurs de **X** pour chaque observation.

Dans le cas de l'estimation en population finie, les modèles sont habituellement construits en émettant une hypothèse de superpopulation. Supposons qu'un modèle soit raisonnablement bien ajusté pour la majeure partie de la population. Par souci de commodité, nous l'appellerons le modèle « vrai ». Cependant, l'objectif est habituellement de trouver un modèle possédant une certaine puissance descriptive ou prédictive, en se souvenant qu'aucun modèle n'est réellement « vrai ». Les diagnostics d'influence devraient permettre aux analystes de détecter les points qui font s'écarter les paramètres estimés du modèle vrai. Dans la régression linéaire en utilisant des données d'enquête complexes, les estimations des paramètres sont souvent calculées par la méthode du pseudo-maximum de vraisemblance, décrite par Skinner, Holt et Smith (1989, chapitre 3) en s'inspirant des idées de Binder (1983). Dans le présent article, nous supposons que l'analyste a décidé qu'un estimateur faisant intervenir les poids de sondage convient pour son problème. Comme nous le montrons plus loin, la matrice chapeau et les effets de levier pondérés par les poids de sondage sont utiles pour détecter les observations éventuellement influentes causées non seulement par les valeurs de **X** extrêmes, mais aussi par les grands poids de sondage.

La littérature existante sur les sondages offre des discussions de l'effet des valeurs aberrantes sur certaines estimations d'après des données d'enquête, mais n'accorde que peu d'attention aux diagnostics pour les modèles de régression linéaires. Deville et Särndal (1992) et Potter (1990, 1993) discutent de certains moyens de localiser ou d'élaguer les poids de sondage extrêmes quand l'objectif est d'estimer des totaux de population et d'autres statistiques descriptives simples. Hulliger (1995) et Moreno-Rebollo, Muñoz-Reyes et Muñoz-Pichardo (1999) abordent la question de l'effet des valeurs aberrantes sur l'estimateur d'Horvitz-Thompson d'un total de population. Smith (1987) fait la démonstration de diagnostics basés sur la suppression de cas et une forme de fonction d'influence. Zaslavsky, Schenker et Belin (2001), ainsi que Beaumont et Alavi (2004) utilisent des stratégies fondées sur l'estimation M pour sous-pondérer les grappes ou les unités influentes. Chambers (1986), Gwet et Rivest (1992), Welsh et Ronchetti (1998), ainsi que Duchesne (1999) étudient des techniques d'estimation robustes aux valeurs aberrantes pour les totaux.

L'une des éternelles questions que se posent les analystes des données d'enquête est celle de savoir s'il faut ou non utiliser les poids de sondage dans l'ajustement des modèles. Les recueils publiés sous la direction de Skinner et coll. (1989) et de Chambers et Skinner (2003) traitent de cette question en détail. Binder et Roberts (2003, chapitre 3), Chambers, Dorfman et Sverchkov (2003, sections 11.2.3, 11.6), Chambers et Skinner (2003, chapitre 1), Korn et Graubard (1999, sections 4.3, 4.4), Pfeffermann (1996) et

1. Jianzhu Li, Westat, 1650 Research Boulevard, Rockville MD 20850; Richard Valliant, Survey Research Center, University of Michigan, and Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742.

Smith (1989, chapitre 6) exposent les arguments pour et contre. Les détails peuvent être assez mathématiques et abstraits, mais sont résumés succinctement par Skinner (2003, section 6.2.3).

Nous paraphrasons Skinner (2003, section 6.2.3) ici dans le contexte de l'ajustement d'un modèle linéaire pour prédire une réponse Y basée sur un ensemble de variables explicatives $\mathbf{X}$. Si le modèle linéaire est spécifié correctement et que l'échantillonnage dépend uniquement des variables explicatives comprises dans le modèle, alors les estimations non pondérées des paramètres de régression seront sans biais par rapport au modèle. En particulier, les conditions hypothétiques requièrent que les poids de sondage soient indépendants de Y sachant les valeurs des variables explicatives $\mathbf{X}$. Cependant, si l'échantillonnage dépend de facteurs susceptibles d'être reliés à Y, même après conditionnement sur les valeurs des variables explicatives, les estimateurs non pondérés des paramètres seront entachés d'un biais à la fois par rapport au modèle vrai et au sens de l'échantillonnage répété, fondé sur le plan de sondage. On dit alors que l'on a un plan de sondage *informatif* dans lequel la distribution des valeurs de Y dans l'échantillon diffère de la distribution dans la population. Un exemple de cette situation est donné par Chambers, Dorfman et Sverchkov (2003, section 11.2.3). Si les unités de l'échantillon sont sélectionnées avec des probabilités proportionnelles à une certaine mesure x de leur taille et que Y est reliée à x, la distribution de Y dans l'échantillon sera étalée vers la droite par rapport à sa distribution dans la population. Cette situation est semblable à celle de l'étude empirique que nous décrivons à la section 5.

L'utilisation des poids de sondage protège contre le biais qui pourrait résulter du fait de ne pas tenir compte que l'échantillon est informatif. En outre, si le modèle n'est pas spécifié correctement, la régression pondérée par les poids de sondage estime encore un paramètre de recensement. Autrement dit, les estimations pondérées sont approximativement sans biais pour le modèle linéaire le mieux ajusté qui serait obtenu si l'on disposait de la population finie entière. Dans le présent article, nous supposons qu'un analyste a décidé d'utiliser les poids de sondage dans l'ajustement d'un modèle, éventuellement pour les raisons susmentionnées, et qu'il fournit un type de diagnostic pour évaluer les effets de certains points de données.

La matrice chapeau et les effets de levier que nous présentons sont les mêmes que ceux produits par les progiciels standard quand une régression par les moindres carrés pondérés est exécutée. Cependant, on ne trouve dans la littérature aucune discussion de leur utilisation et de leur interprétation dans le contexte de la régression pondérée par les poids de sondage. Korn et Graubard (1999) est l'une des rares références traitant d'une sorte de diagnostics pour les modèles ajustés d'après des données d'enquête. Les effets de levier font partie d'une série diagnostique et sont plus efficaces s'ils sont évalués avec les résidus. De nombreuses statistiques diagnostiques, telles que la fameuse distance de Cook (Cook 1977), s'avèrent posséder à la fois des effets de levier et des résidus comme composantes.

Les conseils donnés dans la littérature quant à la façon de traiter les observations influentes une fois qu'elles ont été détectées sont quelque peu ambigus. Une solution évidente, et peut-être naïve, consiste à éliminer les valeurs aberrantes et à rajuster le modèle, ce qui est sensé si ces valeurs aberrantes résultent de données incorrectement enregistrées. Une extension naturelle consisterait à élaborer une approche automatique en vertu de laquelle certaines règles seraient appliquées pour détecter les points influents, les supprimer et réajuster le modèle. Dans le présent article, nous présumons qu'après avoir décelé les points influents et examinés prudemment les raisons de leur influence, un analyste déterminera si les points doivent être exclus de l'ajustement, au lieu d'établir une procédure qui exclurait automatiquement les points sur la base de certaines valeurs seuils.

La suite de l'article est présentée comme il suit. À la section 2, nous décrivons la matrice chapeau obtenue par la méthode des moindres carrés ordinaires, les effets de levier et certaines de leurs propriétés. Aux sections 3 et 4, nous traitons de la matrice chapeau et des effets de levier pondérés par les poids de sondage, ainsi que d'une décomposition qui montre comment les points peuvent posséder un effet de levier important. Les extensions aux données d'enquête s'appliquent aux plans de sondage à un degré ainsi qu'à plusieurs degrés. À la section 5, nous donnons un exemple numérique s'appuyant sur un échantillon à un degré d'organismes spécialisés dans les soins de santé mentale. À la dernière section, nous résumons nos résultats et donnons certaines orientations que pourraient prendre des travaux de recherche supplémentaires.

2. Matrice chapeau fondée sur les MCO

Un modèle *de travail* est un modèle qui est considéré provisoirement par un analyste comme étant la structure qui décrit le mieux une superpopulation conceptuelle. Il peut être révisé après une évaluation plus approfondie par ajout de variables explicatives, suppression de variables explicatives ou d'autres changements ayant trait à sa forme. Supposons que le modèle linéaire de travail est

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} \tag{1}$$

où $\mathbf{Y} = (Y_1, ..., Y_n)^T$, $\mathbf{X}^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$ avec $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, et $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$. En supposant que la matrice $\mathbf{X}$ est de plein rang, l'estimation par le moindres carrés ordinaires (MCO) de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{A}^{-1} \mathbf{X}^T \mathbf{Y}, \qquad (2)$$

où $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ est une matrice carrée et inversible. Les valeurs ajustées $\hat{\mathbf{Y}}$ correspondant aux valeurs observées $\mathbf{Y}$ sont

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

où $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T$ est appelé la matrice chapeau. Ce nom a été utilisé pour la première fois par Tukey (Belsley, Kuh et Welsch 1980, chapitre 2 ; Hoaglin et Welsch 1978). L'effet de levier, $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i$, est le i^{e} élément sur la diagonale de cette matrice chapeau, qui mesure l'effet de Y_i sur sa propre valeur ajustée, puisque $\hat{Y}_i = \sum_j h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$. Si h_{ii} s'approche de 1, Y_i joue un rôle crucial dans la détermination de la valeur de $\hat{Y}_i$.

La matrice chapeau et les effets de levier basés sur les MCO possèdent de nombreuses propriétés spéciales et utiles :

i) $\mathbf{H}$ est symétrique, ou $h_{ij} = h_{ji}$;
ii) $\mathbf{H}$ est idempotente, ou $\mathbf{H} = \mathbf{H}^2$, ou $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$;
iii) $\mathbf{H}\mathbf{X} = \mathbf{X}$ ou $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$;
iv) $0 \leq h_{ii} \leq 1$;
v) $\sum_i h_{ii} = \text{rang}(\mathbf{X}) = p$, ce qui implique que l'effet de levier moyen est $\bar{h} = p/n$;

si le modèle (1) contient une constante, les deux propriétés suivantes sont vérifiées :

vi) $\sum_i h_{ij} = 1$;
vii) $h_{ii} = 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{A}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$, où $\bar{\mathbf{x}} = \sum_n \mathbf{x}_i / n$.

Dans un ensemble de données raisonnablement grand, une valeur d'effet de levier individuel h_{ii} est habituellement considérée comme extrême si elle est au moins égale à deux fois la moyenne, $\bar{h} = p/n$ (Belsley et coll., 1980, chapitre 2). L'existence d'un écart important entre la plupart des cas et quelques cas inhabituels dans la distribution empirique des effets de levier donne aussi la preuve de l'existence d'unités aberrantes.

3. Matrice chapeau pondérée par les poids de sondage

Dans l'approche du pseudo-maximum de vraisemblance, la première étape consiste à former le système d'équations d'estimation qui conviendrait pour un modèle si la population finie entière était observée. Ce système est un type de total de population qui est alors estimé en utilisant des méthodes fondées sur le plan de sondage. Supposons que le modèle structurel sous-jacent est un modèle linéaire à effets fixes :

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind}\, N(0,\ v_i \sigma^2) \qquad (3)$$

où ε_i suit indépendamment une loi normale de moyenne 0 et de variance $v_i \sigma^2$, qui est connue à part la constante σ^2. L'estimateur du pseudo-maximum de vraisemblance (EPMV) de $\boldsymbol{\beta}$ est la solution du système d'équations d'estimation $\mathbf{X}^T \mathbf{W}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$, avec $\mathbf{V} = \text{diag}(v_1, ..., v_n)$ et $\mathbf{W} = \text{diag}(w_1, ..., w_n)$. Les poids de sondage, qui, dans les échantillons probabilistes, sont habituellement inversement proportionnels aux probabilités d'inclusion, sont utilisés dans l'EPMV pour tenir compte d'un plan informatif dans lequel la distribution de Y dans l'échantillon diffère vraisemblablement de celle dans la population finie. Ces équations peuvent être résolues explicitement comme $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W}\mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}\mathbf{V}^{-1} \mathbf{Y}$. Si nous supposons que $\mathbf{V} = \mathbf{I}$, le modèle (3) se réduit à (1) et l'estimateur pondéré par les poids de sondage (PPS) $\hat{\boldsymbol{\beta}}$ prendra par conséquent la forme d'un estimateur par les moindres carrés pondérés, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}\mathbf{Y}$.

S'il est tenu compte des poids de sondage dans la régression, les valeurs prédites deviennent $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, où la matrice chapeau inclut les poids de sondage et est définie par

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} = \mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T \mathbf{W}$$

avec $\mathbf{A} = \mathbf{X}^T \mathbf{W}\mathbf{X}$. Les effets de levier pour la diagonale de la matrice chapeau sont $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i w_i$. Dans cette formulation, il est supposé que l'analyste n'intègre pas une matrice $\mathbf{V}$ dans la régression. Cependant, il est possible de modifier les résultats afin d'intégrer $\mathbf{V}$ en utilisant simplement $\mathbf{W}^* = \mathbf{W}\mathbf{V}^{-1}$ au lieu de $\mathbf{W}$. Contrairement à la matrice chapeau non pondérée, la matrice chapeau PPS n'est plus symétrique pour les plans d'échantillonnage avec probabilités de sélection inégales (ou, de manière plus générale, poids inégaux). Les propriétés (ii) à (vi) énoncées à la section 2 restent vérifiées (par exemple, voir Valliant, Dorfman et Royall 2000, chapitre 5) à condition que les matrices chapeau non pondérées soient remplacées par leur version pondérée. En outre, la matrice chapeau PPS possède les propriétés supplémentaires utiles, et faciles à vérifier, suivantes :

a) $\mathbf{W}\mathbf{H} = \mathbf{W}\mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} = \mathbf{H}^T \mathbf{W}$;
b) $\mathbf{X}^T \mathbf{W}(\mathbf{I} - \mathbf{H}) = \mathbf{X}^T \mathbf{W} - \mathbf{X}^T \mathbf{H}^T \mathbf{W} = \mathbf{0}$;
c) $w_{i'} h_{i'i} = w_{i'} \mathbf{x}_{i'}^T \mathbf{A}^{-1} \mathbf{x}_i w_i = w_i h_{ii'}$.

La définition des effets de levier pondérés indique qu'un grand effet de levier peut être causé par des valeurs de $\mathbf{X}$ aberrantes, un poids aberrant, ou les deux. Notons que les formules de la matrice chapeau et des effets de levier pondérés par les poids de sondage s'appliquent que le plan de sondage soit stratifié, à un degré ou à plusieurs degrés,

contrairement aux diagnostics, telle la distance de Cook, qui requièrent que les erreurs-types estimées ou les matrices de covariance soient spécialisées de manière à être ajustées au plan d'échantillonnage.

4. Décomposition des effets de levier

Les effets de levier peuvent être décomposés en éléments qui séparent les effets du poids et des valeurs de $\mathbf{X}$ pour une unité. Supposons que le modèle de travail est (1) et qu'il contient une constante, de sorte que

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^T \end{pmatrix} \equiv (\mathbf{1}\ \mathbf{X}_1) \text{ et } \mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

où $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{i,\,p-1})$ représente des vecteurs de dimension $1 \times (p-1)$, $\mathbf{1}$ est un vecteur de dimension $n \times 1$ dont tous les éléments sont égaux à 1, et $\mathbf{X}_1$ est une matrice de dimensions $n \times (p-1)$. La matrice $\mathbf{A}$ est calculée sous la forme

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{X}_1^T \end{pmatrix} \mathbf{W} (\mathbf{1}\ \mathbf{X}_1) = \begin{pmatrix} \mathbf{1}^T\mathbf{W}\mathbf{1} & \mathbf{1}^T\mathbf{W}\mathbf{X}_1 \\ \mathbf{X}_1^T\mathbf{W}\mathbf{1} & \mathbf{X}_1^T\mathbf{W}\mathbf{X}_1 \end{pmatrix} \equiv \begin{pmatrix} \hat{N} & \hat{\mathbf{t}}_X^T \\ \hat{\mathbf{t}}_X & \mathbf{A}_1 \end{pmatrix},$$

où $\hat{\mathbf{t}}_X$ est un vecteur de dimension $(p-1) \times 1$ dont les éléments sont $\hat{\mathbf{t}}_{Xj} = \sum_{i \in s} w_i x_{ij}$ et $\mathbf{A}_1$ est une matrice de dimensions $(p-1) \times (p-1)$. En utilisant l'inverse d'une matrice partitionnée,

$$\begin{aligned} \mathbf{A}^{-1} &= \begin{pmatrix} \frac{1}{\hat{N}} + \frac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \hat{\mathbf{t}}_X \frac{1}{\hat{N}} & -\frac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \\ -\frac{1}{\hat{N}} \mathbf{S}^{-1} \hat{\mathbf{t}}_X & \mathbf{S}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\hat{N}} + \bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \bar{\mathbf{x}}_W & -\bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \\ -\mathbf{S}^{-1} \bar{\mathbf{x}}_W & \mathbf{S}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\hat{N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\bar{\mathbf{x}}_W^T \\ \mathbf{I} \end{pmatrix} \mathbf{S}^{-1} (-\bar{\mathbf{x}}_W\ \mathbf{I}) \end{aligned}$$

où $\bar{\mathbf{x}}_W = \hat{\mathbf{t}}_X / \hat{N}$ est un vecteur de dimension $(p-1) \times 1$, et $\mathbf{S} = \mathbf{A}_1 - \hat{\mathbf{t}}_X \hat{\mathbf{t}}_X^T / \hat{N}$ est une matrice de dimensions $(p-1) \times (p-1)$. En simplifiant la matrice chapeau en nous servant de la matrice inverse susmentionnée, nous obtenons

$$\begin{aligned} \mathbf{H} &= \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W} \\ &= \left\{ \frac{1}{\hat{N}} \mathbf{1}\mathbf{1}^T + (\mathbf{X}_1 - \mathbf{1}\bar{\mathbf{x}}_W^T)\, \mathbf{S}^{-1} (-\bar{\mathbf{x}}_W \mathbf{1}^T + \mathbf{X}_1^T) \right\} \mathbf{W} \\ &= \left\{ \frac{1}{\hat{N}} \mathbf{1}\mathbf{1}^T + \begin{pmatrix} \mathbf{x}_1^T - \bar{\mathbf{x}}_W^T \\ \vdots \\ \mathbf{x}_n^T - \bar{\mathbf{x}}_W^T \end{pmatrix} \mathbf{S}^{-1} (\mathbf{x}_1 - \bar{\mathbf{x}}_W, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}_W) \right\} \mathbf{W}. \end{aligned}$$

Puis, en utilisant le fait que $\hat{N} = n\bar{w}$ avec $\bar{w} = \sum_{i=1}^n w_i / n$, l'effet de levier de la i^{e} observation, ou le i^{e} élément diagonal de la matrice chapeau pondérée $\mathbf{H}$, est

$$h_{ii} = \frac{1}{n} \frac{w_i}{\bar{w}} [1 + \hat{N} (\mathbf{x}_i - \bar{\mathbf{x}}_W)^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_W)].$$

La forme quadratique, $(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\, \mathbf{S}^{-1}\, (\mathbf{x}_i - \bar{\mathbf{x}}_W)$, définit un ellipsoïde centré à $\bar{\mathbf{x}}_W$ (par exemple, voir Weisberg 2005, chapitre 8), et $\hat{N}\, (\mathbf{x}_i - \bar{\mathbf{x}}_W)^T \mathbf{S}^{-1}\, (\mathbf{x}_i - \bar{\mathbf{x}}_W)$ est la distance de Mahalanobis de $\mathbf{x}_i$ à $\bar{\mathbf{x}}_W$. Par conséquent, un effet de levier peut être grand si 1) w_i est grand, surtout comparativement au poids moyen $\bar{w}$ ou que 2) $\mathbf{x}_i$ est très éloigné de la moyenne pondérée, $\bar{\mathbf{x}}_W$, des $\mathbf{X}$, dans la métrique déterminée par la matrice $\mathbf{S}$.

Par exemple, dans un modèle linéaire simple ne contenant qu'une seule variable auxiliaire, $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2)$, l'effet de levier de la i^{e} observation est

$$h_{ii}^W = \frac{1}{n} \frac{w_i}{\bar{w}} \left[1 + \hat{N} \frac{(x_i - \bar{x}_W)^2}{\sum_{j=1}^n w_j (x_j - \bar{x}_W)^2} \right].$$

où $\bar{x}_W = \sum_i w_i x_i / \hat{N}$.

Si les termes d'erreur du modèle ont une structure de variance générale $\boldsymbol{\varepsilon} \sim (0, \mathbf{V})$ et que $\mathbf{V}$ est connu, la matrice chapeau est alors définie par $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}$ avec

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} \mathbf{1}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{1} & \mathbf{1}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1 \\ \mathbf{X}_1^T\mathbf{V}^{-1}\mathbf{W}\mathbf{1} & \mathbf{X}_1^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1 \end{pmatrix} \\ &= \begin{pmatrix} \sum_s w_i / v_i & \sum_s w_i \mathbf{x}_i^T / v_i \\ \sum_s w_i \mathbf{x}_i / v_i & \sum_s w_i \mathbf{x}_i \mathbf{x}_i^T / v_i \end{pmatrix}. \end{aligned}$$

Une formule pour $\mathbf{A}^{-1}$ semblable à celle qui précède s'applique avec $\hat{\mathbf{t}}_{XV} = \sum_s w_i \mathbf{x}_i / v_i$, $\hat{N}_V = \sum_s w_i / v_i$ et $\mathbf{S}_V = \mathbf{X}_1^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1 - \hat{\mathbf{t}}_{XV}\hat{\mathbf{t}}_{XV}^T / \hat{N}_V$. Si nous utilisons une $\mathbf{V}$ générale, $\hat{\mathbf{t}}_{XV}$ et $\hat{N}_V$ ne sont plus des estimations fondées sur le plan de sondage de $\mathbf{T}_X$ et N, mais sont des estimations de $\mathbf{T}_{XV} = \sum_1^N \mathbf{x}_i / v_i$ et $N_V = \sum_1^N 1/v_i$. L'effet de levier de la i^{e} observation sous ce modèle général est

$$h_{ii} = \frac{w_i}{v_i \hat{N}_V} [1 + \hat{N}_V\, (\mathbf{x}_i - \bar{\mathbf{x}}_{WV})^T\, \mathbf{S}_V^{-1}\, (\mathbf{x}_i - \bar{\mathbf{x}}_{WV})].$$

5. Exemple numérique

Comme nous l'avons mentionné à la section 1, des arguments peuvent être avancés pour justifier le fait d'ignorer les caractéristiques du plan de sondage en général et la pondération en particulier dans l'ajustement des modèles. Grosso modo, quand un modèle est conditionné

sur toutes les variables du plan de sondage déterminant le schéma d'échantillonnage et qu'il est correct pour la population ainsi que l'échantillon, la régression par les MCO peut être utilisée. Les analystes peuvent formuler des objections à l'intégration des variables du plan de sondage dans un modèle, parce que certaines ne sont pas scientifiquement intéressantes en tant que variables explicatives. En outre, le conditionnement sur toutes les variables du plan de sondage est parfois impossible, surtout quand le « schéma d'échantillonnage » inclut une non-réponse non contrôlée qui, elle-même, peut être reliée à la variable réponse. Comme nous le mentionnons dans la section 1, la pondération par les poids de sondage offre un minimum de protection contre la spécification incorrecte du modèle quand la distribution des Y dans l'échantillon n'est pas la même que dans la population à cause du type de plan d'échantillonnage utilisé. Néanmoins, certains analystes soutiendront que le plan d'échantillonnage et les poids de sondage peuvent être ignorés dans des applications particulières et que les MCO sont appropriés. Donc, il est intéressant de voir dans quelle mesure les diagnostics pour les MCO et pour la PPS diffèrent dans une application réelle. Cependant, étant donné un mode d'action, un analyste devrait utiliser des diagnostics en harmonie avec la méthode d'ajustement. Si les MCO sont utilisés, les diagnostics standard pour les MCO doivent être examinés ; si la régression PPS est utilisée, les diagnostics pour la pondération par les poids de sondage sont ceux qu'il convient d'appliquer. Il se pourrait fort bien que différents points soient influents selon que l'on utilise la régression par les MCO ou la régression PPS.

À la présente section, nous examinons la matrice chapeau et les effets de levier dans un exemple de régression. Nous utilisons la Survey of Mental Health Organizations (SMHO) réalisée aux États-Unis en 1998, durant laquelle des données ont été recueillies sur les organismes spécialisés dans les soins de santé mentale et sur les services de soins de santé mentale des hôpitaux généraux. L'échantillon de cette enquête a été tiré selon un plan stratifié à un seul degré avec probabilité proportionnelle à la taille (PPT) (Manderscheid et Henderson 2002 ; Choudhry 2000). La mesure de taille (MT) utilisée dans l'échantillonnage était le nombre d'« épisodes », défini comme étant le nombre de patients/clients d'un organisme au début de 1998 plus le nombre de nouveaux patients/clients ajoutés durant l'année civile 1998. Nombre de variables analysées dans l'enquête sont reliées à la MT et leur distribution non pondérée dans l'échantillon sera différente de la distribution dans la population, puisque l'échantillon a tendance à contenir des unités de plus grande taille. Donc, ce plan est éventuellement informatif, tel qu'il est défini dans Chambers et Skinner (2003).

Les tailles variées des organismes de soins de santé mentale ont produit de grandes fourchettes de valeurs pour les variables étudiées dans l'échantillon, si bien que certaines observations pourraient avoir une influence relativement grande sur les estimations des paramètres d'une régression linéaire. Dans la présente étude, le modèle d'intérêt est la régression des dépenses totales d'un organisme de santé, en millier de dollars, sur le nombre de lits installés et dotés en personnel pour l'utilisation et le nombre de patients ou clients ajoutés durant l'année de déclaration. Nous avons utilisés l'estimateur PPS, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$. Imitant la procédure employée par la plupart des analystes, nous n'avons pas intégré de matrice de variance $\mathbf{V}$ du modèle dans l'estimation des paramètres de la régression. Au total, nous avons utilisé 875 observations dans la régression, pour chacune desquelles aucune valeur ne manquait pour les variables indépendantes et dépendante.

Le tableau 1 résume les valeurs des quantiles des variables incluses dans la régression, y compris les poids de sondage. Le total des dépenses possède un maximum de 519 863,3, qui vaut près de 30 000 fois le minimum, 16,6. Bien qu'ils ne soient pas aussi extrêmes que le total des dépenses, le nombre de lits et le nombre de patients ajoutés présentent aussi un écart significatif entre leurs valeurs maximale et minimale. Comme l'échantillon a été sélectionné selon un plan PPT, les poids de sondage étaient associés aux tailles des organismes de santé mentale, la fourchette allant de 1 à 158,86. Les poids que nous utilisons dans l'analyse comprennent une correction de la non-réponse qui a été effectuée séparément par strate du plan de sondage. Dans certains cas, les unités qui avaient été sélectionnées avec certitude dans l'échantillon initial n'ont pas répondu et certaines unités sélectionnées avec certitude répondantes avaient un poids corrigé supérieur à 1. En tout, 157 organismes avaient un poids de 1 après la correction de la non-réponse.

Tableau 1
Quantiles des variables dans la régression pour la SMHO

	Quantiles				
Variables	**0 %**	**25 %**	**50 %**	**75 %**	**100 %**
Dépenses (milliers de dollars)	16,6	2 932,5	6 240,5	11 842,6	519 863,3
N^{bre} de lits	0	6,5	36	93	2 405
N^{bre} d'ajouts	0	558,5	1 410	2 406	79 808
Poids	1	1,42	2,48	7,76	158,86

Dans la régression qui suit, nous avons inclus les unités ayant un poids de 1 dans l'estimation de l'erreur-type. Au lieu de les exclure, conformément à l'approche suivie pour traiter les unités sélectionnées avec certitude dans une estimation purement fondée sur le plan de sondage. Inclure les unités sélectionnées avec certitude concorde avec l'idée qu'un modèle de superpopulation est estimé et que les

coefficients de pente auraient encore une variance, même si l'on procédait à un recensement. L'esquisse de la justification mathématique de cette façon de faire, qui dépend du modèle (et n'est pas fondée sur le plan) est donnée en annexe

La figure 1 montre les diagrammes de dispersion des dépenses en fonction des nombres de lits et d'ajouts pour l'échantillon de 875 établissements (en omettant un très grand établissement décrit plus loin). Dans la première rangée, les points dont l'effet de levier sous les MCO est supérieur à $2p/n = 0{,}007$ sont mis en relief. La deuxième rangée donne les graphiques à bulles dans lesquels les bulles sont proportionnelles au poids de chaque cas. Les points à effet de levier élevé sous la pondération par les poids de sondage (PPS) sont mis en relief en utilisant le même seuil de 0,007. Les distributions des variables explicatives sont assez asymétriques comme l'indique le tableau 1. Il existe également un très grand établissement qui n'est pas représenté dans la figure 1 parce qu'il fausse l'échelle du graphique. Pour cet établissement (désigné par observation 818 ici), (dépenses en milliers de dollars ; lits ; ajouts) = (519 863,3 $; 2 405 ; 79 808) et le poids de sondage est égal à 2,22. (L'observation 818 est l'un des cas mentionnés plus haut dont la sélection dans l'échantillon initial était certaine, mais qui ont fait l'objet d'une correction de la non-réponse et qui ont par conséquent un poids final supérieur à 1.) Parce que ses valeurs données sortent fortement de l'alignement de celles obtenues pour les autres organismes, ce point pourrait affecter les estimations.

Le tableau 2 donne les 20 observations pour lesquelles les effets de levier sous PPS sont les plus grands. Les valeurs de ces effets de levier varient de 0,022 à 0,389, chiffres considérablement plus élevés que le niveau de 0,007 établi empiriquement. Ce tableau montre aussi, pour ces 20 cas, les effets de levier non pondérés sous les MCO, le ratio du poids de sondage individuel au poids de sondage moyen, ainsi que la distance absolue relative entre les valeurs de X individuelles et leurs moyennes pondérées. Nous notons que l'unité 818 possède les effets de levier pondérés et non pondérés les plus grands, principalement à cause de ses nombres très élevés de lits et de patients ajoutés. Puisque ce cas possède un poids d'échantillonnage inférieur à la moyenne, l'effet de levier sous les MCO est encore plus grand que l'effet pondéré. Il existe d'autres cas semblables, dont les unités 271, 179, 820, 157, 163, 156 et 154, qui sont associés à un nombre extrême de lits, ou un nombre extrême de patients ajoutés, ou aux deux, mais dont les poids sont faibles. Un autre type de valeurs aberrantes découle de poids de sondage extrêmes, même si les valeurs de leurs variables auxiliaires diffèrent peu les unes des autres. Les unités 672, 613, 711, 801 et 611 possèdent toutes un poids de sondage valant plus de 15 fois le poids moyen. Leurs effets de levier pondérés s'avèrent être grands, tandis que les effets de levier non pondérés ne le sont pas. Il existe également un écart appréciable entre les effets de levier pondérés pour le cas 331 ($h_{ii} = 0{,}075$) et pour le cas 271 ($h_{ii} = 0{,}046$).

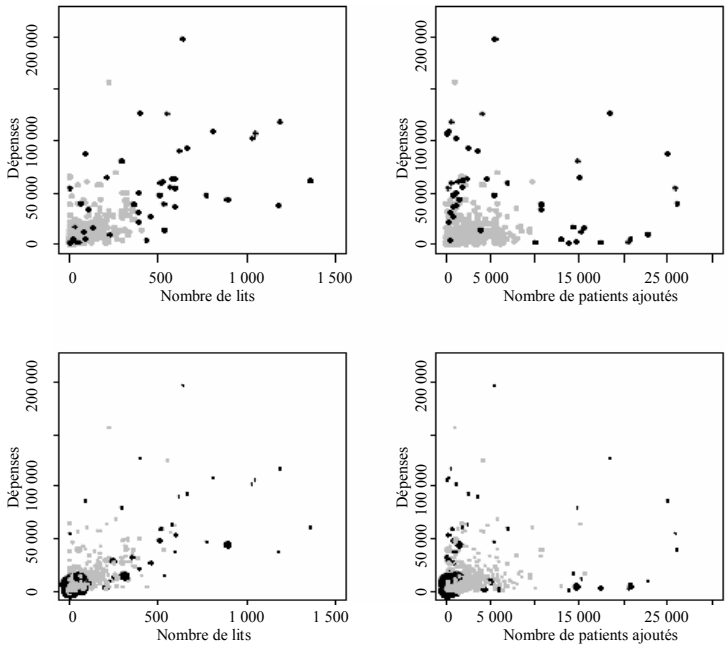


Figure 1 Diagrammes de dispersion des dépenses en fonction du nombre de lits et du nombre de patients ajoutés. Les points dont l'effet de levier est élevé selon les MCO (la PPS) sont mis en relief dans la rangée supérieure (inférieure)

Tableau 2
Observations ayant les 20 effets de levier pondérés les plus importants

			Poids	Lits	Patients ajoutés
ID de l'obs.	h_{ii} **sous MCO**	h_{ii} **pondéré**	$w_i / \bar{w}$	$\lvert x_{1i} - \bar{x}_1^W \rvert / \bar{x}_1^W$	$\lvert x_{2i} - \bar{x}_2^W \rvert / \bar{x}_2^W$
818	0,513	0,389	0,3	49,3	64,7
189	0,037	0,245	3,4	17,7	0,3
346	0,035	0,157	2,2	0,6	16,1
366	0,017	0,105	3,0	0,7	11,1
331	0,024	0,075	1,5	0,1	13,4
271	0,068	0,046	0,4	23,7	0,0
830	0,004	0,045	5,8	5,4	0,1
628	0,056	0,045	0,4	1,0	20,3
179	0,089	0,038	0,2	27,4	0,5
672	0,002	0,034	24,2	1,0	0,8
820	0,048	0,034	0,3	0,8	19,6
207	0,012	0,030	1,3	9,5	0,3
157	0,069	0,030	0,2	23,8	0,5
163	0,017	0,027	0,8	11,4	0,8
613	0,002	0,026	18,5	1,0	0,7
711	0,002	0,024	16,8	1,0	0,9
801	0,002	0,024	17,5	0,6	0,9
156	0,055	0,023	0,2	20,9	0,9
611	0,002	0,023	15,9	1,0	0,8
154	0,051	0,022	0,2	20,5	0,1
			$\bar{w} = 6,57$	$\bar{x}_1^W = 47,83$	$\bar{x}_2^W = 1\,214,13$

Nota: l'ID de l'observation est le numéro de ligne de cette observation dans l'échantillon.

Les tailles des poids de sondage peuvent mener les analystes à tirer des conclusions différentes selon qu'ils utilisent les effets de levier pondérés ou non pondérés pour repérer les observations éventuellement influentes. La figure 2 donne un diagramme de dispersion des effets de levier pondérés en fonction de leur version non pondérée. Les deux droites de référence ont été tracées aux valeurs de 0,007. L'observation 818 est omise, parce qu'elle fausserait l'échelle du graphique. Manifestement, les points à effet de levier élevé détectés par la méthode PPS uniquement, situés dans l'aire A, ont tous un poids significativement plus élevés que ceux compris dans l'aire B, qui sont détectés par la méthode de MCO uniquement.

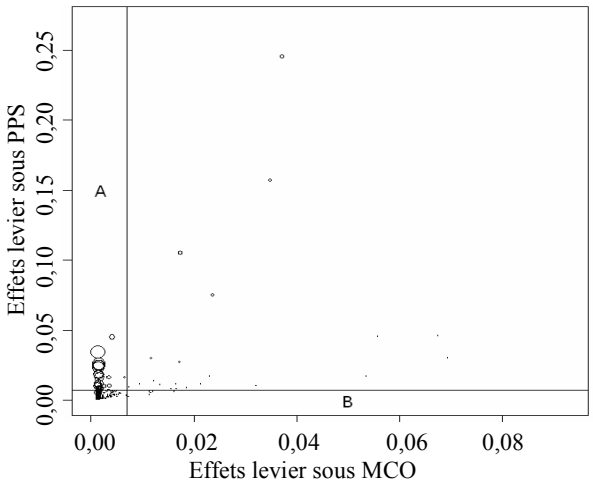


Figure 2 Tracé des effets de levier pondérés par les poids de sondage en fonction des effets de levier sous les MCO non pondérés

Étant donné que certains cas éventuellement influents ont été décelés, l'étape suivante consiste à voir quel effet ils ont sur les estimations des paramètres. Le tableau 3 donne les estimations des paramètres fondées sur les MCO et sur la PPS en utilisant tous les cas. Le tableau 4 donne les estimations fondées sur les MCO et sur la PPS i) en omettant les cas à effet de levier élevé et ii) en omettant l'observation 818. Les points à effet de levier élevé sont ceux pour lesquels $h_{ii} > 0{,}007$. Toutefois, il convient de souligner que les ensembles de points ayant un effet de levier élevé ne sont pas les mêmes dans les régressions par les MCO et par PPS. Les erreurs-types sont estimées au moyen de la formule des MCO habituelle et au moyen de l'estimateur sandwich (Binder 1983) pour les estimations PPS.

Si nous comparons les tableaux 3 et 4, nous constatons que les estimations par les MCO changent considérablement après l'élimination des points à effet de levier élevé (section (i) du tableau 4). L'ordonnée à l'origine de la droite des MCO, qui est significative dans les deux tableaux, passe d'une valeur négative à une valeur positive. La pente de la droite des MCO pour le nombre de lits diminue d'environ 26 % (de 94,16 à 69,27) quand les points à effet de levier élevé sont éliminés. Dans le cas de la pente pour le nombre de patients ajoutés, la diminution est d'environ 59 %. Les estimations PPS pour les nombres de lits et de patients ajoutés sont également sensibles à la présence de points à effet de levier élevé, les pentes diminuant de 7 % et de 46 %, respectivement. Dans tous les cas, les pentes sont significatives, de sorte que la conclusion qualitative selon laquelle les dépenses sont corrélées au nombre de lits et au nombre de patients ajoutés est vérifiée avec ou sans les points à effet de levier élevé. Cependant, les valeurs prédites avant et après omission de ces points diffèrent assez bien.

Les erreurs-types (E.-T.) diminuent aussi considérablement quand les points à effet de levier élevé sont omis. Par exemple, l'erreur-type sous PPS pour le nombre de lits passe de 13,14 à 6,75 (une réduction de 49 %) ; pour le nombre de patients ajoutés, elle passe de 0,76 à 0,21 (une réduction de 72 %). Ce résultat est dû au fait que certains points ayant un poids extrême sont éliminés dans la régression sous PPS. Par contre, l'erreur-type des estimations par les MCO augmente effectivement quand les points à effet de levier élevé sous les MCO sont omis, parce que la variance d'échantillon des x diminue. Il s'agit d'un autre exemple des écarts importants qui peuvent survenir quand le même type de diagnostic est appliqué aux régressions par les MCO et sous PPS.

Vu que le point 818 est si manifestement extrême, nous avons également ajusté la régression après avoir éliminé cette observation uniquement. Les résultats sont présentés à la section (ii) du tableau 4. Omettre ce point unique cause des changements appréciables dans les estimations des paramètres tant par les MCO que par PPS. Cela illustre aussi qu'un seul point peut avoir une incidence sur les erreurs-types des pentes estimées dans une régression pondérée par les poids de sondage, comme c'est également le cas dans une régression par les MCO. L'observation 818 possède un grand résidu (voir figure 3) ; son omission fait baisser l'erreur-type pour le nombre de lits, de 13,14 dans le tableau 3 à 8,04 dans le tableau 4. Notons que, si l'unité 818 avait un grand poids, son résidu serait vraisemblablement plus petit, puisqu'il aurait plus d'effet sur l'ajustement. Le cas échéant, l'erreur-type pourrait actuellement être plus faible que quand l'unité 818 est incluse.

Tableau 3
Estimation par MCO et PPS des paramètres de la régression pour la SMHO en utilisant la totalité des 875 cas échantillonnés

Variables	Estimation par les MCO			Estimation par PPS		
indépendantes	Coefficient	E.-T.	*t*	Coefficient	E.-T.	*t*
Ordonnée à l'origine	-1 201,73	526,19	-2,28	514,08	1 157,71	0,44
N^{bre} de lits	94,16	3,03	31,08	81,23	13,14	6,18
N^{bre} de patients ajoutés	2,31	0,13	18,50	1,84	0,76	2,43

Tableau 4
Estimation par MCO et par PPS des paramètres de la régression pour la SMHO

Variables	Estimation par les MCO			Estimation par PPS		
indépendantes	Coefficient	E.-T.	*t*	Coefficient	E.-T.	*t*
i) Suppression des observations dont l'effet de levier est supérieur à 0,007						
Ordonnées à l'origine	2 987,55	490,54	6,09	1 993,86	353,71	5,64
N^{bre} de lits	69,27	4,35	15,94	75,82	6,75	11,23
N^{bre} de patients ajoutés	0,95	0,20	4,71	1,00	0,21	4,73
ii) Suppression de l'observation 818						
Ordonnées à l'origine	1 979,51	537,93	3,68	2 281,17	460,35	4,96
N^{bre} de lits	81,80	2,92	27,98	68,69	8,04	8,54
N^{bre} de patients ajoutés	1,19	0,14	8,41	0,79	0,29	2,75

Un autre point qui se dégage des tableaux 3 et 4 est que les estimations par les MCO et par PPS sont nettement plus proches les unes des autres après l'élimination des points à effet de levier élevé qu'avant. Comme l'illustre le tableau 5, les estimations par les MCO sont supérieures de 16 % et de 26 % aux estimations par PPS quand tous les points sont retenus, mais sont inférieures de 9 % et de 5 % à ces estimations après avoir laissé tomber certains points.

Tableau 5
Ratios des estimations des paramètres par MCO et par PPS avant et après l'élimination des observations dont l'effet de levier est supérieur à 0,007 dans la régression pour la SMHO

	Ratio des estimations par MCO aux estimations par PPS	
	Avec tous les points	Avec élimination des points à effet de levier élevé
Lits	1,16	0,91
Patients ajoutés	1,26	0,95

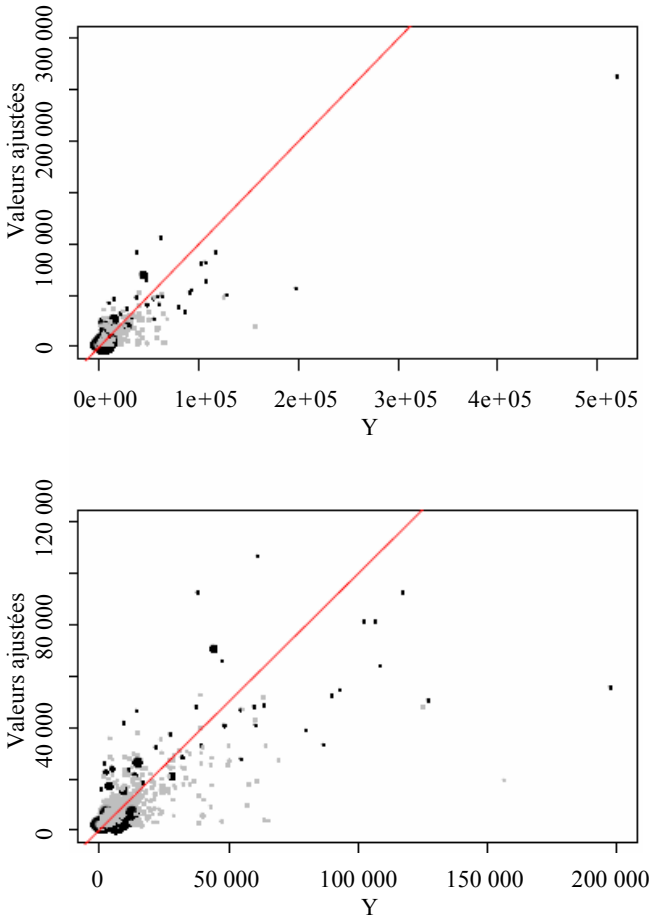


Figure 3 Tracé des valeurs ajustées en fonction des valeurs de Y. La droite de référence est tracée à $Y = \hat{Y}$. Le graphique supérieur comprend tous les points. Dans le graphique inférieur, l'observation extrême 818 est omise. Les points à effet de levier élevé basés sur la PPS sont représentés par des cercles pleins, foncés, dans chaque graphique

Les effets de levier sont habituellement combinés aux résidus pour déterminer quels points sont influents dans l'ajustement du modèle de régression, parce que les résidus peuvent être utilisés pour détecter les valeurs de Y divergentes. Un diagramme de dispersion des valeurs ajustées provenant de la régression sous PPS en fonction des valeurs de Y est représenté à la figure 3. Les points à effet de levier élevé sont présentés par des cercles pleins foncés. Les distances verticales des points jusqu'à la droite à 45 degrés correspondent à la taille des résidus. Le graphique supérieur comprend chacun des 875 points de l'échantillon ; dans le graphique inférieur, l'observation 818 est omise pour améliorer la résolution pour les autres points. Notons que certaines observations ont un effet de levier élevé et un petit résidu, tandis que d'autres ont un effet de levier faible et un grand résidu. L'influence de ces points sur la régression peut être étudiée plus en détail en utilisant divers outils que nous ne décrirons pas ici. Par exemple, la distance de Cook, qui comporte implicitement l'effet de levier et le résidu, est conçue pour mesurer l'effet de la suppression d'une seule observation sur les estimations globales des paramètres. L'adaptation de certaines statistiques diagnostiques élémentaires des MCO aux données d'enquête, DFBETAS et DFFITS, a été discutée sous un plan d'échantillonnage à un seul degré dans Li et Valliant (2006).

6. Conclusion

Les effets de levier et les résidus sont des composantes essentielles des statistiques diagnostiques destinées à détecter l'influence importante d'une seule observation ou d'un groupe d'observations sur un modèle linéaire ajusté. Les ensembles de données d'enquête peuvent contenir des observations influentes, que l'on soutienne que le plan d'échantillonnage est ignorable et que les moindres carrés ordinaires peuvent être utilisés, ou que le plan doive être pris en compte et les poids de sondage utilisés. Les points qui sont influents dans les deux cas ne sont pas nécessairement les mêmes, comme nous l'avons illustré ici.

Une fois que les points dont l'effet de levier est grand sont identifiés, une question importante est celle de savoir comment les traiter pour l'inférence. Deux options consistent soit à les sous-pondérer, soit à les éliminer entièrement de l'ajustement du modèle. La sous-pondération semble insatisfaisante en général, puisqu'un point peut avoir un effet de levier élevé non pas à cause d'un grand poids, mais plutôt parce qu'il possède une ou plusieurs valeurs X inhabituelles. La sous-pondération peut être raisonnables dans le contexte d'une approche fondée sur un modèle, en supposant que le modèle proprement dit est correctement spécifié. Cependant, la notion fondée sur le plan de sondage d'estimer un paramètre dans des conditions de recensement pourrait alors se perdre. Si un point possède un effet de levier important à cause de valeurs X extrêmes, il est possible qu'il ne suive pas du tout le modèle et qu'il doive être éliminé.

Cependant, utiliser une procédure mécanique qui rejette automatiquement de nombreuses observations influentes dont l'effet de levier est élevé peut donner lieu à des estimations de l'erreur-type trop petites, ce qui se traduit par

des intervalles de confiance dont la couverture est inférieure au taux nominal et à des taux d'erreur de type I exagérés dans les tests de vérification d'hypothèses (Li 2007). Ce phénomène est semblable aux problèmes bien connus qui se posent dans la régression séquentielle (Hurvich et Tsai 1990, Zhang 1992). Donc, un sujet de recherche intéressant semble être l'élaboration de procédures d'inférence pour la construction d'intervalles de confiance et l'exécution de tests de vérification d'hypothèse qui tiennent compte des effets de l'élimination ou de la sous-pondération de certains points.

Pour les données d'enquête complexes, la matrice chapeau ne fait intervenir aucune caractéristique du plan de sondage, sauf les poids de sondages et peut être utilisée pour détecter les cas dont les poids ou les valeurs des variables explicatives sont atypiques. D'autres statistiques diagnostiques, comme la distance de Cook, contiennent des estimations de la variance qui doivent tenir compte des caractéristiques du plan d'échantillonnage complexe, telles que la stratification et la mise en grappes. L'adaptation et l'extension d'approches diagnostiques supplémentaires à l'analyse des données d'enquête seront étudiées dans l'avenir.

7. Remerciements

Le présent article est fondé sur des travaux financés par la National Science Foundation des États-Unis en vertu de la subvention n° 0617081. Toutes les opinions, constatations et conclusions ou recommandations exprimées dans l'article sont celles des auteurs et ne reflètent pas forcément celles de la National Science Foundation. Nous remercions le rédacteur en chef et les examinateurs de leurs commentaires constructifs, qui nous ont permis d'améliorer considérablement l'article.

Annexe

Inclusion des unités sélectionnées avec certitude dans l'estimation de l'erreur-type

Dans l'étude empirique décrite à la section 5, nous avons inclus les unités sélectionnées avec certitude dans les calculs de l'erreur-type. Nous esquissons ici la justification de cette façon de procéder. Sous le modèle général (3), la variance sous le modèle de $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, l'estimateur utilisé dans l'étude empirique, est $\text{var}_M(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\mathbf{A}^{-1}\sigma^2$ où $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ et $\mathbf{V} = \text{diag}(v_i)_{i\in s}$. L'estimateur sandwich de variance utilisé dans l'étude présentée à la section 5 est défini comme étant

$$v(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\frac{n}{n-1}\sum_{i\in s}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{A}^{-1} \qquad (4)$$

où $\mathbf{z}_i = w_i e_i \mathbf{x}_i$ avec $e_i = Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ et $\bar{\mathbf{z}} = \sum_{i\in s} w_i e_i \mathbf{x}_i/n$. Cet estimateur est convergent par rapport au plan (voir Binder 1983) dans l'échantillonnage à un seul degré si les unités sont échantillonnées avec remise avec probabilité égale à w_i^{-1}, et qu'aucune unité n'est sélectionnée avec certitude. Si l'échantillon contient des unités sélectionnées avec certitude, la formule de $v(\hat{\boldsymbol{\beta}})$ doit être modifiée pour estimer la variance par rapport au plan : les unités sélectionnées avec certitude doivent être exclues des sommes figurant dans (4) et de $\bar{\mathbf{z}}$, et n doit être remplacé par n_{nc}, le nombre d'unités sélectionnées sans certitude. Dans le cas extrême d'un recensement, l'estimateur de variance par rapport au plan se réduirait à zéro.

L'estimateur donné par (4) est approximativement sans biais par rapport au modèle sous (3), que l'échantillon contienne ou non des unités sélectionnées avec certitude. Dans (4), la matrice du milieu peut être développée sous la forme $\sum_{i\in s}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T = \sum_{i\in s}\mathbf{z}_i\mathbf{z}_i^T - n\bar{\mathbf{z}}\bar{\mathbf{z}}^T$. En supposant que $e_i \approx Y_i - \mathbf{x}_i^T\boldsymbol{\beta}$, l'espérance par rapport au modèle sous (3) du premier terme est $E_M(\sum_{i\in s}\mathbf{z}_i\mathbf{z}_i^T) = \mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\sigma^2$, tandis que $E_M(n\bar{\mathbf{z}}\bar{\mathbf{z}}^T) = n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\sigma^2$. La substitution de ces espérances donne $E_M[v(\hat{\boldsymbol{\beta}})] = \text{var}_M(\hat{\boldsymbol{\beta}})$, expression qui est vérifiée même si certaines unités sont sélectionnées avec certitude. Cela montre aussi que $v(\hat{\boldsymbol{\beta}})$ est robuste au sens où il reflète correctement la contribution des variances hétéroscédastiques dans (3) à la variance par rapport au modèle de $\hat{\boldsymbol{\beta}}$, même si $\mathbf{V}$ peut être inconnue et non prise en compte dans l'estimation de $\boldsymbol{\beta}$.

Bibliographie

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

Belsley, D.A., Kuh, E. et Welsch, R. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. New York : John Wiley & Sons, Inc.

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

Binder, D.A., et Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters, Chapitre 3 dans *Analysis of Survey Data*, (Éds. R. Chambers et C. Skinner). New York : John Wiley & Sons, Inc.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Chambers, R.L., Dorfman, A.H. et Sverchkov, M.Y. (2003). Nonparametric regression with complex survey data, Chapitre 11 dans *Analysis of Survey Data*, (Éds. R. Chambers et C. Skinner). New York : John Wiley & Sons, Inc.

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

Choudhry, G. (2000). The 1998 Survey of Mental Health Organizations Survey Design. Westat technical report prepared for Center for Mental Health Services, Substance Abuse and Mental Health Services Administration (SAMHSA), disponible sur demande au SAMHSA.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.

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

Duchesne, P. (1999). Estimateurs de calage robuste. *Techniques d'enquête*, 25, 47-60.

Gwet, J., et Rivest, L. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

Hoaglin, D.C., et Welsch, R.E. (1978). The hat matrix in regression and ANOVA (Corr: 78V32 p146). *The American Statistician*, 32, 17-22.

Hulliger, B. (1995). Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes. *Techniques d'enquête*, 21, 89-97.

Hurvich, C.M., et Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.

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

Li, J. (2007). *Regression Diagnostics for Complex Survey Data*: *Identification of Influential Observations*. Thèse de doctorat non-publiée, University of Maryland.

Li, J., et Valliant, R. (2006). Influence analysis in linear regression with sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3330-3337.

Manderscheid, R.W., et Henderson, M.J. (2002). Mental Health, United States, 2002. DHHS Publication No. SMA04-3938. Rockville MD USA : Substance Abuse and Mental Health Services Administration. Disponible au http://mentalhealth.samhsa.gov/publications/allpubs/SMA04-3938/AppendixA.asp

Moreno-Rebollo, J.L., Muñoz-Reyes, A. et Muñoz-Pichardo, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.

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

Potter, F.J. (1993). The effect of weight trimming on nonlinear survey estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.

Skinner, C.J. (2003). Introduction to Part B, Chapitre 6 dans *Analysis of Survey Data*, (Éds. R. Chambers et C. Skinner). New York : John Wiley & Sons, Inc.

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

Smith, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14 , 143-152.

Smith, T.M.F. (1989). Introduction to Part B, Chapitre 6 dans *Analysis of Complex Surveys*, (Éds. C.J. Skinner, D. Holt et T.M.F. Smith). New York : John Wiley & Sons, Inc.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York : John Wiley & Sons, Inc.

Weisberg, S. (2005). *Applied Linear Regression*, Troisième édition. New York : John Wiley & Sons, Inc.

Welsh, A.H., et Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society*, Series B, Methodological, 60, 413-428.

Zaslavsky, A.M., Schenker, N. et Belin, T.R. (2001). Downweighting influential clusters in surveys: Application to the 1990 post enumeration Survey. *Journal of the American Statistical Association*, 96 , 858-869.

Zhang, P. (1992). Influence after variable selection in linear regression models. *Biometrika*, 79, 741-746.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 29-40
Statistique Canada, N° 12-001-X au catalogue

Une méthode bootstrap pratique pour les tests d'hypothèses à partir des données d'enquête

Jean-François Beaumont et Cynthia Bocci [1]

Résumé

Le recours à la méthode bootstrap est de plus en plus répandu dans le contexte des enquêtes par sondage réalisées par les organismes statistiques nationaux. Dans la plupart des applications, plusieurs ensembles de poids bootstrap sont fournis aux analystes avec le fichier de microdonnées d'enquête. Jusqu'à présent, l'utilisation de la méthode en pratique semble avoir été limitée principalement aux problèmes d'estimation de la variance. Dans le présent article, nous proposons une méthode bootstrap pour les tests d'hypothèses au sujet d'un vecteur de paramètres inconnus d'un modèle quand l'échantillon a été tiré d'une population finie. Le plan d'échantillonnage probabiliste utilisé pour sélectionner l'échantillon peut être informatif ou non. Notre méthode s'appuie sur des statistiques de test fondées sur un modèle dans lesquelles sont intégrés les poids de sondage. Ces statistiques sont habituellement faciles à calculer en se servant de progiciels statistiques classiques. Nous approximons la distribution sous l'hypothèse nulle de ces statistiques pondérées fondées sur un modèle en utilisant des poids bootstrap. L'un des avantages de notre méthode bootstrap par rapport aux méthodes existantes de test d'hypothèses à partir des données d'enquête est qu'après avoir reçu les ensembles de poids bootstrap, les analystes peuvent l'appliquer très facilement, même s'ils ne disposent pas de logiciels spécialisés pour le traitement des données d'enquêtes complexes. En outre, nos résultats de simulation laissent entendre que, dans l'ensemble, la méthode donne des résultats comparables à ceux de la méthode de Rao-Scott et meilleurs que ceux des méthodes de Wald et de Bonferroni quand on teste des hypothèses au sujet d'un vecteur de paramètres d'un modèle de régression linéaire.

Mots clés : Poids bootstrap ; analyse de données d'enquête ; test d'hypothèses ; échantillonnage informatif ; régression linéaire ; paramètres d'un modèle.

1. Introduction

La méthode bootstrap est utilisée de plus en plus fréquemment dans le contexte des enquêtes par sondage réalisées par les organismes statistiques nationaux. Les principales raisons semblent être qu'elle permet de traiter plusieurs situations qu'il serait difficile de résoudre autrement (par exemple ajustement des poids pour la non-réponse, calage, statistiques non lisses) et qu'elle est commode pour les analystes. Dans la plupart de ses applications, plusieurs ensembles de poids bootstrap sont fournis aux analystes avec le fichier de microdonnées d'enquête ; aucune autre information sur le plan n'est donnée. Ces poids sont habituellement obtenus en supposant que les fractions d'échantillonnage de premier degré sont suffisamment faibles pour qu'un plan d'échantillonnage sans remise puisse être approximé correctement par un plan d'échantillonnage avec remise. Le lecteur trouvera dans Rao, Wu et Yue (1992) une description succincte, mais claire, d'une méthode de construction de poids bootstrap sous cette hypothèse dans le cas d'un plan d'échantillonnage stratifié à plusieurs degrés.

Jusqu'ici, l'utilisation de la méthode en pratique semble avoir été limitée principalement aux problèmes d'estimation de la variance (par exemple, Langlet, Faucher et Lesage 2003 ; Yeo, Mantel et Liu 1999 ; Hughes et Brodsky 1994). Dans le domaine de la recherche, les travaux ont visé principalement à découvrir une méthode bootstrap appropriée pour estimer la variance quand l'échantillon est tiré sans remise d'une population finie (voir Sitter 1992 ou Shao et Tu 1995, chapitre 6, pour une revue des méthodes). Certains auteurs ont également étudié le problème de la détermination des intervalles de confiance bootstrap pour un paramètre de population finie (par exemple, Rao et Wu 1988 ; Kovar, Rao et Wu 1988 ; Sitter 1992 ; Rao et coll. 1992). Autant que nous sachions, il ne semble exister aucune publication sur les tests d'hypothèses utilisant la méthode bootstrap dans le cas des sondages, quoique le problème ait été étudié dans le contexte de la statistique classique. Le lecteur trouvera dans Hall et Wilson (1991) une discussion des tests d'hypothèses bootstrap et dans Efron et Tibshirani (1993), un excellent compte rendu de la méthode bootstrap en statistique classique. Les travaux de Graubard, Korn et Midthune (1997), qui ont appliqué la méthode bootstrap paramétrique classique à des données d'enquête pour tester l'ajustement d'un modèle de régression logistique, méritent également d'être soulignés. Leur procédure est valide quand l'échantillonnage n'est pas informatif.

1. Jean-François Beaumont, Statistique Canada, Division de la recherche et de l'innovation en statistique, Pré Tunney, immeuble R.-H.-Coats, 16e étage, Ottawa (Ontario), Canada, K1A 0T6. Courriel : Jean-Francois.Beaumont@statcan.gc.ca ; Cynthia Bocci, Statistique Canada, Division des méthodes d'enquête auprès des entreprises, Pré Tunney, immeuble R.-H.-Coats, 11e étage, Ottawa (Ontario), Canada, K1A 0T6. Courriel : Cynthia.Bocci@statcan.gc.ca.

Le problème des tests d'hypothèses à partir des données d'enquête complexes a été bien étudié les 30 dernières années (par exemple Rao et Scott 1981 ; Fay 1985 ; Thomas et Rao 1987 ; Korn et Graubard 1990 ; Korn et Graubard 1991 ; Graubard et Korn 1993 ; Thomas, Singh et Roberts 1996 ; Rao et Thomas 2003). Toutefois, sauf peut-être pour estimer les variances-covariances inconnues intervenant dans ces méthodes, la méthode bootstrap ne semble pas encore avoir été étudiée pour les tests d'hypothèses. L'objectif du présent article est donc de proposer une méthode bootstrap pour les tests d'hypothèses au sujet d'un vecteur de paramètres inconnus d'un modèle quand l'échantillon a été tiré d'une population finie. Le plan d'échantillonnage probabiliste utilisé pour sélectionner l'échantillon peut être informatif ou non. Informellement parlant, l'échantillonnage est informatif quand le modèle vérifié pour l'échantillon sélectionné diffère de celui qui tient pour l'ensemble de la population ; sinon, l'échantillonnage n'est pas informatif.

Notre méthode s'appuie sur des statistiques de tests fondées sur un modèle dans lesquelles sont intégrés les poids de sondage. Habituellement, ces statistiques se calculent facilement à l'aide de progiciels classiques. Pour approximer la distribution sous l'hypothèse nulle de ces statistiques pondérées fondées sur un modèle, nous utilisons des poids bootstrap. L'un des avantages de notre méthode bootstrap par rapport aux méthodes existantes de tests d'hypothèses à partir des données d'enquête est que, après avoir reçu les ensembles de poids bootstrap, les analystes peuvent l'appliquer très facilement, même s'ils ne disposent pas de logiciels spécialisés pour le traitement des données d'enquêtes complexes.

À la section 2, nous présentons la notation et le problème. À la section 3, nous décrivons et justifions la méthode bootstrap que nous proposons pour les tests d'hypothèses à partir des données d'enquête. À la section 4, nous donnons un exemple de régression linéaire pour illustrer la théorie. À la section 5, nous décrivons brièvement les méthodes de Rao-Scott (Rao et Scott 1981), de Wald et de Bonferroni, qui sont des méthodes de rechange, dans le cas de tests d'hypothèses au sujet d'un vecteur des paramètres d'un modèle de régression linéaire. À la section 6, au moyen d'une étude par simulation, nous comparons ces méthodes et les comparons à la méthode bootstrap que nous proposons. Enfin, à la dernière section, nous concluons par un bref résumé et une discussion.

2. Préliminaires

Nous supposons qu'une population finie U de taille N a été créée conformément à un modèle, spécifié par l'analyste, qui décrit la distribution conditionnelle $F(\mathbf{y}_U|\mathbf{X}_U\,;\,\boldsymbol{\beta},\,\boldsymbol{\theta})$. Le vecteur $\mathbf{y}_U$ de dimension N contient les valeurs dans la population d'une variable dépendante y, $\mathbf{X}_U$ est une matrice à N lignes qui contient les valeurs dans la population d'un vecteur de variables indépendantes $\mathbf{x}$, $\boldsymbol{\beta}$ est un vecteur de dimension r des paramètres inconnus du modèle et $\boldsymbol{\theta}$ est un vecteur potentiel de paramètres inconnus supplémentaires du modèle. Nous souhaitons tester des hypothèses au sujet de $\boldsymbol{\beta}$, mais non de $\boldsymbol{\theta}$. Nous supposons également que, si l'on pouvait observer la population complète U, on utiliserait une statistique de test $t\,(U\,;\,\mathbf{c})$ pour tester l'hypothèse linéaire multiple $\mathrm{H}_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ contre l'hypothèse alternative $\mathrm{H}_1 : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$. La matrice $\mathbf{H}$ de dimensions $Q \times r$ est utilisée pour définir l'hypothèse à tester et $\mathbf{c}$ est un vecteur de dimension Q de constantes spécifiées par l'analyste. Idéalement, $t\,(U\,;\,\mathbf{c})$ est asymptotiquement un pivot ; autrement dit, elle suit une distribution asymptotique qui ne dépend d'aucun paramètre inconnu. Nous considérons les statistiques qui ont la forme quadratique suivante :

$$t\,(U\,;\,\mathbf{c}) = (\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c})'\,\{\mathbf{A}(U)\}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c}), \qquad (2.1)$$

où $\hat{\boldsymbol{\beta}}_U$ est un estimateur convergent de $\boldsymbol{\beta}$ sous le modèle et $\mathbf{A}(U)$ est une matrice de mise à l'échelle. Habituellement, $\mathbf{A}(U)$ est symétrique et définie positive.

À titre d'exemple, supposons que les y_k, pour toutes les unités de la population $k \in U$, sont des variables aléatoires indépendantes et identiquement distribuées de moyenne β et de variance θ, et que nous souhaitons tester l'hypothèse nulle $\mathrm{H}_0 : \beta = c$. Dans cet exemple, $Q = 1$, $r = 1$, $\mathbf{H} = 1$ et $\mathbf{X}_U = \mathbf{1}_U$, où $\mathbf{1}_U$ est un vecteur de population de valeurs un. Une statistique de test courante pour ce problème est

$$t\,(U\,;\,\mathbf{c}) = \frac{(\hat{\beta}_U - c)^2}{\hat{\theta}_U / N}, \qquad (2.2)$$

où $\hat{\beta}_U = \sum_{k\in U} y_k/N$ et $\hat{\theta}_U = \sum_{k\in U} (y_k - \hat{\beta}_U)^2/(N-1)$. La statistique (2.2) est de la même forme que (2.1) si nous posons que $A\,(U) = \hat{\theta}_U/N$. On suppose habituellement que cette statistique suit la distribution χ^2_1 ou $F_{1,\,N-1}$ sous l'hypothèse nulle.

Comme cela est ordinairement le cas, un échantillon aléatoire s de taille n est tiré de la population finie U conformément à un plan d'échantillonnage probabiliste donné $p(s)$. Puisque la variable dépendante y et, éventuellement, les variables indépendantes $\mathbf{x}$ ne sont pas observées pour les unités non échantillonnées, il pourrait être préférable d'utiliser la statistique $t\,(s\,;\mathbf{c})$ plutôt que $t\,(U\,;\mathbf{c})$. Dans l'exemple susmentionné, cela mènerait à $t\,(s\,;\mathbf{c}) = n\,(\hat{\beta}_s - c)^2/\,\hat{\theta}_s$, où $\hat{\beta}_s = \sum_{k\in s} y_k/n$ et $\hat{\theta}_s = \sum_{k\in s} (y_k - \hat{\beta}_s)^2/(n-1)$. Cependant, si l'échantillonnage est informatif par rapport au modèle, il pourrait être plus approprié, ce qui est indubitablement plus fréquent, d'utiliser une statistique de test pondérée de la forme

$$\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c}) = (\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'\{\hat{\mathbf{A}}(s,\ \mathbf{w}_s)\}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c}). \quad (2.3)$$

Le vecteur $\mathbf{w}_s$ de dimension n contient le poids de sondage de l'unité échantillonnée k dans son k^{e} élément, désigné par w_k, $\hat{\boldsymbol{\beta}}_{ws}$ est un estimateur pondéré de $\boldsymbol{\beta}$ et $\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ est un analogue pondéré de $\mathbf{A}\,(s)$ en ce sens que chaque unité échantillonnée k est pondérée par son poids de sondage w_k, alors qu'aucune pondération n'a lieu avec $\mathbf{A}\,(s)$. Nous avons donc $\hat{\mathbf{A}}\,(s,\ \mathbf{1}_s) = \mathbf{A}\,(s)$, où $\mathbf{1}_s$ est un vecteur échantillon de valeurs un. Par conséquent, la statistique $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ est aussi un analogue pondéré de $t\,(s\,;\ \mathbf{c})$ et nous avons $\hat{t}\,(s,\ \mathbf{1}_s\ ;\ \mathbf{c}) = t\,(s\,;\ \mathbf{c})$. Si la statistique $t\,(s\,;\ \mathbf{c})$ peut être calculée en se servant d'un progiciel classique qui n'a pas nécessairement été développé pour traiter des données d'enquête, la statistique $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ peut aussi être calculée en utilisant le même progiciel, à condition que ce dernier permette la pondération de chaque observation par son poids de sondage.

Ordinairement, le poids de sondage w_k, pour une unité $k \in s$, est égal à l'inverse de sa probabilité de sélection, qui peut alors être calée pour tenir compte de l'information externe connue (par exemple, Deville et Särndal 1992). Nous supposons que le plan d'échantillonnage et les poids de sondage sont construits de manière que les deux hypothèses suivantes soient vérifiées :

Hypothèse 1 : $\sqrt{n}\ (\hat{\boldsymbol{\beta}}_{ws} - \boldsymbol{\beta}) \xrightarrow{mp} N(\mathbf{0}, \boldsymbol{\Sigma})$, où $\xrightarrow{mp}$ désigne la convergence en distribution sous le modèle et le plan d'échantillonnage, et $\boldsymbol{\Sigma}$ est la matrice de variance-covariance asymptotique de $\sqrt{n}\,\hat{\boldsymbol{\beta}}_{ws}$ sous le modèle et le plan d'échantillonnage. La notation « m » désigne le modèle, tandis que la notation « p » désigne le plan d'échantillonnage probabiliste.

Hypothèse 2 : $n\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ est symétrique, définie positive et convergente sous m et p pour une matrice de mise à l'échelle fixe, définie positive et symétrique, $\tilde{\mathbf{A}}$.

Notons que l'hypothèse 2 n'exige pas que $\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ soit convergente sous p pour $\mathbf{A}\,(U)$. En effet, $N\mathbf{A}\,(U)$ sera habituellement convergente sous m pour $\tilde{\mathbf{A}}$. D'autres choix de la matrice de mise à l'échelle $\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ dans (2.3) sont possibles. Par exemple, elle pourrait être remplacée par un estimateur de la variance par rapport au plan de $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ sous échantillonnage aléatoire simple (par exemple, Rao et Scott 1981). Un autre choix est la statistique de Wald courante. On l'obtient en remplaçant $\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ dans (2.3) par $\hat{\mathbf{V}}_{mp}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, qui est un estimateur convergent sous m et p de $\mathbf{V}_{mp}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, la variance de $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ évaluée par rapport au modèle et au plan d'échantillonnage. Comme nous le mentionnons au paragraphe qui suit l'équation (2.3), l'un des avantages de l'utilisation d'une matrice de mise à l'échelle $\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ telle que $\hat{\mathbf{A}}\,(s,\ \mathbf{1}_s) = \mathbf{A}\,(s)$ est que la statistique de test résultante $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ peut alors être calculée directement en utilisant des progiciels classiques, à condition qu'ils permettent la pondération de chaque observation par son poids de sondage, ce qui est plus commode pour les utilisateurs de données d'enquête.

Poursuivant l'exemple susmentionné, nous pouvons définir notre statistique de test pondérée sous la forme

$$\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c}) = \frac{(\hat{\beta}_{ws} - c)^2}{\{(\hat{N} - 1)/(n - 1)\}\,(\hat{\theta}_{ws}/\hat{N})}, \quad (2.4)$$

où $\hat{N} = \sum_{k\in s} w_k$, $\hat{\beta}_{ws} = \sum_{k\in s} w_k y_k / \sum_{k\in s} w_k$ et $\hat{\theta}_{ws} = \sum_{k\in s} w_k (y_k - \hat{\beta}_{ws})^2/(\hat{N} - 1)$. Dans (2.4), la matrice de mise à l'échelle pondérée sous-jacente est $\hat{A}\,(s,\ \mathbf{w}_s) = \{(\hat{N} - 1)/(n - 1)\}\ (\hat{\theta}_{ws}/\hat{N})$, qui ne dépend pas de la façon dont les poids sont rééchelonnés. S'ils le sont de manière que $\sum_{k\in s} w_k = n$, ce que font habituellement les analystes, le facteur $(\hat{N} - 1)/(n - 1)$ disparaît. Le rôle de ce facteur, conjugué à d'autres conditions de régularité, est de satisfaire l'hypothèse 2. Si l'on choisit d'utiliser le système SAS®, la statistique de test (2.4) s'obtient en utilisant l'instruction WEIGHT dans les procédures standard. Si l'hypothèse nulle est vérifiée, il est bien connu que (2.4) ne suit malheureusement pas la distribution χ^2_1 ou $F_{1,\,n-1}$ sous le modèle et le plan d'échantillonnage.

Pour obtenir une procédure de test valide, nous devons approximer la distribution de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ sous l'hypothèse nulle. Nous pouvons pour cela utiliser le résultat suivant :

Résultat 1 : $\hat{t}\,(s, \mathbf{w}_s\,; \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q$, où λ_q pour $q = 1, \ldots, Q$, sont les valeurs propres de $\boldsymbol{\Lambda} = (\tilde{\mathbf{A}}^{-1})\,(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$ et Ω_q sont des variables aléatoires khi-carré possédant un degré de liberté.

La preuve du résultat 1, qui s'appuie sur les hypothèses 1 et 2, est présentée en annexe. Quand l'hypothèse nulle est vérifiée (c'est-à-dire $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$), nous avons donc

$$\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c}) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q\,\Omega_q. \quad (2.5)$$

Rao et Scott (1981) ont utilisé un résultat comparable pour construire leur procédure de test. Ils ont approximé une distribution telle que (2.5) par une distribution du khi-carré mise à l'échelle qui concorde avec les deux premiers moments estimés du deuxième membre de (2.5). Au lieu de cela, nous approximons la distribution de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ sous l'hypothèse nulle en utilisant des poids bootstrap, ce que nous décrivons à la section suivante.

Avant de fournir des détails sur notre procédure de test, il convient de souligner que, dans (2.3), $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ peut s'écrire

$$\begin{aligned}\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c}) &= \hat{t}(s,\ \mathbf{w}_s\ ;\ \mathbf{H}\boldsymbol{\beta}) \\ &\quad + 2(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})'\{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{c}) \\ &\quad + (\mathbf{H}\boldsymbol{\beta} - \mathbf{c})'\,\{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}\,(\mathbf{H}\boldsymbol{\beta} - \mathbf{c}). \quad (2.6)\end{aligned}$$

Sous l'hypothèse nulle, les deux derniers termes du deuxième membre de (2.6) disparaissent et nous avons $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c}) = \hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{H}\boldsymbol{\beta})$. Si l'hypothèse nulle est fausse, le troisième terme du deuxième membre de (2.6) domine les autres à mesure que la taille d'échantillon augmente, puisque les premier, deuxième et troisième termes sont d'ordre $O_p(1)$, $O_p(\sqrt{n})$ et $O_p(n)$ respectivement, à condition que les hypothèses 1 et 2 soient vérifiées. En outre, puisque $\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$ est définie positive, le troisième terme est toujours positif. Par conséquent, l'observation d'une grande valeur positive de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ comparativement à un centile élevé de la distribution de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{H}\boldsymbol{\beta})$ est une indication que l'hypothèse nulle pourrait être fausse.

3. La méthode bootstrap proposée

Soit w_k^* un poids bootstrap aléatoire pour l'unité k, obtenu en utilisant une méthode bootstrap telle que celle de Rao et coll. (1992), et soit $\mathbf{w}_s^*$ le vecteur de dimension n qui contient le poids bootstrap aléatoire w_k^* dans son k^{e} élément. L'estimateur par le bootstrap $\hat{\boldsymbol{\beta}}_{w^*s}$ s'obtient de la même manière que $\hat{\boldsymbol{\beta}}_{ws}$ en remplaçant le poids de sondage w_k par sa version bootstrap w_k^* pour chaque unité échantillonnée. Nous désignons aussi par $\mathbf{w}_s^{*b}$, pour $b = 1, \ldots, B$, les B vecteurs de dimension n contenant les poids bootstrap w_k^{*b} dans leur k^{e} élément. Ces B vecteurs sont tirés indépendamment et ont la même distribution que $\mathbf{w}_s^*$; celle-ci est appelée distribution bootstrap et est désignée par le symbole « * ». Le b^{e} estimateur par le bootstrap $\hat{\boldsymbol{\beta}}_{w^{*b}s}$ est défini d'une manière évidente.

Avant de décrire notre procédure de test bootstrap, nous commençons par introduire trois hypothèses supplémentaires relatives à la construction des poids bootstrap :

Hypothèse 3 : $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_{w^*s} - \hat{\boldsymbol{\beta}}_{ws}) \xrightarrow{\ *\ } N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, où $\xrightarrow{\ *\ }$ désigne la convergence en distribution bootstrap et $\hat{\boldsymbol{\Sigma}}$ est la matrice de variance-covariance bootstrap asymptotique de $\sqrt{n}\,\hat{\boldsymbol{\beta}}_{w^*s}$.

Hypothèse 4 : $n\hat{\mathbf{A}}(s,\ \mathbf{w}_s^*)$ est convergent en distribution bootstrap pour $n\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)$.

Hypothèse 5 : $\hat{\boldsymbol{\Sigma}}$ est convergent sous m et p pour $\boldsymbol{\Sigma}$.

Les hypothèses 3 et 4 sont les analogues bootstrap des hypothèses 1 et 2 et devraient être satisfaites avec la plupart des méthodes bootstrap (par exemple, celles décrites dans l'article de synthèse de Sitter 1992) et modèles (par exemple, modèle de régression linéaire, modèle de régression logistique). Pour des renseignements plus détaillés, le lecteur est invité à consulter Shao et Tu (1995, chapitre 6 ; en particulier la section 6.4.4).

L'hypothèse 5 nécessite un commentaire. Cette hypothèse équivaut à exiger que la variance bootstrap $\mathbf{V}_*(\hat{\boldsymbol{\beta}}_{w^*s})$ soit convergente sous m et p pour

$$\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) = \mathbf{E}_m\mathbf{V}_p(\hat{\boldsymbol{\beta}}_{ws}) + \mathbf{V}_m\mathbf{E}_p(\hat{\boldsymbol{\beta}}_{ws}). \qquad (3.1)$$

Autrement dit, la distribution bootstrap doit refléter la variabilité due au modèle ainsi qu'au plan d'échantillonnage. Malheureusement, les méthodes bootstrap fondées sur le plan standard ne reflètent que la variabilité due au plan d'échantillonnage, de sorte qu'elles ne dépistent que le premier terme du deuxième membre de (3.1). Donc, elles ne satisfont pas l'hypothèse 5 en général. Toutefois, si la fraction globale d'échantillonnage n/N est négligeable, le deuxième terme du deuxième membre de (3.1) devient lui aussi négligeable (par exemple, voir Binder et Roberts 2003), de sorte que l'approximation $\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) \approx \mathbf{E}_m\mathbf{V}_p(\hat{\boldsymbol{\beta}}_{ws})$ est appropriée et que les méthodes bootstrap fondées sur le plan peuvent être appliquées. Dans le cas de nombreuses enquêtes-ménages, la fraction globale d'échantillonnage est effectivement assez faible. En effet, les poids bootstrap sont souvent obtenus sous l'hypothèse que les fractions d'échantillonnage de premier degré sont faibles (par exemple, Rao et coll. 1992). L'élaboration de méthodes bootstrap qui reflètent les deux termes de (3.1) est un domaine sur lequel porteront de futurs travaux de recherche.

Sous les hypothèses 3 et 4, nous obtenons un deuxième résultat :

Résultat 2 : $\hat{t}\,(s,\ \mathbf{w}_s^*\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}) \xrightarrow{\ *\ } \sum_{q=1}^{Q}\hat{\lambda}_q\Omega_q$, où $\hat{\lambda}_q$, pour $q = 1, \ldots, Q$, sont les valeurs propres de $\hat{\boldsymbol{\Lambda}} = [n\hat{\mathbf{A}}\,(s,\ \mathbf{w}_s)]^{-1}(\mathbf{H}\hat{\boldsymbol{\Sigma}}\mathbf{H}')$ et les Ω_q sont de nouveau des variables aléatoires khi-carré indépendantes possédant un degré de liberté.

Nous omettons la preuve du résultat 2, car elle est fort semblable à celle du résultat 1 présentée en annexe. À partir des hypothèses 2 et 5, $\hat{\boldsymbol{\Lambda}}$ est convergent sous m et p pour $\boldsymbol{\Lambda}$. Donc, en utilisant les résultats 1 et 2, la distribution bootstrap de $\hat{t}\,(s,\ \mathbf{w}_s^*\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ est asymptotiquement la même que la distribution sous m et p de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{H}\boldsymbol{\beta})$, qui est elle-même identique à la distribution sous m et p de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ sous l'hypothèse nulle, c'est-à-dire celle que nous voulons approximer. Cela suggère la procédure de test bootstrap suivante :

i) obtenir les poids bootstrap, w_k^{*b}, pour $k \in s$ et $b = 1, \ldots, B$;
ii) calculer $\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, pour $b = 1, \ldots, B$;
iii) puisqu'une grande valeur de $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ donne lieu au rejet de l'hypothèse nulle, calculer le seuil de signification observé (valeur p) selon

$$\frac{\#\{\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}) > \hat{t}(s,\ \mathbf{w}_s\ ;\ \mathbf{c})\}}{B}.$$

L'hypothèse nulle est rejetée si la valeur est inférieure au seuil de signification α (par exemple, 5 %).

Notons que la statistique qui est soumise au bootstrap est $\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ et non $\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{c})$. L'utilisation de cette dernière ne refléterait pas convenablement la distribution sous l'hypothèse nulle et violerait donc la première ligne directrice donnée dans Hall et Wilson (1991).

Si $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ est un pivot, alors la deuxième ligne directrice de Hall et Wilson (1991) est également satisfaite. Le fait que $t\,(U\ ;\ \mathbf{c})$ est asymptotiquement un pivot facilite certainement l'obtention d'une meilleure procédure de test bootstrap. Cela ne garantit malheureusement pas que $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ sera également asymptotiquement un pivot, particulièrement si l'échantillonnage est informatif. Néanmoins, ne pas utiliser une statistique pivot ne rend pas la procédure de test susmentionnée non valide et ne réduit pas nécessairement sa puissance. Toutefois, cela pourrait réduire le niveau d'exactitude du test. Comme l'ont souligné Hall et Wilson (1991), il convient parfois de ne pas tenir compte de la deuxième ligne directrice. Le principal avantage de l'utilisation de la simple statistique (éventuellement non pivot) $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ dans (2.3) et la statistique bootstrap $\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ tient au fait que, une fois que les poids bootstrap ont été fournis dans le fichier de microdonnées, ces statistiques peuvent être calculées facilement au moyen de progiciels classiques qui ne tiennent pas compte des caractéristiques du plan d'échantillonnage. En outre, nous montrons à la section 5, au moyen d'une étude par simulation, que notre méthode de test bootstrap donne des résultats comparables à la méthode de Rao-Scott et de meilleurs résultats que celles de Wald et de Bonferroni.

4. Un exemple de régression linéaire

Afin de mieux illustrer la théorie dans un contexte pratique, supposons que, sachant $\mathbf{X}_U$, les variables aléatoires y_k, pour $k \in U$, suivent indépendamment une distribution de moyenne $\mathrm{E}_m(y_k \mid \mathbf{X}_U) = \mathbf{x}_k'\boldsymbol{\beta}$ et de variance $\mathrm{V}_m(y_k \mid \mathbf{X}_U) = \theta$, où $\mathbf{x}_k$ est un vecteur de dimension r de variables linéairement indépendantes pour l'unité k. Rappelons que nous désirons tester l'hypothèse nulle H_0 : $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ contre l'hypothèse alternative H_1 : $\mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$. Si nous pouvions observer la population entière, nous pourrions utiliser la statistique courante

$$t\,(U\ ;\ \mathbf{c}) = \frac{(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c})'\left(\mathbf{H}\left(\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{H}'\right)^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c})}{Q\,\hat{\theta}_U} \quad (4.1)$$

où

$$\hat{\boldsymbol{\beta}}_U = \left(\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\sum_{k\in U}\mathbf{x}_k y_k$$

et

$$\hat{\theta}_U = \frac{\sum_{k\in U}(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_U)^2}{N - r}.$$

Dans (4.1), la statistique $t\,(U\ ;\ \mathbf{c})$ suit la distribution $F_{Q,\,N-r}$ sous l'hypothèse nulle. Elle se réduit à (2.2) quand $Q = r = \mathbf{H} = \mathbf{x}_k = 1$ dans (4.1).

Une version d'échantillon pondérée de (4.1), qui peut s'écrire sous la forme de (2.3), est

$$\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c}) = \frac{(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'\left(\mathbf{H}\left(\sum_{k\in s}w_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{H}'\right)^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})}{Q\,\hat{\theta}_{ws}\,\{(\hat{N} - r)/(n - r)\}}, \quad (4.2)$$

où

$$\hat{\boldsymbol{\beta}}_{ws} = \left(\sum_{k\in s}w_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\sum_{k\in s}w_k\mathbf{x}_k y_k \quad (4.3)$$

et

$$\hat{\theta}_{ws} = \frac{\sum_{k\in s}w_k(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_{ws})^2}{\hat{N} - r}. \quad (4.4)$$

Par exemple, dans (4.2) la statistique $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$ pourrait être obtenue en utilisant l'instruction WEIGHT dans la procédure REG de SAS, à condition que $w_k > 0$, pour $k \in s$. Notons qu'elle satisfait l'hypothèse 2 et ne dépend pas de la façon dont les poids sont mis à l'échelle. De nouveau, si les poids sont rééchelonnés de manière que $\sum_{k\in s}w_k = n$, le facteur $(\hat{N} - r)/(n - r)$ disparaît dans (4.2). La statistique de test (4.2) se réduit à (2.4) quand $Q = r = \mathbf{H} = \mathbf{x}_k = 1$ dans (4.2), (4.3) et (4.4). La statistique bootstrap $\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, ainsi que $\hat{\boldsymbol{\beta}}_{w^{*b}{}_s}$ et $\hat{\theta}_{w^{*b}{}_s}$ s'obtiennent de manière comparable à $\hat{t}\,(s,\ \mathbf{w}_s\ ;\ \mathbf{c})$, $\hat{\boldsymbol{\beta}}_{ws}$ et $\hat{\theta}_{ws}$ dans (4.2), (4.3) et (4.4) respectivement, excepté que w_k est remplacé par w_k^{*b} et $\mathbf{c}$ est remplacé par $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$.

Remarque 1 : Notons que w_k^{*b} sera vraisemblablement nul pour certaines unités $k \in s$ (voir, par exemple, Rao et coll. 1992). Dans certains progiciels tels que SAS, le nombre d'observations utilisées dans l'analyse de la b^{e} réplique bootstrap, n^{*b}, est égal au nombre d'unités $k \in s$ pour lesquelles $w_k^{*b} > 0$. Ces progiciels peuvent utiliser $n^{*b} - r$ au lieu de $n - r$ pour calculer la statistique bootstrap $\hat{t}\,(s,\ \mathbf{w}_s^{*b}\ ;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$. Il faut s'assurer que $n - r$ est utilisé et, dans la négative, que la statistique bootstrap calculée au moyen de ces progiciels est rajustée correctement avant d'appliquer la méthode de test bootstrap proposée. Un moyen d'éviter ce problème consiste à ajouter une très petite valeur positive (par exemple 1×10^{-10}) à chaque poids bootstrap w_k^{*b}, pour $k \in s$, afin qu'aucune observation ne soit exclue du calcul de $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$.

Remarque 2 : Définissons la statistique bootstrap $\hat{t}_e(s, \mathbf{w}_s^{*b}; \mathbf{0})$ en remplaçant y_k par $e_k = y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_{ws}$ dans $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{0})$, pour chaque $k \in s$. Il n'est pas difficile de montrer que $\hat{t}_e(s, \mathbf{w}_s^{*b}; \mathbf{0}) = \hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ de sorte que notre méthode bootstrap peut être appliquée en utilisant $\hat{t}_e(s, \mathbf{w}_s^{*b}; \mathbf{0})$ ou $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ si l'on se sert d'un modèle de régression linéaire. L'utilisation de la première option est parfois plus commode dans le cas de certains progiciels. Il en a été ainsi dans le cas de notre étude par simulation, puisque l'utilisation de $\hat{t}_e(s, \mathbf{w}_s^{*b}; \mathbf{0})$ nous a permis d'éviter de devoir entrer manuellement les valeurs de $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ pour chaque échantillon sélectionné. Une explication informelle de l'égalité $\hat{t}_e(s, \mathbf{w}_s^{*b}; \mathbf{0}) = \hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ peut être obtenue en traitant $\hat{\boldsymbol{\beta}}_{ws}$ comme une quantité fixe, ce qui est le cas en réalité sous la distribution bootstrap. La statistique bootstrap $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ peut donc être interprétée comme une statistique destinée à tester l'hypothèse nulle $\mathrm{H}_0^*: \mathbf{H}\boldsymbol{\beta} = \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ ou, alternativement, $\mathrm{H}_0^*: \mathbf{H}\boldsymbol{\gamma} = \mathbf{0}$, où $\boldsymbol{\gamma} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ws}$. En supposant encore que $\hat{\boldsymbol{\beta}}_{ws}$ est fixe, nous pouvons réécrire notre modèle linéaire $\mathrm{E}_m(y_k \mid \mathbf{X}_U) = \mathbf{x}_k'\boldsymbol{\beta}$ sous la forme $\mathrm{E}_m(e_k \mid \mathbf{X}_U) = \mathbf{x}_k'\boldsymbol{\gamma}$. Ces observations semblent impliquer que l'utilisation de la statistique bootstrap $\hat{t}_e(s, \mathbf{w}_s^{*b}; \mathbf{0})$ équivaut à utiliser $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, ce qui est en effet vrai.

Remarque 3 : Nous avons déjà mentionné que l'instruction WEIGHT était nécessaire pour obtenir une statistique pondérée si la méthode de test bootstrap proposée est appliquée en utilisant la procédure REG de SAS. De plus, l'instruction TEST est nécessaire pour demander que la statistique souhaitée soit produite, ainsi que l'instruction « ODS OUTPUT TESTANOVA = » pour sauvegarder ces statistiques demandées dans un ensemble de données SAS spécifié par l'utilisateur.

5. Certaines autres méthodes pour la régression linéaire

À la présente section, nous décrivons brièvement certaines méthodes de test dans le contexte de la régression linéaire exposé à la section 4, à savoir deux méthodes naïves qui sont parfois utilisées en pratique, ainsi que des applications spécifiques des méthodes de Rao-Scott, Wald et Bonferroni. Elles seront toutes évaluées dans une étude par simulation à la section 6.

Les méthodes de Bonferroni, de Wald et de Rao-Scott, décrites aux sections 5.2, 5.3 et 5.4 respectivement, nécessitent toutes un estimateur convergent sous m et p $\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$ de $\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$. Dans l'étude par simulation de la section 6, nous avons utilisé l'estimateur de variance bootstrap

$$\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) = \frac{\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}_{w^{*b}s} - \hat{\boldsymbol{\beta}}_{ws})(\hat{\boldsymbol{\beta}}_{w^{*b}s} - \hat{\boldsymbol{\beta}}_{ws})'}{B}. \quad (5.1)$$

Il convient de souligner que la validité de l'hypothèse 5 est donc requise non seulement pour la méthode bootstrap que nous proposons, mais aussi pour les méthodes de Bonferroni, Wald et Rao-Scott.

5.1 Deux méthodes naïves

La version pondérée de la méthode naïve consiste à utiliser la statistique $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ donnée par (4.2), qui est comparée à la queue supérieure de la distribution $F_{Q, n-r}$. La version non pondérée s'appuie sur la statistique $\hat{t}(s, \mathbf{1}_s; \mathbf{c})$, qui est de nouveau comparée à la queue supérieure de la distribution $F_{Q, n-r}$. Bien qu'on ne puisse pas s'attendre à ce que ces deux méthodes donnent de bons résultats sous échantillonnage informatif, elles sont néanmoins utilisées fréquemment en pratique, surtout la version pondérée. Notons que si l'échantillonnage n'est pas informatif, la version non pondérée, qui ne tient pas compte du plan d'échantillonnage, aboutit à un test simple, valide et raisonnablement puissant.

5.2 Méthode de Bonferroni

La méthode de Bonferroni a été étudiée par Korn et Graubard (1990). Elle est simple à appliquer et ces auteurs ont montré qu'elle donnait de bons résultats dans leur étude empirique. Pour décrire cette procédure, représentons par $\mathbf{H}_q'$ la q^{e} ligne de $\mathbf{H}$ et par c_q le q^{e} élément de $\mathbf{c}$. Ensuite, calculons les Q statistiques pondérées

$$\hat{t}_q^{BON}(s; c_q) = \frac{(\mathbf{H}_q'\hat{\boldsymbol{\beta}}_{ws} - c_q)^2}{\mathbf{H}_q'\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})\mathbf{H}_q}. \quad (5.2)$$

Nous comparons la plus grande statistique $\hat{t}_q^{BON}(s; c_q)$, pour $q = 1, \ldots, Q$, à la queue supérieure de la distribution $F_{1, d}$ avec un seuil de signification révisé α / Q au lieu de α. Le nombre de degrés de liberté d est égal au nombre d'unités primaires d'échantillonnage échantillonnées moins le nombre de strates. Notons que cette méthode dépend en général de la paramétrisation du modèle utilisé.

5.3 Méthode F de Wald

Nous pouvons définir une version F de la statistique khi-carré de Wald standard avec nombre de degrés de liberté ajusté au dénominateur comme l'a proposé Fellegi (1980) sous la forme

$$\hat{t}^W(s; \mathbf{c}) = \frac{d - Q + 1}{Qd}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'(\mathbf{H}\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})\mathbf{H}')^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c}). \quad (5.3)$$

Nous comparons la statistique $\hat{t}^{W}(s;\mathbf{c})$ à la queue supérieure de la distribution $F_{Q,\,d-Q+1}$. Cette procédure est exécutée au moyen du progiciel SUDAAN (Research Triangle Institute 2004).

5.4 Méthode F de Rao-Scott

Une autre méthode consiste à utiliser une version F (voir Rao et Thomas 2003) de la statistique khi-carré ajustée de deuxième ordre de Rao et Scott (1981), qui est fondée sur la correction de Satterthwaite pour le nombre de degrés de liberté. Nous utilisons une adaptation de la méthode de ces auteurs pour la régression linéaire, comme elle est exécutée dans le progiciel SUDAAN (Research Triangle Institute 2004). La statistique est définie par

$$\hat{t}^{RS}(s;\mathbf{c}) = \frac{1}{\bar{\lambda}(1+a^2)Q^*} (\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'(\mathbf{H}\hat{\mathbf{V}}_{\mathrm{SRS}}(\hat{\boldsymbol{\beta}}_{ws})\,\mathbf{H}')^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c}), \quad (5.4)$$

où $\hat{\mathbf{V}}_{\mathrm{SRS}}(\hat{\boldsymbol{\beta}}_{ws})$ est un estimateur de la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}_{ws}$ sous un plan d'échantillonnage aléatoire simple, $\bar{\lambda}$ est la moyenne des valeurs propres de la matrice d'effets de plans généralisés $[\hat{\mathbf{V}}_{\mathrm{SRS}}(\hat{\boldsymbol{\beta}}_{ws})]^{-1}\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$, a est le coefficient de variation de ces valeurs propres et $Q^* = Q/(1+a^2)$. La statistique F de Rao-Scott $\hat{t}^{RS}(s;\mathbf{c})$ est comparée à la queue supérieure de la distribution $F_{Q^*,\,d}$.

6. Étude par simulation

Nous avons effectué une étude par simulation pour examiner le niveau et la puissance des méthodes de test susmentionnées dans le cas d'un échantillonnage informatif ainsi que non informatif. Aux sections 6.1 et 6.2, nous décrivons la création des populations et des échantillons, respectivement. Puis nous définissons l'hypothèse nulle qui doit être testée à la section 6.3, décrivons les méthodes évaluées à la section 6.4 et présentons les résultats de la simulation à la section 6.5.

6.1 Création des populations

Nous avons créé quatre populations de $N = 10\,000$ unités. Pour commencer, nous avons créé indépendamment une variable catégorique v_k pour chaque unité de population k telle que $v_k = i$, pour $i = 1, \ldots, I$, avec la probabilité $P(v_k = i) = 1/I$, où I est le nombre de catégories de v_k, qui a été fixé à 5. La variable dépendante y a été créée selon

$$y_k = \alpha_o + \alpha_1\left(v_k - \frac{(I+1)}{2}\right) + \sigma\varphi_k, \quad (6.1)$$

où $\varphi_k \sim N(0,\,1)$, $\alpha_o = 10$ et $\sigma = 3$. Les quatre populations que nous avons créées ne diffèrent que par le choix de α_1, qui contrôle la corrélation entre y et v. Nous avons considéré les valeurs $\alpha_1 = 0$, 0,25, 0,50 et 0,75.

6.2 Création des échantillons et des poids bootstrap

Pour chacune des quatre populations susmentionnées, nous avons sélectionné 5 000 échantillons aléatoires simples de taille 100 sans remise sous deux scénarios de stratification différents destinés à simuler l'échantillonnage informatif, ainsi que non informatif. Dans le cas de l'échantillonnage non informatif, les strates correspondent exactement aux cinq catégories de la variable v définies plus haut. Dans le cas de l'échantillonnage informatif, les strates sont définies par classification croisée de la variable v et d'une autre variable catégorique z qui dépend du terme d'erreur aléatoire $\sigma\varphi_k$ dans (6.1). Pour chaque unité de population k, nous avons créé la variable z de la façon suivante : $z_k = 1$, si $\sigma\varphi_k > 0$, et $z_k = 2$, autrement. Dans le cas informatif, cela donne dix strates qui sont construites par recoupement entre les cinq catégories de v et les deux catégories de z. Chacune des dix strates informatives contient environ 1 000 unités de population, tandis que chacune des cinq strates non informatives contient environ 2 000 unités de population.

En outre, nous avons utilisé deux scénarios distincts de répartition aux strates. Sous le scénario SCHEME_UNEQUAL, les 100 unités échantillonnées sont réparties entre les strates de la façon suivante :

Tableau 1
Tailles d'échantillon pour SCHEME_UNEQUAL

Informatif	v / z	1	2	3	4	5
	1	4	4	16	4	28
	2	4	8	4	24	4

Non informatif	1	2	v 3	4	5
	8	12	20	28	32

Sous le second scénario, désigné SCHEME_EQUAL, le même nombre d'unité est affecté à chaque strate de la façon suivante :

Tableau 2
Tailles d'échantillon pour SCHEME_EQUAL

Informatif	v / z	1	2	3	4	5
	1	10	10	10	10	10
	2	10	10	10	10	10

Non informatif	1	2	v 3	4	5
	20	20	20	20	20

Ces scénarios produisent deux ensembles très différents de poids d'échantillonnage. Les poids résultant de la répartition SCHEME_UNEQUAL sont beaucoup plus variables que ceux obtenus sous le scénario SCHEME_EQUAL.

Notons que nous définissons simplement le poids de sondage w_k comme étant l'inverse de la probabilité de sélection de l'unité k.

Enfin, pour chaque échantillon sélectionné, nous avons calculé 500 poids bootstrap fondés sur le plan pour chaque unité échantillonnée, comme l'ont décrit Rao et coll. (1992), entre autres. Dans notre application de cette méthode, chaque échantillon bootstrap a été sélectionné avec remise par échantillonnage aléatoire simple stratifié avec $n_h - 1$ tirages à partir des n_h unités d'échantillonnage dans la strate h. Cette méthode tient compte de la variabilité liée au plan d'échantillonnage (avec une légère surestimation de la variance par rapport au plan due à l'hypothèse d'échantillonnage avec remise), mais ignore la variabilité du modèle, ce qui est acceptable, puisque la fraction globale d'échantillonnage (1/100) est faible.

6.3 Hypothèses nulles

Pour chaque échantillon sélectionné, nous avons modélisé y_k sous forme d'une fonction de v_k en nous servant d'un modèle d'analyse de variance. Plus précisément, nous avons défini les variables indicatrices

$$x_{ik} = \begin{cases} 1, & \text{si } v_k = i, \\ 0, & \text{autrement,} \end{cases}$$

pour $i = 1, ..., I,$ et ajusté le modèle linéaire $y_k = \beta_0 + \sum_{i=1}^{I-1} \beta_i x_{ik} + \varepsilon_k$ en utilisant la méthode des moindres carrés pondérés, où ε_k est un terme d'erreur aléatoire de moyenne nulle et de variance constante. Nous avons envisagé de tester les deux hypothèses nulles suivantes :

$$\text{TEST1} : H_0 : \beta_1 = 0$$
$$\text{TEST2} : H_0 : \beta_1 = \beta_2 = ... = \beta_{I-1} = 0.$$

Notons que ces hypothèses sont toutes deux vraies pour la population obtenue avec $\alpha_1 = 0,$ tandis qu'elles sont fausses pour les autres populations. Les trois dernières populations sont utilisées pour évaluer la puissance des diverses méthodes de test étudiées.

6.4 Méthodes de test

Pour chaque échantillon sélectionné, nous avons testé les deux hypothèses nulles susmentionnées avec les cinq méthodes suivantes : la méthode bootstrap proposée, la méthode naïve (versions non pondérée et pondérée) décrite à la section 5.1, la méthode de Bonferroni décrite à la section 5.2, la méthode F de Wald décrite à la section 5.3 et la méthode F de Rao-Scott décrite à la section 5.4. Les résultats pour la méthode naïve sont la sortie standard du logiciel SAS, tandis que les statistiques F de Wald et de Rao-Scott sont les sorties standard du logiciel statistique SUDAAN, version 9. La statistique de Bonferroni (5.2) est également obtenue au moyen de SUDAAN. La méthode proposée est programmée dans le logiciel statistique SAS, version 8.

De plus, nous avons exécuté l'étude par simulation en utilisant l'estimateur de variance linéarisé dans les méthodes de Wald, de Rao-Scott et de Bonferroni au lieu de l'estimateur de variance bootstrap (5.1). Les taux de rejet obtenus avec l'estimateur de variance linéarisé étaient légèrement inférieurs, mais assez semblables à ceux obtenus en utilisant (5.1). Étant donné cette observation et notre concentration sur les méthodes bootstrap, nous ne présentons et ne discutons pas ces résultats supplémentaires ici.

6.5 Résultats de simulation

Pour chaque population, scénario de stratification, scénario de répartition, hypothèse nulle et méthode, nous avons calculé le taux de rejet en pourcentage sur les 5 000 échantillons sélectionnés (en utilisant un seuil de signification de 5 %). Les résultats sont présentés ci-après dans les tableaux 3A, 3B, 4A et 4B. Ils sont plus frappants et plus intéressants pour l'hypothèse nulle TEST2 que pour l'hypothèse TEST1. Par conséquent, nous axerons notre discussion sur les résultats concernant la première.

Les tableaux 3A et 3B contiennent les résultats dans le cas de l'échantillonnage informatif, qui présente un plus grand intérêt pour nous. Commençons par examiner les résultats du tableau 3A pour SCHEME_UNEQUAL. Les deux méthodes naïves ont donné des résultats médiocres car elles n'exploitent pas convenablement l'information du plan d'échantillonnage. D'une part, la version non pondérée est définitivement trop libérale, car son taux de rejet est de loin supérieur à 5 % sous l'hypothèse nulle. D'autre part, la version pondérée est trop conservatrice et manque nettement de puissance comparativement aux autres méthodes. La méthode de Wald est trop libérale, le taux de rejet étant de 15,8 % quand H_0 est vraie. La méthode simple de Bonferroni améliore la situation, bien qu'elle soit encore trop libérale, avec un taux de rejet de 11,4 % quand H_0 est vraie. Ce résultat est un peu surprenant, car la méthode de Bonferroni a la réputation d'être (asymptotiquement) conservatrice. Un examinateur a suggéré que nous envisagions une méthode Bonferroni améliorée, telle que celle élaborée par Benjamini et Hochberg (1995). Dans la présente étude par simulation, ce genre de méthode ne serait pas utile, car elle produit systématiquement un taux de rejet plus élevé que la méthode de Bonferroni standard. La méthode de Rao-Scott l'emporte significativement sur les méthodes de Wald et de Bonferroni sous l'hypothèse nulle avec un taux de rejet de 6,8 %. La méthode bootstrap proposée est comparable à la méthode éprouvée, mais plus compliquée, de Rao-Scott, avec peut-être même une légère amélioration

du niveau, le taux de rejet étant de 6,2 % quand H_0 est vraie. Cependant, la méthode de Rao-Scott est un peu plus puissante que la méthode bootstrap proposée.

Le tableau 3B contient les résultats sous le scénario SCHEME_EQUAL dans le cas de l'échantillonnage informatif. Ici, les versions pondérée et non pondérée de la méthode naïve produisent des résultats semblables, puisque la variabilité des poids de sondage est assez faible. Même dans ce cas, la méthode naïve est définitivement trop conservatrice, ce qui se traduit par une puissance extrêmement faible. Toutes les autres méthodes sont comparables en ce qui concerne le niveau (H_0 vraie) et la puissance (H_0 fausse), quoique la méthode de Wald soit encore légèrement trop libérale comparativement à celles de Bonferroni et de Rao-Scott, ainsi qu'à la méthode bootstrap proposée, avec un taux de rejet de 7,9 % quand H_0 est vraie.

Les tableaux 4A et 4B donnent les résultats dans le cas de l'échantillonnage non informatif. De nouveau, commençons par discuter des résultats du tableau 4A pour le scénario SCHEME_UNEQUAL. Comme il fallait s'y attendre, la méthode naïve non pondérée donne de bons résultats ici, tandis que la variante pondérée devient trop libérale avec un taux de rejet de 12,8 % quand H_0 est vraie. En ce qui concerne le niveau, la méthode proposée fait concurrence à la méthode naïve non pondérée et est même légèrement conservatrice. Elle donne de meilleurs résultats que la méthode de Wald et est un peu meilleure que les méthodes de Bonferroni et de Rao-Scott. Sa puissance est toutefois légèrement plus faible que celle de ces deux dernières concurrentes, mais néanmoins acceptable.

Tableau 3A
Taux de rejet au seuil de signification de 5 % sous SCHEME_UNEQUAL et échantillonnage informatif

SCHEME_UNEQUAL	Échantillonnage informatif							
	Ho VRAIE $\alpha_1 = 0$		Ho FAUSSE $\alpha_1 = 0{,}25$		Ho FAUSSE $\alpha_1 = 0{,}50$		Ho FAUSSE $\alpha_1 = 0{,}75$	
Méthode	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve non pondérée	37,5	100,0	85,3	100,0	98,8	100,0	100,0	100,0
Naïve pondérée	1,7	0,4	14,5	4,6	58,0	33,6	90,3	78,6
Wald	8,0	15,8	30,9	37,1	71,8	73,9	93,1	95,4
Rao-Scott	8,0	6,8	30,9	21,1	71,8	61,7	93,1	91,8
Bonferroni	8,0	11,4	30,9	32,6	71,8	68,8	93,1	91,9
Bootstrap proposé	7,4	6,2	29,4	19,7	70,2	59,7	92,8	91,0

Tableau 3B
Taux de rejet au seuil de signification de 5 % sous SCHEME_EQUAL et échantillonnage informatif

SCHEME_EQUAL	Échantillonnage informatif							
	Ho VRAIE $\alpha_1 = 0$		Ho FAUSSE $\alpha_1 = 0{,}25$		Ho FAUSSE $\alpha_1 = 0{,}50$		Ho FAUSSE $\alpha_1 = 0{,}75$	
Méthode	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve non pondérée	0,1	0,0	6,7	0,3	58,1	16,5	97,2	79,7
Naïve pondérée	0,1	0,0	6,3	0,3	56,8	18,2	97,0	81,4
Wald	5,8	7,9	43,6	37,5	93,7	92,3	99,9	100,0
Rao-Scott	5,8	5,5	43,6	32,1	93,7	90,4	99,9	99,9
Bonferroni	5,8	6,2	43,6	33,6	93,7	88,6	99,9	99,8
Bootstrap proposé	2,3	5,1	42,3	31,0	93,6	89,6	99,9	99,9

Tableau 4A
Taux de rejet au seuil de signification de 5 % sous SCHEME_UNEQUAL et échantillonnage non informatif

SCHEME_UNEQUAL	Échantillonnage non informatif							
	Ho VRAIE $\alpha_1 = 0$		Ho FAUSSE $\alpha_1 = 0{,}25$		Ho FAUSSE $\alpha_1 = 0{,}50$		Ho FAUSSE $\alpha_1 = 0{,}75$	
Méthode	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve non pondérée	4,2	4,7	13,5	11,2	39,9	34,6	71,8	70,5
Naïve pondérée	11,4	12,8	24,6	23,0	56,8	50,2	83,8	81,2
Wald	7,6	8,6	16,8	17,8	42,9	42,6	72,5	76,2
Rao-Scott	7,6	6,4	16,8	12,3	42,9	32,1	72,5	72,5
Bonferroni	7,6	7,1	16,8	16,5	42,9	42,1	72,5	75,0
Bootstrap proposé	6,3	4,5	14,4	9,2	38,5	26,4	68,2	56,4

Tableau 4B
Taux de rejet au seuil de signification de 5 % sous SCHEME_EQUAL et échantillonnage non informatif

SCHEME_EQUAL	Échantillonnage non informatif							
	Ho VRAIE $\alpha_1 = 0$		Ho FAUSSE $\alpha_1 = 0{,}25$		Ho FAUSSE $\alpha_1 = 0{,}50$		Ho FAUSSE $\alpha_1 = 0{,}75$	
Méthode	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
Naïve non pondérée	4,9	4,5	17,2	12,4	54,3	42,2	88,2	81,7
Naïve pondérée	5,0	4,5	17,4	12,5	54,7	42,7	88,3	81,9
Wald	5,7	6,9	18,8	16,3	56,6	48,9	88,9	85,0
Rao-Scott	5,7	5,0	18,8	13,1	56,6	49,2	88,9	82,6
Bonferroni	5,7	5,4	18,8	13,7	56,6	45,1	88,9	81,8
Bootstrap proposé	5,0	3,3	16,4	10,0	53,2	36,5	86,8	77,6

Le tableau 4B contient les résultats sous le scénario SCHEME_EQUAL avec échantillonnage non informatif. Dans ce tableau, les méthodes ne semblent pas différer spectaculairement. Comme prévu, la méthode naïve (versions pondérée et non pondérée) donne de bons résultats bien qu'elle ne surpasse pas les méthodes de Rao-Scott et de Bonferroni dans cette étude par simulation. La méthode proposée reste légèrement conservatrice sous ce scénario non informatif et est légèrement moins puissante que les autres.

Pour étudier l'effet de grands échantillons sur les méthodes de test, nous avons également exécuté certaines simulations avec des tailles d'échantillon dix fois plus grandes que dans les conditions originales, comme l'a proposé un examinateur. Nous avons considéré une taille de population de 100 000 et sélectionné 1 000 échantillons de taille 1 000, ce qui nous a permis de maintenir délibérément la même faible fraction d'échantillonnage. Dans ces conditions, nous avons obtenu les résultats quand H_0 est vraie présentés au tableau 5 pour l'échantillonnage informatif ainsi que non informatif sous répartition inégale entre les strates. Comme nous nous y attendions, toutes les méthodes autres que les méthodes naïves ont produit des taux de rejet semblables qui étaient effectivement légèrement inférieurs à 5 %. Ces résultats montrent que les écarts entre les méthodes deviennent moins importants quand la taille d'échantillon augmente.

Dans l'ensemble, la méthode bootstrap que nous proposons était la meilleure en ce qui concerne le niveau, suivie de près par la méthode de Rao-Scott. Elle a toutefois donné des résultats un peu plus conservateurs sous les scénarios d'échantillonnage non informatif, ce qui a été accompagné d'une légère perte de puissance. La méthode de Rao-Scott est une bonne alternative si les utilisateurs ont accès à un logiciel approprié. La méthode de Bonferroni est simple à utiliser, mais est parfois trop libérale et la méthode de Wald est encore pire. Les méthodes naïves peuvent présenter des insuffisances sérieuses en ce qui concerne le niveau ou la puissance, quoique la méthode naïve non pondérée soit viable si l'on est raisonnablement certain que l'échantillonnage n'est pas informatif.

Tableau 5
Taux de rejet au seuil de signification de 5 % sous SCHEME_UNEQUAL

SCHEME_UNEQUAL	Informatif Ho VRAIE $\alpha_1 = 0$		Non informatif Ho VRAIE $\alpha_1 = 0$	
Méthode	Test1	Test2	Test1	Test2
Naïve non pondérée	100,0	100,0	3,7	3,8
Naïve pondérée	1,3	0,7	9,3	10,5
Wald	4,6	4,5	3,2	4,1
Rao-Scott	4,6	3,8	3,2	3,8
Bonferroni	4,6	4,5	3,2	3,6
Bootstrap proposé	4,4	3,6	2,9	3,8

7. Sommaire et discussion

Nous avons proposé une méthode bootstrap générale et simple de test d'hypothèses à partir des données d'enquête qui pourrait également être appliquée dans d'autres domaines que celui des sondages. Notre méthode s'appuie sur des statistiques classiques de test fondées sur un modèle, si bien qu'il est facile pour les analystes de l'appliquer en utilisant des progiciels classiques. Nous avons montré au moyen d'une étude par simulation qu'elle donne de bons résultats dans le contexte d'un modèle de régression linéaire. Ces bons résultats sont encourageants et laissent peut-être entendre que la méthode bootstrap que nous proposons pourrait être utile dans le cas d'autres modèles plus complexes et d'autres statistiques. L'idée pourrait également être adaptée facilement à la construction d'intervalles de confiance bootstrap.

Nous pourrions également envisager de traiter par le bootstrap une statistique asymptotiquement pivot, telle que celle de Rao-Scott (5.4). Cela comporterait cependant un

double bootstrap si $\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$ est estimé par la méthode bootstrap comme dans (5.1). Un double bootstrap nécessiterait la création d'un autre ensemble de répliques bootstrap pour chaque réplique bootstrap initiale. Il permettrait peut-être d'obtenir de meilleures procédures de test, mais risquerait de ne pas être commode pour les analystes. En étant axée sur des statistiques moins compliquées qui ne requièrent pas de procédure bootstrap, notre méthode de test évite le double bootstrap et reste simple.

Les propriétés de notre méthode dépendent non seulement du choix de la statistique de test, mais aussi de la construction des poids bootstrap. Habituellement, ceux-ci reflètent les deux premiers moments de l'erreur d'échantillonnage sous le plan, ce qui devrait suffire dans la plupart des cas à satisfaire nos hypothèses bootstrap 3, 4 et 5. Les poids bootstrap qui reflètent aussi le troisième moment sous le plan seraient peut-être utiles pour améliorer le niveau d'exactitude du test bootstrap. Cette question doit être étudiée plus en profondeur. Enfin, comme nous l'avons déjà mentionné à la section 3, les poids bootstrap standard fondés sur le plan ne satisfont l'hypothèse 5 que si la fraction d'échantillonnage globale est négligeable de sorte que la partie de la variance totale (3.1) due au modèle est négligeable. Les travaux doivent se poursuivre en vue d'élaborer des poids bootstrap appropriés si la fraction d'échantillonnage est non négligeable qui refléteront les composantes dues au modèle et au plan de la variance totale.

Remerciements

Nous remercions sincèrement le rédacteur associé et deux examinateurs de leurs commentaires. Nous remercions aussi J.N.K. Rao, de l'Université Carleton, ainsi que David Binder et Yves Lafortune, de Statistique Canada, de leurs commentaires et discussions intéressantes sur le sujet. Tous ces commentaires et discussions nous ont permis d'améliorer la qualité générale et la clarté de l'article.

Annexe

Preuve du résultat 1

Partant de l'hypothèse 1, nous voyons facilement que

$$\sqrt{n}\,(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} N(\mathbf{0}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'). \quad \text{(A.1)}$$

En utilisant un résultat standard sur des formes quadratiques (par exemple, Seber 1984, page 540) et l'équation (A.1), nous obtenons

$$n\,(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})'\tilde{\mathbf{A}}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q, \quad \text{(A.2)}$$

où λ_q, pour $q = 1, \ldots, Q$, sont les valeurs propres de $\boldsymbol{\Lambda} = (\tilde{\mathbf{A}}^{-1})(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$ et les Ω_q sont des variables aléatoires khi-carré indépendantes possédant un degré de liberté. Par conséquent, à partir de (A.2) et l'hypothèse 2, nous obtenons

$$\hat{t}\,(s, \mathbf{w}_s\,;\, \mathbf{H}\boldsymbol{\beta}) =$$

$$(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})'\{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q.$$

Bibliographie

Benjamini, Y., et Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Séries B, 57, 289-300.

Binder, D.A., et Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of survey data*, (Éds. R.L. Chambers et C.J. Skinner). New-York : John Wiley & Sons, Inc.

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

Efron, B., et Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York : Chapman & Hall.

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

Graubard, B.I., et Korn, E.L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.

Graubard, B.I., Korn, E.L. et Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 170-174.

Hall, P., et Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757-762.

Hughes, A.L., et Brodsky, M.D. (1994). Variance estimation of drug abuse episodes using the bootstrap. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 212-217.

Korn, E.L., et Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *The American Statistician*, 44, 270-276.

Korn, E.L., et Graubard, B.I. (1991). A note on the large sample properties of linearization, jackknife and balanced repeated replication methods for stratified samples. *The Annals of Statistics*, 19, 2275-2279.

Kovar, J.G., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, Supplément, 25-45.

Langlet, É.R., Faucher, D. et Lesage, É. (2003). An application of the bootstrap variance estimation method to the Canadian Participation and Activity Limitation Survey. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2299-2306.

Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

Rao, J.N.K., et Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: An appraisal and update. *Analysis of survey data*, (Éds. R.L. Chambers et C.J. Skinner). New-York : John Wiley & Sons, Inc.

Research Triangle Institute (2004). *SUDAAN language manual*, *release* 9.0. Research Triangle Park, NC : Research Triangle Institute.

Seber, G.A.F. (1984). *Multivariate Observations*. New-York : John Wiley & Sons, Inc.

Shao, J., et Tu, D. (1995). *The jackknife and the bootstrap*. New-York : Springer-Verlag.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, 20, 135-154.

Thomas, D.R., et Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Thomas, D.R., Singh, A.C. et Roberts, G.R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64, 295-311.

Yeo, D., Mantel, H. et Liu, T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 41-56
Statistique Canada, N° 12-001-X au catalogue

Méthodes bayésiennes pour un tableau de contingence à double entrée incomplet avec application aux sondages électoraux de l'État de l'Ohio

Bo-Seung Choi, Jai Won Choi et Yousung Park [1]

Résumé

Nous appliquons une méthode bayésienne pour résoudre le problème des solutions limites de l'estimation du maximum de vraisemblance (MV) dans un tableau de contingence à double entrée incomplet en utilisant un modèle log-linéaire et des lois a priori de Dirichlet. Nous comparons cinq lois a priori de Dirichlet pour estimer les probabilités multinomiales par case sous un modèle de non-réponse non ignorable. Trois de ces lois a priori ont été utilisées dans le cas d'un tableau à simple entrée incomplet et les deux autres sont deux nouvelles lois a priori proposées afin de tenir compte de la différence entre les profils de réponse des répondants et des électeurs indécis. Les estimations bayésiennes obtenues à l'aide des trois premières lois a priori n'ont pas systématiquement de meilleures propriétés que les estimations du MV, contrairement à ce qu'indiquaient des études antérieures, tandis que les deux nouvelles lois a priori donnent de meilleurs résultats que les trois lois a priori antérieures et que les estimations du MV chaque fois qu'est obtenue une solution limite. Nous utilisons quatre jeux de données provenant des sondages électoraux réalisés en 1998 dans l'État de l'Ohio pour illustrer comment il convient d'utiliser et d'interpréter les résultats des estimations pour les élections. Nous procédons à des études par simulation pour comparer les propriétés de cinq estimations bayésiennes sous un modèle de non-réponse non ignorable.

Mots clés : Analyse bayésienne ; non-réponse non ignorable ; tableau de contingence ; solution limite ; algorithme EM.

1. Introduction

Fréquente dans la plupart des sondages, la non-réponse devient un problème sérieux à mesure que son taux augmente (De Heer 1999 ; Groves et Couper 1998). Si l'on résume des données de sondage dans un tableau de contingence à double entrée, ce dernier contient des effectifs de case entièrement classifiés, des effectifs partiellement classifiés (c'est-à-dire non-réponse partielle) et des effectifs non classifiés (c'est-à-dire non-réponse totale). Par exemple, dans le sondage électoral de l'Ohio (Chen et Stasny 2003), l'une des catégories est la préférence de l'électeur (candidat A, B ou C, ou indécis) et l'autre est la probabilité de voter (votera vraisemblablement, ne votera vraisemblablement pas, et indécis). La première marge supplémentaire contient uniquement des données sur les préférences des électeurs, la deuxième contient uniquement des données sur la probabilité de voter et la troisième donne uniquement le nombre de non-réponses totales (cas où les deux réponses sont inconnues). Nous désirons intégrer ces observations manquantes dans l'estimation de l'appui réel pour chaque candidat et présenter des modèles bayésiens pour prédire le gagnant.

Dans certains sondages, les réponses indécises sont traitées comme une catégorie de réponse valide si les répondants ne manifestent pas de préférence nette pour un candidat ni d'intention nette de voter (Smith 1984 ; Rubin, Stern et Vehovar 1995). Néanmoins, de nombreuses études ont révélé que le comportement de vote des électeurs indécis peut avoir une incidence importante sur le résultat final et que l'on peut améliorer l'exactitude de la prédiction des résultats de l'élection en tenant compte de ces électeurs (Perry 1979 ; Fenwick, Wiseman, Becker et Heiman 1982 ; Myers et O'Connor 1983 ; Kim 1995 ; Chen et Stasny 2003 ; Martin, Traugott et Kennedy 2005). Parmi ces auteurs, Perry (1979) a montré au moyen de données empiriques obtenues selon une approche de scrutin secret que le pourcentage d'électeurs indécis observé dans un sondage est vraisemblablement plus élevé que le pourcentage réel. Kim (1995) a également indiqué que le rôle de ces électeurs indécis est critique, surtout quand leur nombre est supérieur à l'écart entre les deux candidats en tête dans une course électorale. Trois de nos études empiriques décrites à la section 3 appartiennent à ce cas critique. Fenwick et coll. (1982) ainsi que Kim (1995) ont procédé à une analyse discriminante des données du sondage électoral d'octobre 1980 au Massachusetts et des données de l'élection présidentielle américaine de 1992 d'après laquelle ils ont réparti les électeurs indécis entre les candidats afin de montrer qu'en général, ces électeurs indécis ne se rendent pas aux urnes dans les mêmes proportions que leurs homologues décidés. Si l'on met l'accent sur le candidat pour lequel pourrait voter l'électeur indécis, il est préférable de traiter les réponses indécises comme des données manquantes (Myers et O'Connor 1983). Comme l'ont mentionné Flannelly, Flannelly et McLeod (2000) et Lau

1. Bo-Seung Choi, directeur de recherche, Institut d'économie, Université de la Corée, Séoul 136-701, Corée ; Jai Won Choi, professeur, Département de biostatistique, Medical College of Georgia, Augusta, GA 30912 ; Yousung Park, professeur, Département de statistique, Université de la Corée, Séoul 136-701, Corée. Courriel : yspark@korea.ac.kr.

(1994), l'erreur de prédiction des résultats réels de l'élection augmente parallèlement au taux d'électeurs indécis. Afin de surmonter ce problème, Monterola, Lim, Garcia et Saloma (2001) ont adopté une approche basée sur des réseaux neuronaux pour classer les électeurs indécis dans un sondage d'opinion publique. Smith, Skinner et Clarke (1999) et Molenberghs, Kenward et Goetghebeur (2001) ont utilisé des méthodes d'imputation fondées sur un modèle pour le British General Election Panel Survey de 1992 et pour le sondage d'opinion publique réalisé à l'occasion du plébiscite de 1991 en Slovénie. Notre objectif principal étant d'obtenir des prédictions plus exactes grâce à l'affectation des électeurs indécis aux cases appropriées, nous avons traité ces derniers comme des observations manquantes à l'instar des chercheurs susmentionnés.

Les non-réponses (ou, de manière équivalente, les électeurs indécis) peuvent être réparties en trois catégories (Little et Rubin 2002, page 11), à savoir les données manquant complètement au hasard (MCAR pour *missing completely at random*), ce qui signifie que la probabilité d'une non-réponse pour une variable d'intérêt est indépendante de toute variable étudiée, y compris la variable elle-même, les données manquant au hasard (MAR pour *missing at random*), ce qui signifie que la probabilité d'une non-réponse dépend uniquement des données observées, et les données ne manquant pas au hasard (MNAR pour *missing not at random*), ce qui signifie que la probabilité d'une non-réponse dépend des valeurs non observées. Les modèles qui correspondent aux cas MCAR ou MAR sont appelés modèles de non-réponse ignorable, tandis que ceux correspondant aux cas MNAR sont appelés modèles de non-réponse non ignorable. Par exemple, dans un sondage préélectoral, si les personnes qui répondent n'indiquent pas leur préférence pour un candidat bien qu'elles appuient un candidat particulier, le profil des préférences pour les candidats pourrait ne pas être le même pour les répondants et les non-répondants. Dans ces conditions, le mécanisme de non-réponse est non ignorable. Si l'on suppose que le mécanisme de création des données manquantes est MCAR, l'effet de la non-réponse peut être éliminé dans l'inférence de la vraisemblance (Little et Rubin 2002, page 11). Cependant, si le profil de réponse des non-répondants diffère de celui des répondants, le fait d'écarter les non-réponses ou de spécifier incorrectement le mécanisme de création de la non-réponse donne lieu à des estimations entachées d'une plus grand variance et d'un plus grand biais (Chen 1972 ; Park et Brown 1994).

Dans les tableaux de contingence, si la non-réponse est non ignorable, l'estimation du MV donne fréquemment des solutions limites où il est estimé que la probabilité de non-réponse est nulle dans certaines cases. Ces solutions limites fournissent souvent un maximum local de la fonction de vraisemblance. Dans ce cas, les estimations du maximum de vraisemblance (MV) des paramètres du modèle log-linéaire ne peuvent pas avoir de solution unique et présentent habituellement de grands écarts-types (voir la section 4 ou Baker, Rosenberger et Dersimonian (1992), ainsi que Park et Brown (1994) pour des discussions plus détaillées).

Les conditions dans lesquelles l'estimation du MV se situe sur la solution limite ont été proposées dans le cas d'un tableau de contingence à simple entrée (Baker et Laird 1988 ; Michiels et Molenbergs 1997). L'explication géométrique de la solution limite de l'estimation du MV a été présentée (Smith et coll. 1999 ; Clark 2002). Baker et coll. (1992) ont énoncé une condition suffisante et nécessaire sous laquelle l'estimation du MV peut avoir une solution limite dans un tableau de contingence à double entrée.

Afin de contourner ce genre de problème de solution limite dans l'estimation du MV en présence de non-réponse non ignorable, Park et Brown (1994) et Park (1998) ont proposé une approche bayésienne au moyen de lois a priori empiriques basées uniquement sur l'information fournie par les répondants. Clogg, Rubin, Schenker et Schultz (1991) ont utilisé une loi a priori constante pour un tableau de contingence à simple entrée incomplet. Ils ont montré qu'en cas de non-réponse non ignorable, les méthodes bayésiennes donnent des erreurs quadratiques moyennes (EQM) plus petites que l'estimation du MV pour les espérances par case, mais notre étude par simulation révèle que cela n'est généralement pas vérifié dans un tableau de contingence à double entrée incomplet. Donc, nous présentons deux modèles bayésiens dont les lois a priori dépendent de l'information provenant à la fois des répondants et des électeurs indécis. Puis, nous appliquons chacun d'eux pour analyser le tableau de contingence à double entrée incomplet. L'extension à un tableau à multiples entrées est simple. Nous pouvons appliquer facilement cette extension à des données pondérées issues d'un échantillonnage stratifié ou en grappes en utilisant les covariables appropriées (voir la section 2.2).

La suite de l'article est divisée en quatre sections. À la section 2, nous considérons des modèles bayésiens avec cinq lois a priori différentes et présentons un algorithme Espérance-Maximisation (EM) généralisé pour estimer les probabilités par case. À la section 3, nous appliquons les modèles bayésiens à quatre jeux de données empiriques provenant du sondage électoral de l'État de l'Ohio et nous comparons les estimations bayésiennes à l'estimation du MV ainsi qu'aux résultats réels de l'élection. À la section 4, nous recourons à des études par simulation pour comparer les EQM et les biais des estimations bayésiennes fondées sur différents pourcentages de données manquantes et profils de réponse des répondants et des non-répondants. Dans cette

section, nous calculons également la probabilité de couverture, afin d'examiner la performance des estimations bayésiennes. À la section 5, nous présentons certaines conclusions.

2. Modèles bayésiens

Nous discutons de cinq estimations bayésiennes pour traiter la non-réponse non ignorable dans un tableau de contingence à double entrée incomplet. À la section 2.1, nous présentons un algorithme EM pour aborder le problème de la non-réponse dans un tableau de contingence à double entrée. Puis, à la section 2.2, nous spécifions cinq lois a priori et étendons notre approche à un tableau de contingence à multiples entrées.

Soit X_1 et X_2 ayant pour indice I et J catégories respectivement, dans un tableau de contingence à double entrée. Posons aussi que $R_1 = 1$ quand X_1 est observée et $R_1 = 2$ quand X_1 manque. De même, $R_2 = 1$ quand X_2 est observée et $R_2 = 2$ quand X_2 manque. Alors, la série complète des X_1, X_2, R_1, et R_2 produit un tableau de contingence $I \times J \times 2 \times 2$ contenant des effectifs entièrement classifiés, des effectifs partiellement classifiés et des effectifs non classifiés. Afin de distinguer ces trois catégories d'observations, représentons par y_{ijkl} l'effectif appartenant à la i^{e} catégorie de X_1, la j^{e} catégorie de X_2, la k^{e} valeur de R_1 et la l^{e} valeur de R_2. Donc, nous utilisons y_{ij11} pour les effectifs complètement classifiés, y_{i+12} et y_{+j21} pour les marges supplémentaires de colonne et de ligne respectives, et y_{++22} pour les effectifs non classifiés. Nous supposons que ces trois catégories d'observations suivent une loi multinomiale afin d'obtenir la log-vraisemblance suivante :

$$l = \sum_i \sum_j y_{ij11} \cdot \log(\pi_{ij11}) + \sum_i y_{i+12} \cdot \log(\pi_{i+12}) + \sum_j y_{+j21} \cdot \log(\pi_{+j21}) + y_{++22} \cdot \log(\pi_{++22}) \qquad (1)$$

où $\pi_{ijkl} = \Pr[X_1 = i,\ X_2 = j,\ R_1 = k,\ R_2 = l]$ et $N = \sum_{i,j,k,l} y_{ijkl}$ est fixé.

Comme cette fonction de vraisemblance contient plus de paramètres que le nombre de degrés de liberté disponibles pour l'estimation, nous relions π_{ijkl} à des covariables pertinentes en utilisant une fonction log-linéaire. Puisque nous ne disposons pas de variables explicatives, nous n'en utilisons aucune. Cependant, des variables explicatives peuvent être intégrées facilement dans le modèle log-linéaire de la même façon qu'y sont intégrées les variables catégoriques (voir Baker et Laird 1988, ainsi que Park et Brown 1994 pour plus de précisions).

Nous définissons un modèle de non-réponse non ignorable pour toutes les variables X_1, X_2, R_1 et R_2 par

$$\log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}$$

$$\text{pour } i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, 2, \text{ et } l = 1, 2 \qquad (2)$$

où $m_{ijkl} = N \cdot \pi_{ijkl}$ est la fréquence (ou effectif) par case attendue pour la $(i, j, k, l)^{\text{e}}$ catégorie et la somme de chaque terme β sur n'importe lequel de ses indices supérieurs respectifs est nulle.

Ce modèle log-linéaire est saturé, puisque le nombre de paramètres est exactement le même que le nombre de cases observées provenant du tableau de convergence à double entrée incomplet. Il s'agit aussi d'un modèle de non-réponse non ignorable à cause des termes d'interaction entre X_1 et R_1 et entre X_2 et R_2, ce qui sous-entend que, pour chaque variable de réponse, la non-réponse dépend de la situation de cette variable. Le modèle log-linéaire est un outil utilisé fréquemment pour analyser les tableaux de contingence incomplets avec non-réponse non ignorable. Soit p le nombre de paramètres (c'est-à-dire β) à estimer. Nous introduisons le $p \times 1$ vecteur de plan d'expérience $\mathbf{z}_{ijkl}$ pour préciser l'affiliation de l'observation appartenant à la $(i, j, k, l)^{\text{e}}$ catégorie. Alors, le modèle log-linéaire donné en (2) peut être réécrit sous la forme

$$\log \mathbf{m} = Z\boldsymbol{\beta} \qquad (3)$$

où le vecteur $\mathbf{m}$ de dimensions $I \times J \times 2 \times 2$ est l'espérance par case et $\boldsymbol{\beta}$ est la représentation vectorielle des β. Afin d'éviter une solution limite de l'estimation du MV dans le modèle (2), nous imposons aux probabilités par case $(\pi_{ij11}, \pi_{ij12}, \pi_{ij21}, \pi_{ij22})$ des lois a priori de Dirichlet donnés par

$$\prod_i \prod_j \pi_{ij11}^{\delta_{ij11}} \cdot \pi_{ij12}^{\delta_{ij12}} \cdot \pi_{ij21}^{\delta_{ij21}} \cdot \pi_{ij22}^{\delta_{ij22}} \qquad (4)$$

où les hyperparamètres δ_{ijkl} sont spécifiés à la section 2.2. Ces lois a priori de Dirichlet produisent une forme explicite et commode d'une loi a posteriori, parce qu'ils sont conjugués à une loi multinomiale (Clogg et coll. 1991 ; Park et Brown 1994 ; Forster et Smith 1998). Avec (3), la loi multinomiale de (1) pour les observations et la loi a priori (4), nous obtenons la log-distribution a posteriori :

$$l_{pos} = \sum_i \sum_j y_{ij11} \cdot (\mathbf{z}_{ij11} \cdot \boldsymbol{\beta}) - \sum_i \sum_j y_{ij11} \cdot \log\left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right) + \sum_i y_{i+12} \cdot \log\left(\sum_j \exp(\mathbf{z}_{ij12} \cdot \boldsymbol{\beta}) \right)$$

$$
\begin{aligned}
&- \sum_i y_{i+12} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ \sum_j y_{+j21} \cdot \log\Big(\sum_i \exp(\mathbf{z}_{ij21} \cdot \boldsymbol{\beta}) \Big) \\
&- \sum_j y_{+j21} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ y_{++22} \cdot \log\Big(\sum_i \sum_j \exp(\mathbf{z}_{ij22} \cdot \boldsymbol{\beta}) \Big) \\
&- y_{++22} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ \sum_{i,\,j,\,k,\,l} \delta_{ijkl} \cdot (\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \\
&- \sum_{i,\,j,\,k,\,l} \delta_{ijkl} \cdot \log\Big(\sum_{i,\,j,\,k,\,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big). \qquad (5)
\end{aligned}
$$

L'équation (5) est assez complexe et nous utilisons donc l'algorithme EM pour estimer les paramètres (c'est-à-dire $\boldsymbol{\beta}$).

2.1 L'algorithme EM

Nous maximisons la loi a posteriori donnée en (5) sur le paramètre $\boldsymbol{\beta}$ en utilisant l'algorithme Espérance-Maximisation généralisé (EMG) (Dempster, Laird et Rubin 1977) comportant les étapes E et M qui suivent.

Étape E : En utilisant les variables augmentées y_{ij12}, y_{ij21}, et y_{ij22} pour $i = 1, ..., I$ et $j = 1, ..., J$, la loi a posteriori (5) peut s'écrire sous la forme

$$
\begin{aligned}
l_{a.\text{pos}} &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \log(\pi_{ij11}) \\
&+ \sum_i \sum_j (y_{ij12} + \delta_{ij12}) \log(\pi_{ij12}) \\
&+ \sum_i \sum_j (y_{ij21} + \delta_{ij21}) \log(\pi_{ij21}) \\
&+ \sum_i \sum_j (y_{ij22} + \delta_{ij22}) \log(\pi_{ij22}). \qquad (6)
\end{aligned}
$$

Pour déterminer l'espérance de la loi a posteriori du log donné par (6), nous calculons la moyenne sur les effectifs manquants y_{ij12}, y_{ij21} et y_{ij22}, sachant les estimations courantes des paramètres, π_{ijkl}^{old} et les sommes marginales y_{i+12}, y_{+j21} et y_{++22} :

$$
\begin{aligned}
E_{\text{old}}[l_{a.\text{pos}}] &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \cdot \log(\pi_{ij11}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij12} \mid \pi_{ijkl}^{\text{old}}, y_{i+12}] + \delta_{ij12}) \cdot \log(\pi_{ij12}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij21} \mid \pi_{ijkl}^{\text{old}}, y_{+j21}] + \delta_{ij21}) \cdot \log(\pi_{ij21}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij22} \mid \pi_{ijkl}^{\text{old}}, y_{++22}] + \delta_{ij22}) \cdot \log(\pi_{ij22}). \qquad (7)
\end{aligned}
$$

Puisque y_{ij12}, y_{ij21} et y_{ij22} sont des variables aléatoires multinomiales conditionnées sur les sommes marginales respectives y_{i+12}, y_{+j21} et y_{++22}, dans l'équation (7), les espérances conditionnelles sont données par

$$
E_{\text{old}}(y_{ij12} \mid \pi_{ijkl}^{\text{old}}, y_{i+12}) = y_{i+12} \frac{m_{ij12}^{\text{old}}}{m_{i+12}^{\text{old}}},
$$

$$
E_{\text{old}}(y_{ij21} \mid \pi_{ijkl}^{\text{old}}, y_{+j21}) = y_{+j21} \frac{m_{ij21}^{\text{old}}}{m_{+j21}^{\text{old}}},
$$

et

$$
E_{\text{old}}(y_{ij22} \mid \pi_{ijkl}^{\text{old}}, y_{++22}) = y_{++22} \frac{m_{ij22}^{\text{old}}}{m_{++22}^{\text{old}}}
$$

où $m_{ijkl}^{\text{old}} = N \cdot \pi_{ijkl}^{\text{old}}$.

Étape M : À cette étape, nous maximisons l'espérance de la loi a posteriori du log (7) en utilisant les pseudo-observations $\tilde{y}_{ij11} = y_{ij11} + \delta_{ij11}$, $\tilde{y}_{ij12} = y_{i+12}\, m_{ij12}^{\text{old}} / m_{i+12}^{\text{old}} + \delta_{ij12}$, $\tilde{y}_{ij21} = y_{+j21}\, m_{ij21}^{\text{old}} / m_{+j21}^{\text{old}} + \delta_{ij21}$ et $\tilde{y}_{ij22} = y_{++22}\, m_{ij22}^{\text{old}} / m_{++22}^{\text{old}} + \delta_{ij22}$. Nous imposons à ces pseudo-observations que leurs sommes marginales soient égales aux sommes marginales correspondantes des observations $\tilde{y}_{++11} = y_{++11}$, $\tilde{y}_{i+12} = y_{i+12}$, $\tilde{y}_{+j21} = y_{+j21}$ et $\tilde{y}_{++22} = y_{++22}$. Sous ces contraintes, les pseudo-observations sont alors

$$
y_{ijkl}^{*} = \begin{cases}
\tilde{y}_{ij11} \dfrac{y_{++11}}{y_{++11} + \delta_{++11}} & \text{pour k = 1 et l = 1} \\
\tilde{y}_{ij12} \dfrac{y_{i+12}}{y_{i+12} + \delta_{i+12}} & \text{pour k = 1 et l = 2} \\
\tilde{y}_{ij21} \dfrac{y_{+j21}}{y_{+j21} + \delta_{+j21}} & \text{pour k = 2 et l = 1} \\
\tilde{y}_{ij22} \dfrac{y_{++22}}{y_{++22} + \delta_{++22}} & \text{pour k = 2 et l = 2.}
\end{cases}
$$

Alors, l'espérance de la fonction a posteriori du log (7) devient

$$E_{\text{old}}[l_{a.\text{pos}}] = \sum_i \sum_j y^*_{ij11} \cdot \log(\pi_{ij11}) + \sum_i \sum_j y^*_{ij12} \cdot \log(\pi_{ij12}) + \sum_i \sum_j y^*_{ij21} \cdot \log(\pi_{ij21}) + \sum_i \sum_j y^*_{ij22} \cdot \log(\pi_{ij22}).$$

Cette équation a la même forme que la vraisemblance obtenue à partir d'un tableau de contingence à quatre entrées quand les effectifs par case y^*_{ijkl} sont entièrement observés. Donc, en utilisant la méthode itérative des moindres carrés repondérés (Agresti 2002, page 342), nous obtenons l'estimateur du maximum a posteriori (EMP) de $\boldsymbol{\beta}$ comme il suit :

$$\boldsymbol{\beta}^{(t+1)} = (Z^T \hat{V}_t^{-1} Z)^{-1} Z^T \hat{V}_t^{-1} \gamma^{(t)},$$

où $\gamma^{(t)}$ possède l'élément $\gamma^{(t)}_{ijkl} = \log m^{(t)}_{ijkl} + (y_{ijkl} - m^{(t)}_{ijkl})/m^{(t)}_{ijkl}$ et $\hat{V}_t = [\text{diag}(\mathbf{m}^{(t)})]^{-1}$. Enfin, nous itérons ces étapes E et M jusqu'à ce qu'un critère de convergence soit atteint. Nous choisissons comme critère de convergence $\varepsilon \leq 10^{-6}$, où ε est la différence entre deux fonctions a posteriori du log consécutives.

Soit $Y_{\text{obs}} = (y_{ij11}, y_{i+12}, y_{+j21}, y_{++22})$ et $Y_{\text{manq}} = (y_{ij12}, y_{ij21}, y_{ij22})$ pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$ le vecteur des effectifs observés et le vecteur des effectifs manquants, respectivement. Alors, la log-distribution a posteriori (5) peut s'écrire

$$l_{\text{pos}} = l(\boldsymbol{\beta} \mid Y_{\text{obs}}) = l(\boldsymbol{\beta} \mid Y_{\text{obs}}, Y_{\text{manq}}) - \log f(Y_{\text{manq}} \mid Y_{\text{obs}}, \boldsymbol{\beta}). \quad (8)$$

Par double dérivation par rapport à $\boldsymbol{\beta}$, (8) donne

$$\frac{\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \frac{\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}}, Y_{\text{manq}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - \frac{\partial^2 \log f(Y_{\text{manq}} \mid Y_{\text{obs}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$$

$$= -\mathbf{Z}^T [\text{diag}(\mathbf{m}) - \mathbf{m}\mathbf{m}^T/N] \mathbf{Z} + \mathbf{Z}^T [\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T] AB\mathbf{Z}, \quad (9)$$

où $\boldsymbol{\pi}$ est l'expression vectorielle des probabilités par case π_{ijkl} et A, B sont données par

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \text{diag}\left(\dfrac{y^2_{i+12}}{y_{i+12} + \delta_{i+12}} \dfrac{m_{ij12}}{m_{i+12}}\right) & 0 & 0 \\ 0 & 0 & \text{diag}\left(\dfrac{y^2_{+j21}}{y_{+j21} + \delta_{+j21}} \dfrac{m_{ij21}}{m_{+j21}}\right) & 0 \\ 0 & 0 & 0 & \text{diag}\left(\dfrac{y^2_{++22}}{y_{++22} + \delta_{++22}} \dfrac{m_{ij22}}{m_{++22}}\right) \end{pmatrix}$$

et

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_{IJ} - B^{12} & 0 & 0 \\ 0 & 0 & I_{IJ} - B^{21} & 0 \\ 0 & 0 & 0 & I_{IJ} - B^{22} \end{pmatrix}.$$

Ici, faute d'espace et puisque l'extension pour i et j généraux n'est pas difficile, nous illustrons B^{12}, B^{21} et B^{22} uniquement pour $I = 2$ et $J = 3$:

$$B^{12} = \begin{pmatrix} \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2112}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} \end{pmatrix},$$

$$B^{21} = \begin{pmatrix} \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \\ \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \end{pmatrix},$$

et

$$B^{22} = \begin{pmatrix} \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \end{pmatrix}.$$

Nous constatons que l'information provenant des données observées $\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T$ est égale à la différence entre l'information provenant des données augmentées et celle provenant des données manquantes. Comme le montrent Gelman, Carlin, Stern et Rubin (2004, page 103), l'inverse de l'information provenant des données observées évalué à l'EMP de $\boldsymbol{\beta}$ est la variance de l'EMP de $\boldsymbol{\beta}$.

2.2 Spécification des lois a priori

Afin d'achever l'algorithme EM, nous devons déterminer les hyperparamètres δ_{ijkl}. Nous posons que la somme des lois a priori $\sum_{i,j,k,l} \delta_{ijkl}$ est égale au nombre de paramètres intervenant dans le modèle log-linéaire, p, comme l'ont suggéré Clogg et coll. (1991). Sous cette contrainte, nous proposons cinq types de lois a priori de la façon suivante. Nous commençons par attribuer δ_{ijkl} de manière que l'EMP de m_{ijkl} diminue et se rapproche de l'EMV obtenu sous non-réponse ignorable. Autrement dit, nous déterminons les lois a priori δ_{ijkl} uniquement en fonction des effectifs de réponses connues y_{ij11} et nous les appelons lois a priori axées sur les répondants.

Le premier type de loi a priori axée sur les répondants est, pour tout $i = 1, \ldots, I$ et $j = 1, \ldots, J$,

$$\begin{aligned} \delta_{ij11} = \nabla_{11} \frac{y_{ij11}}{y_{++11}}, \delta_{ij12} &= \nabla_{12} \frac{y_{ij11}}{y_{++11}}, \delta_{ij21} \\ &= \nabla_{21} \frac{y_{ij11}}{y_{++11}}, \text{ et } \delta_{ij22} = \nabla_{22} \frac{y_{ij11}}{y_{++11}} \end{aligned} \qquad (10)$$

où $\nabla_{kl} = p \cdot y_{++kl} / y_{++++}$ pour $k = 1, 2$ et $l = 1, 2$. Le deuxième type de loi a priori axée sur les répondants ne donne aucune loi a priori (c'est-à-dire, comme il est décrit plus bas, qu'aucune loi a priori n'est nécessaire) sur π_{ij11} dans le premier type de lois a priori. En d'autres termes, le deuxième type est identique au premier type excepté que $\delta_{ij11} = 0$ pour tout i et j. Dans le cas d'un tableau de contingence à simple entrée (c'est-à-dire que X_1 ou X_2 est entièrement observée sans information manquante) et $y_{++22} = 0$, le premier type se réduit aux lois a priori utilisées dans Park (1998), tandis que le deuxième type se réduit aux lois a priori utilisées dans Park et Brown (1994). Ces deux types de lois a priori axées sur les répondants pourraient être trop simplistes, parce que l'on suppose généralement que le profil de réponse des non-répondants diffère de celui des répondants sous un modèle de non-réponse non ignorable. Par exemple, le candidat préféré des non-répondants pourrait ne pas être le même que celui des répondants dans un sondage préélectoral.

Afin de définir le troisième type de loi a priori, désignons par $\hat{m}_{ijkl}$ l'EMV de m_{ijkl}. Nous pouvons tirer la forme explicite de $\hat{m}_{ijkl}$ de Baker et coll. (1992) où certains $\hat{m}_{ijkl}$ pourraient être nuls à cause des solutions limites. Par exemple, quand une marge supplémentaire de colonne possède une solution limite dans un tableau 2×2 incomplet, les EMV sont $\hat{m}_{1j11} = y_{1j11}$,

$$\hat{m}_{2j11} = \frac{y_{2+11}(y_{2j11} + y_{+j21})}{y_{2+11} + y_{++21}}, \; \hat{m}_{ij12} = \hat{m}_{ij11} b_j$$

où b_j est la solution de $\sum_{j=1}^{2} y_{ij11} b_j = y_{i+12}$, $\hat{m}_{1j21} = 0$,

$$\hat{m}_{2j21} = \hat{m}_{2j11} \frac{y_{++21}}{y_{2+11}},\ \hat{m}_{1j22} = 0,$$

et $\hat{m}_{2j22} = \hat{m}_{2j12}\, y_{++22} / y_{2+12}$. Par conséquent, ces estimations du MV tiennent compte à la fois de l'information des répondants et des non-répondants. Nous pouvons aussi obtenir les estimations du MV au moyen de notre algorithme EM décrit à la section 2.1 en posant que $\delta_{ijkl} = 0$ pour tout i, j, k et l. En utilisant ces estimations du MV, nous définissons le troisième type de loi a priori sous la forme

$$\begin{aligned}
\delta_{ij11} &= \nabla_{11} \cdot \left(\frac{\hat{m}_{ij11}}{\hat{m}_{++11}} \right), \delta_{ij12} \\
&= \nabla_{12} \cdot \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}, \\
\delta_{ij21} &= \nabla_{21} \cdot \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}, \\
\delta_{ij22} &= \nabla_{22} \cdot \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}
\end{aligned} \qquad (11)$$

et

où $\nabla_{kl} = p \cdot \hat{m}_{++kl} / \hat{m}_{++++}$ pour $k, l = 1, 2,$ et le terme $1/IJ$ est la loi a priori constante de Clogg et coll. (1991) afin d'éviter d'éventuelles solutions limites pour m_{ij12}, m_{ij21}, et m_{ij22} (voir aussi la cinquième loi a priori plus loin). Donc, nous attribuons la troisième loi a priori de δ_{ijkl} pour que l'EMP de m_{ijkl} diminue et se rapproche de l'EMV obtenu sous le modèle de non-réponse non ignorable, tandis que la première loi a priori est obtenue sous le modèle de non-réponse ignorable.

Nous définissons le quatrième type de loi a priori en posant que $\delta_{ij11} = 0$ dans (11) comme nous l'avons fait pour obtenir le deuxième type de loi a priori à partir du premier type. Le dernier type de loi a priori, tiré de Clogg et coll. (1991) est défini comme il suit

$$\delta_{ij11} = 0, \delta_{ij12} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right), \delta_{ij21} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right),$$

et

$$\delta_{ij22} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right). \qquad (12)$$

Ces cinq types de lois a priori sont résumés au tableau 1 et sont comparés à la section suivante en utilisant des données empiriques et des études par simulation.

Tableau 1
Cinq types de lois a priori δ_{ijkl} ($\hat{m}_{ijkl}$ est l'EMV, I et J sont les nombres de lignes et de colonnes dans un tableau à double entrée, et p est le nombre de paramètres)

	δ_{ij11}	δ_{ij12}	δ_{ij21}	δ_{ij22}	
Type I	$\nabla_{11} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{12} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{21} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{22} \frac{y_{ij11}}{y_{++11}}$,	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y_{++++}}$
Type II	0	$\nabla_{12} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{21} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{22} \frac{y_{ij11}}{y_{++11}}$,	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y^*_{++++}}$
Type III	$\nabla_{11} \cdot \left(\frac{\hat{m}_{ij11}}{\hat{m}_{++11}} \right)$	$\frac{\nabla_{12}}{2} \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{21}}{2} \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{22}}{2} \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right)$,	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}_{++++}}$
Type IV	0	$\frac{\nabla_{12}}{2} \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{21}}{2} \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{22}}{2} \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right)$,	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}^*_{++++}}$
Type V	0	$\nabla_{12} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{21} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{22} \left(\frac{1}{I \cdot J} \right)$,	$\nabla_{kl} = \frac{p}{3}$

$y^*_{++++} = y_{++++} - y_{++11}$ et $\hat{m}^*_{++++} = \hat{m}_{++++} - \hat{m}_{++11}$

Jusqu'ici, nous avons présenté des méthodes pour un tableau à double entrée, et y_{ijkl} est défini pour l'effectif de la case (i, j) de la i^{e} ligne et de la j^{e} colonne (c'est-à-dire $X_1 = i, X_2 = j$), et l'indicateur R_1 pour une ligne manquante et R_2 pour une colonne manquante (c'est-à-dire $R_1 = k, R_2 = l$). L'extension à un tableau à triple entrée est facile. Soit y_{ijklmn} l'effectif de la $(i, j, k)^{\text{e}}$ case pour les trois variables de réponse (c'est-à-dire $X_1 = i$, $X_2 = j$ et $X_3 = k$) et les lignes et colonnes manquantes respectives (c'est-à-dire $R_1 = l$, $R_2 = m$ et $R_3 = n$ pour $l, m, n = 1, 2$). Donc, $lmn = 111$ implique que chacune des trois variables est observée, $lmn = 112$ implique que X_1 et X_2 sont observées, mais que X_3 manque ; de même pour $lmn = 121, 122, 211, 212, 221, 222, 1$ indique que la variable est observée et 2, qu'elle manque. Par conséquent, nous pouvons définir l'algorithme EM et les lois a priori pour un tableau de contingence à triple entrée incomplet. À l'étape E, l'espérance conditionnelle pour la $(i, j, k)^{\text{e}}$ case avec l'information inconnue de la marge k est donnée par

$$E_{\text{old}}(y_{ijk112} \mid \pi^{\text{old}}_{ijklmn}, y_{ij+112}) = y_{ij+112} \frac{m^{\text{old}}_{ijk112}}{m^{\text{old}}_{ij+112}}.$$

De même,

$$E_{\text{old}}(y_{ijk122} \mid \pi^{\text{old}}_{ijklmn}, y_{i++122}) = y_{i++122} \frac{m^{\text{old}}_{ijk122}}{m^{\text{old}}_{i++122}}$$

et

$$E_{\text{old}}(y_{ijk222} \mid \pi^{\text{old}}_{ijklmn}, y_{+++222}) = y_{+++222} \frac{m^{\text{old}}_{ijk222}}{m^{\text{old}}_{+++222}}.$$

Nous pouvons définir de la même manière d'autres espérances et cinq types de lois a priori.

Le sondage électoral de l'État de l'Ohio est réalisé par la méthode de composition aléatoire (CA). Si le sondage par CA est strictement autopondéré, aucune modification ne doit être apportée aux méthodes bayésiennes (Lavrakas 1993 ; Potthoff 1994). Cependant, la CA n'est pas toujours exécutée selon un plan autopondéré. Par exemple, un échantillon téléphonique est constitué de ménages et non de personnes. Si une personne est interviewée dans un ménage, la réponse doit être pondérée par le nombre de personnes dans le ménage. Un poids est également nécessaire pour les ménages possédant plus d'un numéro de téléphone. Si l'on dispose d'une estimation exacte du nombre total de ménages, on peut procéder à une stratification par région ou par État et il convient d'envisager une pondération dans une analyse complète. Les sondages électoraux réalisés en Ohio en 1998 ont été effectués par CA. Dans la présente étude, nos méthode et modèles ne comprennent pas de pondération pour tenir compte de la stratification, de la mise en grappes et d'autres facteurs donnant lieu à des probabilités de sélection différentes dans un sondage téléphonique.

Toutefois, il est possible de les étendre afin d'introduire ce genre de pondération. La simple extension qui suit montre comment tenir compte d'une stratification type. Dans un tableau à triple entrée, soit X_3 la troisième variable de réponse marquée de l'indice h ($h = 1, \ldots, H$) que nous supposons être toujours observée. Les H catégories peuvent être des strates dans un échantillonnage stratifié. Puisque X_3 est toujours observée, la variable indicatrice de données manquantes correspondante R_3 est égale à 1 et son observation peut être désignée par y_{ijhlm1}. Alors, pour chaque strate h, nous pouvons écrire la log-vraisemblance suivante :

$$l_h = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijh111} \log(\pi_{ijh11}) + \sum_{i=1}^{I} y_{i+h121} \log(\pi_{i+h12}) + \sum_{j=1}^{J} y_{+jh211} \log(\pi_{+jhk21}) + y_{++h221} \log(\pi_{++h22})$$

où $\pi_{ijhlm} = P[X_1 = i, X_2 = j, R_1 = l, R_2 = m \mid X_3 = h]$. Donc, la terminologie X_3 utilisée pour un tableau à triple entrée joue le rôle d'un indicateur pour les strates. Pour chaque strate h, la vraisemblance de (13) est exactement la même que celle d'un tableau à double entrée.

Alors, nous pouvons définir un modèle log-linéaire pour l'espérance par case $m_{ijhlm} = N_h \cdot \pi_{ijhlm}$ de la même façon qu'en (2), où $N_h = \sum_{i, j, l, m} y_{ijhlm}$ pour chaque $h = 1, 2, ..., H$. Un modèle de non-réponse non ignorable est donné par

$$\log(m_{ijhlm}) = \beta_{0h} + \beta^{i}_{X_1 h} + \beta^{j}_{X_2 h} + \beta^{l}_{R_1 h} + \beta^{m}_{R_2 h} + \beta^{ij}_{X_1 X_2 h} + \beta^{il}_{X_1 R_1 h} + \beta^{jm}_{X_2 R_2 h}. \quad (13)$$

Afin d'éviter le problème des solutions limites décrit à la section 2, nous utilisons les lois a priori de Dirichlet pour π_{ijhlm}

$$\prod_i \prod_j \pi^{\delta_{ijh11}}_{ijh11} \cdot \pi^{\delta_{ijh12}}_{ijh12} \cdot \pi^{\delta_{ijh21}}_{ijh21} \cdot \pi^{\delta_{ijh22}}_{ijh22}.$$

Ensuite, nous suivons exactement les mêmes procédures que celles illustrées à la section 2 pour estimer l'espérance par case m_{ijhlm} pour chaque $h = 1, 2\ldots, H$. L'estimation de l'espérance dans la $(i, j)^{\text{e}}$ case est

$$\hat{E}(y_{ij}) = \sum_{h=1}^{H} w_h \sum_{l, m} \hat{m}_{ijhlm}$$

où w_h est le poids connu pour la h^{e} strate et $\hat{m}_{ijhlm}$ est m_{ijhlm} évalué à l'EMP de $\boldsymbol{\beta}$. Par exemple, $w_h = N_h / \sum_h N_h$ est le poids pour un échantillon stratifié où N_h est la taille de la population de la h^{e} strate.

La matrice de variance-covariance d'une approximation de la distribution de $\hat{\mathbf{m}}$ est

$$\frac{\partial \hat{\mathbf{m}}^T}{\partial \boldsymbol{\beta}} \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{EMP}}) \frac{\partial \hat{\mathbf{m}}}{\partial \boldsymbol{\beta}} \qquad (14)$$

où $\hat{\mathbf{m}}$ est une expression vectorielle des estimations par case $\hat{m}_{ijhlm}$, $\hat{\boldsymbol{\beta}}_{\mathbf{EMP}}$ est l'EMP de $\boldsymbol{\beta}$ dont la variance $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{EMP}})$ est donnée par l'inverse de (9), et $\partial \mathbf{m} / \partial \boldsymbol{\beta} = N_h \times [\mathrm{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}^T]\mathbf{Z}$ où $\hat{\boldsymbol{\pi}}$ possède

$$\hat{\pi}_{ijhlm} = \pi_{ijhlm}(\hat{\boldsymbol{\beta}}_{\mathbf{EMP}}) = \frac{\exp(\mathbf{z}_{ijhlm}\hat{\boldsymbol{\beta}}_{\mathbf{EMP}})}{\sum_{k\in(i,j,h,l,m)} \exp(\mathbf{z}_k \hat{\boldsymbol{\beta}}_{\mathbf{EMP}})}$$

comme élément type.

3. Une application à un sondage électoral dans l'État de l'Ohio

Lorsque l'on veut prédire par sondage le gagnant d'une élection, l'exactitude du résultat dépend souvent de la façon dont sont traités les électeurs indécis qui voteront vraisemblablement, mais qui n'ont pas encore décidé quel est leur candidat préféré. Nous comparons les estimations bayésiennes basées sur les cinq types de lois a priori à l'estimation du MV en nous servant des données du Buckeye State Poll (BSP), ou sondage électoral dans l'État de l'Ohio, réalisé en 1998 par le Center for Survey Research de la Ohio State University. Les sondages préélectoraux du BSP ont produit des tableaux de contingence à double entrée incomplets dont une catégorie était le candidat préféré et l'autre, la probabilité de voter aux élections de novembre 1998 en vue d'élire le gouverneur, le procureur général, le maire de Columbus et le trésorier de l'Ohio. Le tableau 2, qui résume ces quatre sondages, révèle un nombre important d'électeurs indécis.

Pour la comparaison, nous considérons le modèle 1 de non-réponse ignorable et les modèles 2 et 3 de non-réponse non ignorable qui suivent.

$$\text{Modèle 1}: \quad \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2},$$

$$\text{Modèle 2}: \quad \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2},$$

$$\text{Modèle 3}: \quad \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{il}_{X_1R_2} + \beta^{jk}_{X_2R_1} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}.$$

Dans le modèle 1, les données manquent complètement au hasard et les cas pour lesquels des données manquent peuvent être ignorés dans les inférences de la vraisemblance. Les modèles 2 et 3 sont des modèles de non-réponse non ignorable où la probabilité qu'une variable manque dépend de la variable elle-même dans le modèle 2, tandis qu'elle dépend de l'autre variable dans le modèle 3. Notons que, sous les modèles 1 et 3, les estimations du MV ne se situent pas sur la limite de l'espace des paramètres comme l'ont montré Baker et coll. (1992). En outre, ayant constaté que, sous les modèles 1 et 3, les cinq estimations bayésiennes des effectifs attendus par case sont non seulement assez proches de l'estimation du MV, mais ont aussi presque le même écart-type, nous ne présentons les estimations du MV que pour les modèles 1 et 3.

Nous désignons les estimations du MV sous le modèle 1 de non-réponse ignorable, le modèle 2 de non-réponse non ignorable et le modèle 3 de non-réponse non ignorable par $IG1_{ML}$, $NON2_{ML}$ et $NON3_{ML}$, respectivement. *IG* et *NON* signifient ignorable et non ignorable, respectivement. Soit aussi $NON2_i^{BE}$ l'estimateur bayésien utilisant le i^e type de lois a priori sous le modèle 2. Autrement dit, $NON2_1^{BE}$ utilise les lois a priori axées sur les répondants de (10) et $NON2_2^{BE}$ utilise les mêmes lois a priori que $NON2_1^{BE}$, excepté que $\delta_{ij11} = 0$. De même, $NON2_3^{BE}$ est donné par (11) et $NON2_4^{BE}$ utilise les mêmes lois a priori excepté que $\delta_{ij11} = 0$. $NON2_5^{BE}$ est l'estimation bayésienne obtenue en utilisant les lois a priori constantes de (12). De plus, nous pouvons utiliser la méthode de Stasny (1986, 1988) pour estimer les effectifs prévus par case sous les modèles 1 et 3 qu'elle a supposés implicitement. Cependant, ses estimations semblent être exactement les mêmes que $IG1_{ML}$.

Tableau 2
Données observées pour les sondages préélectoraux du BSP

	Course à l'élection du gouverneur				Course à l'élection du procureur général		
	Fisher	Taft	Autre	Indécis	Montgomery	Cordray	Indécis
Votera probablement	112	140	23	61	197	82	57
Ne votera probablement pas	96	108	21	73	161	65	75
Indécis	7	11	1	4	15	4	0
	Course à l'élection du maire				Course à l'élection du trésorier		
	Coleman	Teater	Espy	Indécis	Deters	Donofrio	Indécis
Votera probablement	40	32	25	30	127	119	90
Ne votera probablement pas	37	47	41	56	127	90	84
Indécis	0	2	1	0	10	7	0

La partie supérieure du tableau 3 donne les résultats prédits des élections obtenus en utilisant uniquement les réponses « votera vraisemblablement » pour les quatre courses, ainsi que les écarts-types entre parenthèses. Les écarts-types sont proches les uns des autres et révèlent des écarts importants entre les premier et deuxième candidats en tête, excepté dans le cas de la course à l'élection du maire. Ce tableau contient aussi les résultats réels des élections et indique si les estimations du MV se situent ou non dans les solutions limites.

La partie inférieure du tableau donne les prédictions des résultats des élections en utilisant les réponses « votera vraisemblablement » et « ne votera vraisemblablement pas » pour voir ce qui se passerait si les personnes ayant déclaré qu'elles ne « voteraient vraisemblablement pas » se rendaient effectivement aux urnes. La comparaison des deux tableaux nous permet de conclure que les gagnants des élections du gouverneur, du procureur général et du trésorier restent les mêmes quelle que soit la probabilité déclarée de voter, tandis que le gagnant de l'élection du maire aurait pu changer si la plupart des personnes ayant déclaré qu'elles « ne voteraient vraisemblablement pas » avaient effectivement voté.

En nous basant sur le tableau 3, nous pouvons classer les sept estimations, sauf $NON2_{ML}$, en deux groupes : $NON2_3^{BE}$, $NON2_4^{BE}$ et $NON2_5^{BE}$ dans le premier groupe et les quatre autres estimations, $NON2_1^{BE}$, $NON2_2^{BE}$, $IG1_{ML}$ et $NON3_{ML}$, dans le deuxième groupe. Comme il fallait s'y attendre, puisque les lois a priori δ_{ijkl} pour $NON2_1^{BE}$ et $NON2_2^{BE}$ sont définis de façon telle que l'estimation de m_{ijkl} diminue et se rapproche du MV sous un modèle de non-réponse ignorable, ces deux estimations bayésienne sont très proches de $IG1_{ML}$ et offrent par conséquent peu d'avantage par rapport à cette dernière estimation. Il est également intéressant de souligner que $NON3_{ML}$ est presque identique à $IG1_{ML}$, bien que leurs modèles log-linéaires soient spécifiés différemment.

Il n'existe aucun critère général pour déterminer s'il convient d'utiliser un modèle de non-réponse ignorable ou un modèle de non-réponse non ignorable. Cependant, comme l'indique Chen et Stasny (2003), l'hypothèse de non-ignorabilité d'une non-réponse peut être raisonnable dans l'étude fondée sur le sondage électoral de l'État de l'Ohio parce que les gens hésitent à indiquer leur appui pour un candidat impopulaire ou que leurs préférences au moment du sondage ne sont pas fermes ou exactes. À cet égard, les estimations $NON2_1^{BE}$, $NON2_2^{BE}$ et $NON3_{ML}$ ne sont peut-être pas appropriées dans ces études de cas particulières, parce qu'elles sont presque les mêmes que les estimations $IG1_{ML}$ du modèle 1.

Tableau 3
Prédiction des résultats des élections basée sur les sondages électoraux de l'État de l'Ohio d'octobre 98 et d'avril 98 (les résultats sont exprimés en pourcentage et les chiffres entre parenthèses sont les écarts-types)

	Gouverneur			Maire			Procureur général		Trésorier	
	Fisher	Taft	Autre	Coleman	Teater	Espy	Mongomery	Cordray	Deters	Donofrio
Utilisation des réponses « votera vraisemblablement » seulement										
$NON2_{ML}$	33,2(2,75)	42,1(3,00)	24,8	31,5(4,65)	25,3(4,23)	43,2	75,6(3,71)	24,4	57,0(3,48)	43,0
$NON2_1^{BE}$	40,6(3,04)	48,5(3,27)	10,9	38,1(5,14)	34,2(4,78)	27,7	72,1(3,61)	27,9	52,7(3,36)	47,3
$NON2_2^{BE}$	40,9(3,01)	50,7(3,20)	8,40	39,9(5,04)	33,6(4,83)	26,5	71,0(3,59)	29,0	52,1(3,34)	47,9
$NON2_3^{BE}$	35,8(2,85)	44,5(3,08)	19,7	35,6(4,87)	29,3(4,51)	35,1	63,0(3,67)	37,0	54,3(3,41)	45,7
$NON2_4^{BE}$	36,3(2,87)	45,2(3,11)	18,6	35,9(4,91)	29,4(4,52)	34,6	63,0(3,64)	37,0	53,9(3,40)	46,1
$NON2_5^{BE}$	38,9(2,99)	47,4(3,20)	13,7	37,7(4,99)	33,6(4,77)	28,7	66,0(3,54)	34,0	51,5(3,32)	48,5
$IG1_{ML}$	40,6(3,03)	51,2(3,28)	8,20	40,8(5,16)	33,4(4,76)	25,8	70,9(3,59)	29,1	51,8(3,32)	48,2
$NON3_{ML}$	40,6(3,03)	51,2(3,28)	8,20	40,9(5,16)	33,3(4,75)	25,8	70,9(3,58)	29,1	51,7(3,32)	48,3
Résultat réel	45	50	5	39	37	24	63	37	57	43
Limite	oui			oui			oui		non	
Utilisation des réponses « votera vraisemblablement » + « ne votera vraisemblablement pas »										
$NON2_{ML}$	32,7(1,83)	39,4(1,91)	27,8	24,8(2,45)	26,2(2,49)	49,0	77,0(1,64)	23,0	60,2(1,93)	39,8
$NON2_1^{BE}$	41,3(1,93)	46,4(1,96)	12,3	30,7(2,68)	37,1(2,75)	32,2	72,8(1,74)	27,2	56,0(1,96)	44,0
$NON2_2^{BE}$	41,9(1,93)	49,2(1,95)	8,90	32,7(2,63)	36,5(2,76)	30,8	71,4(1,77)	28,6	55,3(1,96)	44,7
$NON2_3^{BE}$	35,4(1,87)	41,8(1,93)	22,7	27,8(2,55)	30,5(2,62)	41,7	61,0(1,72)	39,0	57,6(1,95)	42,4
$NON2_4^{BE}$	36,0(1,88)	42,6(1,93)	21,4	28,7(2,57)	30,6(2,62)	40,7	60,9(1,75)	39,1	57,2(1,95)	42,8
$NON2_5^{BE}$	39,1(1,91)	45,1(1,95)	15,8	30,7(2,63)	35,8(2,74)	33,5	64,8(1,88)	35,2	54,8(1,96)	45,2
$IG1_{ML}$	41,5(1,96)	49,8(1,96)	8,70	33,9(2,70)	36,1(2,74)	29,9	71,2(1,78)	28,8	55,0(1,96)	45,0
$NON3_{ML}$	41,5(1,96)	49,8(1,96)	8,70	34,1(2,71)	36,0(2,74)	29,9	71,1(1,78)	28,9	55,0(1,96)	45,0

Comparativement aux résultats réels des élections, $NON2_{ML}$ donne la pire prédiction pour le gouverneur, le maire et le procureur général, parce que les valeurs se situent sur une solution limite ; par contre, elle donne la meilleure prédiction pour le trésorier, parce que les valeurs ne se situent pas sur une solution limite. Dans le cas de l'élection du procureur général, $NON2_3^{BE}$ et $NON2_4^{BE}$ prédisent non seulement le résultat réel exact, mais diffèrent aussi assez bien des autres estimations. Puisque $NON2_3^{BE}$ et $NON2_4^{BE}$ utilisent les lois a priori destinées à refléter les profils de réponse différents des répondants et des électeurs indécis, nous pouvons inférer que la préférence de ces derniers pour un candidat diffère assez fortement de celle des répondants (autrement dit, $NON2_3^{BE}$ et $NON2_4^{BE}$ affectent 19,4 % des électeurs indécis qui voteront vraisemblablement à Montgomery et 80,6 % à Cordray, tandis que dans le tableau 2, les données indiquent que le pourcentage allant à Montgomery par opposition à Cordray est de 29,4 % contre 70,6 % chez les répondants qui voteront vraisemblablement).

Afin de visualiser cette différence entre les répondants et les électeurs indécis en ce qui a trait aux estimations des paramètres et d'examiner l'effet de l'occurrence de la solution limite sur les estimations sous le modèle 2 de non-réponse non ignorable, nous présentons au tableau 4 les estimations du MV et les estimations $NON2_3^{BE}$, ainsi que les écarts-types correspondants pour l'élection du procureur général. Comme il existe une solution limite, toutes les estimations du MV ont un écart-type trop grand comme il fallait s'y attendre. Par ailleurs, $NON2_3^{BE}$ est très stable. Puisque $\beta_{X_1X_2}^{11} = 0,0472$ est la plus petite et que son écart-type est relativement grand, nous négligeons le terme $\beta_{X_1X_2}^{11}$ pour rendre l'interprétation moins complexe. Sous $\beta_{X_1X_2}^{11} = 0$, il n'est pas difficile de montrer qu'en utilisant les estimations de $NON2_3^{BE}$ du tableau 4,

$$\log\frac{m_{1j1l}}{m_{2j1l}} = 2(\beta_{X_1}^{1} + \beta_{X_1R_1}^{11}) = 0,09$$

et

$$\log\frac{m_{1j2l}}{m_{2j2l}} = 2(\beta_{X_1}^{1} - \beta_{X_1R_1}^{11}) = 1,3916$$

pour chaque valeur fixée de j et l, et

$$\log\frac{m_{i1k1}}{m_{i2k1}} = 2(\beta_{X_2}^{1} + \beta_{X_2R_2}^{11}) = 0,8982$$

et

$$\log\frac{m_{i1k2}}{m_{i2k2}} = 2(\beta_{X_2}^{1} - \beta_{X_2R_2}^{11}) = -1,4942$$

pour chaque valeur fixée de i et k. Donc, en vertu de

$$\log\frac{m_{1j1l}}{m_{2j1l}} = 2(\beta_{X_1}^{1} + \beta_{X_1R_1}^{11}) = 0,09,$$

les personnes qui voteront vraisemblablement (c'est-à-dire $i = 1$) sont 1,09 fois (c'est-à-dire $e^{0,09}$) plus nombreuses que celles qui ne voteront vraisemblablement pas (c'est-à-dire $i = 2$) parmi les répondants ($k = 1$), tandis qu'en vertu de

$$\log\frac{m_{1j2l}}{m_{2j2l}} = 2(\beta_{X_1}^{1} - \beta_{X_1R_1}^{11}) = 1,3916,$$

les personnes qui voteront vraisemblablement ($i = 1$) sont 4,02 fois (c'est-à-dire $e^{1,3916}$) plus nombreuses que les personnes qui ne voteront vraisemblablement pas $i = 2$ parmi les électeurs indécis ($k = 2$) ; en vertu de

$$\log\frac{m_{i1k1}}{m_{i2k1}} = 2(\beta_{X_2}^{1} + \beta_{X_2R_2}^{11}) = 0,8982,$$

les personnes qui votent pour Montgomery sont 2,46 fois plus nombreuses que celles qui votent pour Cordray parmi les répondants, tandis qu'en vertu de

$$\log\frac{m_{i1k2}}{m_{i2k2}} = 2(\beta_{X_2}^{1} - \beta_{X_2R_2}^{11}) = -1,4942,$$

les personnes qui ne voteront vraisemblablement pas sont 4,46 fois plus nombreuses que celles qui voteront vraisemblablement parmi les électeurs indécis. Cela implique que le profil de réponse des répondants et des électeurs indécis est fort différent.

Tableau 4
MV et le troisième type d'estimation bayésienne sous le modèle 2 de non-réponse non ignorable pour l'élection du procureur général (les écarts-types sont indiqués entre parenthèses)

	β_0	$\beta_{X_1}^1$	$\beta_{X_2}^1$	$\beta_{R_1}^1$	$\beta_{R_2}^1$	$\beta_{X_1R_1}^{11}$	$\beta_{X_2R_2}^{11}$	$\beta_{X_1X_2}^{11}$	$\beta_{R_1R_2}^{11}$
$NON2_{ML}$	-3,3735	-1,9487	3,2134	4,8496	4,8186	2,0283	-2,7594	-0,0452	-1,5588
		(3,120)	(8,515)	(3,996)	(8,871)	(3,120)	(8,512)	(0,045)	(2,501)
$NON2_3^{BE}$	0,6860	0,3704	-0,1490	3,3024	2,2942	-0,3254	0,5981	0,0472	-1,5450
		(0,118)	(0,052)	(2,501)	(2,501)	(0,117)	(0,052)	(0,041)	(2,501)

La grandeur de cette différence peut être mesurée à l'aide des termes les plus importants, $\beta^{11}_{X_1R_1}$ et $\beta^{11}_{X_2R_2}$, dans le modèle 2 de non-réponse non ignorable. Puisque

$$\beta^{11}_{X_1R_1} = \frac{1}{4}\log\frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = -0,3254$$

et

$$\beta^{11}_{X_2R_2} = \frac{1}{4}\log\frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = 0,5981,\ \beta^{11}_{X_1R_1}$$

est le logarithme du rapport des cotes qui montre la différence logarithmique entre le ratio du nombre de ceux « qui voteront vraisemblablement » à ceux qui « ne voteront vraisemblablement pas » parmi les électeurs décidés pour Montgomery et le même ratio parmi les électeurs indécis qui préfèrent Montgomery, mais qui n'expriment pas leur probabilité de se rendre aux urnes. Par contre, $\beta^{11}_{X_2R_2}$ est le logarithme du rapport des cotes qui montre la log-différence entre le ratio du nombre d'électeurs en faveur de Montgomery au nombre d'électeurs en faveur de Cordray parmi les électeurs décidés qui voteront vraisemblablement et le même ratio parmi les électeurs indécis qui voteront vraisemblablement, mais qui n'indiquent pas quel est leur candidat préféré. Donc, parmi les électeurs en faveur de Montgomery, la possibilité que les électeurs indécis votent par rapport à celle qu'ils ne votent pas est d'environ 3,67 fois

$$\left(\text{c'est-à-dire } \frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = e^{4\times -0,3254} = 3,67^{-1}\right)$$

plus grande que la possibilité pour les électeurs décidés, ce qui veut dire que Montgomery doit mettre en œuvre une stratégie en vue d'accroître la participation des électeurs au scrutin. Par ailleurs, parmi les personnes qui voteront vraisemblablement, le taux de soutien pour Montgomery chez les électeurs décidés est environ 10,94 fois

$$\left(\text{c'est-à-dire } \frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = e^{4\times 0,5981} = 10,94\right)$$

plus grand que celui des électeurs indécis à l'égard de Montgomery, ce qui implique que la plupart des électeurs indécis n'indiquant pas qui est leur candidat préféré voteront vraisemblablement pour Cordray comme procureur général. Cela confirme également la croyance populaire selon laquelle les électeurs ont tendance à demeurer « indécis » dans un sondage s'ils appuient un candidat qui est considéré comme étant inférieur dans une course électorale et ils sont enclins à s'abstenir de voter s'ils appuient le candidat qui domine la course avec certitude.

4. Étude par simulation

Nous considérons un tableau de contingence 2×2 avec marges supplémentaires pour comparer la performance des cinq estimations bayésiennes décrites à la section 2 pour divers pourcentages de données manquantes et divers profils de réponse sous le modèle de non-réponse non ignorable suivant (c'est-à-dire le modèle 2) :

$$\log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}.$$

Donc, nous comparons uniquement $NON2_{ML}$ et $NON2^{BE}_i$ pour $i = 1, ..., 5$ dans cette étude par simulation.

Puisque chacune des variables X_1, X_2, R_1 et R_2 comprend deux niveaux, huit paramètres doivent être déterminés pour l'étude par simulation. D'après les équations de

$$4\beta^{11}_{X_1R_1} = \log\frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} \quad \text{et} \quad 4\beta^{11}_{X_2R_2} = \log\frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}},$$

$$\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0$$

signifie qu'il n'y a aucune différence entre les profils de réponse des répondants et des électeurs indécis. Ces profils de réponse diffèrent d'autant plus que les paramètres $\beta^{11}_{X_1R_1}$ et $\beta^{11}_{X_2R_2}$ sont grands. Nous faisons varier ces deux paramètres de 0,2 à 0,8 par incrément de 0,2. Nous fixons le pourcentage de valeurs manquantes à 20 % et à 30 % en ajustant $\beta^1_{X_1}$ et $\beta^1_{R_1}$ et en posant que

$$\frac{m_{1111}/m_{1211}}{m_{2111}/m_{2211}} = 5,\ \frac{m_{1111}/m_{1112}}{m_{1112}/m_{1122}} = 2,$$

et

$$N = \sum_{ijkl} m_{ijkl} = 1\,000.$$

Cela signifie que la taille et le pourcentage de valeurs manquantes pour la case de $X_1 = 1$ et $X_2 = 1$ sont approximativement égaux à cinq fois et deux fois la taille des trois autres cases, respectivement.

Nous produisons un grand nombre d'échantillons $\{y_{ijkl}, i, j, k, l = 1, 2\}$ dans les conditions susmentionnées jusqu'à ce que nous obtenions 1 000 échantillons aléatoires avec des solutions limites et 1 000 autres sans solution limite. L'occurrence d'une solution limite est déterminée par le critère donné dans Michiels et Molenberghs (1997) (voir aussi Clarke 2002, ainsi que Smith et coll. 1999 pour plus de précisions). En utilisant $\{y_{ij11}, y_{i+12}, y_{+j21}, y_{++22}, i, j, = 1, 2\}$ obtenu d'après les données générées, nous estimons les effectifs attendus par case m_{ijkl} au moyen des

cinq estimations bayésiennes et de l'estimation du MV décrites à la section 2.

Nous calculons les erreurs quadratiques moyennes (EQM) et les biais absolus de $NON2_{ML}$, $NON2_1^{BE}$, ..., $NON2_5^{BE}$ pour $\{\sum_{kl} m_{ijkl},\ i,\ j = 1, 2\}$. Puis, nous prenons la moyenne sur les quatre EQM et sur les quatre biais absolus que nous obtenons à partir de chaque estimation pour voir la performance globale de l'estimation. De même, nous calculons les EQM moyennes et les biais absolus moyens pour $\{m_{ij12} + m_{ij21} + m_{ij22}, i,\ j = 1, 2\}$ pour voir la performance de chaque estimation en présence d'imputation des non-réponses.

Le tableau 5 donne les ratios des EQM moyennes et des biais absolus moyens des cinq estimations bayésiennes (c'est-à-dire $NON2_1^{BE}$, ..., $NON2_5^{BE}$) par rapport à l'estimation du MV (c'est-à-dire $NON2_{ML}$) quand surviennent des solutions limites, tandis que le tableau 6 donne les mêmes ratios quand aucune solution limite ne se présente. Donc, les valeurs inférieures à 1 impliquent que l'estimation bayésienne correspondante possède une EQM moyenne ou un biais absolu moyen plus faible que l'estimation du MV. Les deux tableaux ne présentent que les cas pour $\beta_{X_1R_1}^{11} < \beta_{X_2R_2}^{11}$ et pour un pourcentage de valeurs manquantes de 20 %, parce que les EQM et les biais sont presque symétriques autour de la coordonnée de $(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11})$. Ils augmentent si nous faisons passer le pourcentage de valeurs manquantes à 30 % en maintenant les mêmes profils d'EQM et de biais que pour le cas où 20 % de données manquent.

Le tableau 5, dans lequel survient une solution limite, montre que $NON2_1^{BE}$, $NON2_3^{BE}$, $NON2_4^{BE}$ ont une EQM plus faible que l'estimation du MV (c'est-à-dire $NON2_{ML}$) pour toutes les valeurs de $\beta_{X_1R_1}^{11}$ et $\beta_{X_2R_2}^{11}$, sauf $(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11}) = (0{,}8,\ 0{,}8)$. Ici, $NON2_3^{BE}$ a une plus petite EQM que l'estimation du MV. Cela est vrai pour les biais absolus. Par ailleurs, le tableau 6, où ne survient aucune solution limite, montre que seule $NON2_3^{BE}$ est comparable à l'estimation du MV en ce qui concerne l'EQM, mais qu'elle est légèrement biaisée. En particulier, $NON2_3^{BE}$ a une EQM plus faible que l'estimation du MV à condition que $\beta_{X_1R_1}^{11} \neq 0{,}8$ ou $\beta_{X_2R_2}^{11} \neq 0{,}8$ (c'est-à-dire que les profils de réponse des répondants et des non-répondants diffèrent peu).

Tableau 5
Ratios des EQM moyennes et des biais absolus moyens des estimations bayésiennes relativement à l'estimation du MV quand surviennent des solutions limites sous un pourcentage de valeurs manquantes de 20 % (les ratios des biais absolus figurent entre parenthèses)

	$(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
Pour $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0,2, 0,2)	0,68(0,66)	0,47(0,22)	0,76(0,76)	0,65(0,48)	0,42(0,05)
	(0,2, 0,4)	0,68(0,48)	0,57(0,20)	0,77(0,68)	0,60(0,29)	0,56(0,30)
	(0,2, 0,6)	0,67(0,23)	0,73(0,66)	0,77(0,57)	0,64(0,10)	0,69(0,64)
	(0,2, 0,8)	0,77(0,26)	1,08(1,55)	0,83(0,43)	0,76(0,28)	0,95(1,34)
	(0,4, 0,4)	0,65(0,32)	0,69(0,57)	0,76(0,63)	0,61(0,17)	0,65(0,52)
	(0,4, 0,6)	0,58(0,14)	0,83(0,90)	0,71(0,56)	0,56(0,06)	0,69(0,71)
	(0,4, 0,8)	0,75(0,36)	1,46(2,07)	0,78(0,36)	0,74(0,42)	1,12(1,61)
	(0,6, 0,6)	0,66(0,22)	1,35(1,73)	0,73(0,43)	0,66(0,16)	1,01(1,29)
	(0,6, 0,8)	0,85(0,87)	2,27(3,19)	0,76(0,17)	0,83(0,81)	1,52(2,35)
	(0,8, 0,8)	1,12(1,93)	3,58(5,49)	0,83(0,24)	1,04(1,67)	2,18(3,95)
Pour $\{m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0,2, 0,2)	0,57(0,63)	0,27(0,13)	0,69(0,74)	0,41(0,40)	0,28(0,31)
	(0,2, 0,4)	0,54(0,46)	0,37(0,34)	0,68(0,68)	0,42(0,24)	0,44(0,57)
	(0,2, 0,6)	0,51(0,19)	0,69(0,94)	0,65(0,55)	0,47(0,10)	0,69(0,88)
	(0,2, 0,8)	0,63(0,35)	1,39(2,08)	0,71(0,34)	0,62(0,47)	1,11(1,52)
	(0,4, 0,4)	0,49(0,35)	0,54(0,64)	0,65(0,64)	0,42(0,17)	0,57(0,76)
	(0,4, 0,6)	0,48(0,17)	0,98(1,24)	0,62(0,51)	0,45(0,17)	0,85(1,04)
	(0,4, 0,8)	0,62(0,44)	1,81(2,33)	0,67(0,35)	0,61(0,55)	1,35(1,81)
	(0,6, 0,6)	0,55(0,42)	1,70(1,90)	0,63(0,41)	0,54(0,40)	1,28(1,51)
	(0,6, 0,8)	0,78(0,92)	2,91(3,43)	0,69(0,14)	0,75(0,92)	1,96(2,64)
	(0,8, 0,8)	1,13(1,96)	4,63(5,72)	0,75(0,33)	1,02(1,77)	2,86(4,24)

Tableau 6
Ratios des EQM moyennes et des biais absolus moyens des estimations bayésiennes relativement à l'estimation du MV quand ne survient aucune solution limite sous un pourcentage de valeurs manquantes de 20 % (les ratios des biais absolus figurent entre parenthèses)

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
Pour $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0,2, 0,2)	0,99(3,37)	1,05(7,00)	0,94(2,51)	0,93(4,89)	1,06(8,96)
	(0,2, 0,4)	0,98(2,57)	1,21(5,13)	0,97(1,89)	1,00(3,26)	1,24(5,56)
	(0,2, 0,6)	1,04(2,18)	1,52(3,84)	0,95(1,67)	1,06(2,38)	1,43(3,71)
	(0,2, 0,8)	1,12(2,04)	1,75(3,53)	1,00(1,48)	1,13(2,14)	1,52(3,21)
	(0,4, 0,4)	1,03(2,40)	1,49(4,66)	0,97(1,69)	1,05(2,74)	1,39(4,46)
	(0,4, 0,6)	1,20(2,17)	2,11(3,85)	1,00(1,52)	1,22(2,24)	1,78(3,42)
	(0,4, 0,8)	1,28(2,09)	2,36(3,67)	1,05(1,45)	1,26(2,09)	1,86(3,12)
	(0,6, 0,6)	1,22(2,16)	2,49(3,90)	0,96(1,48)	1,21(2,15)	1,90(3,32)
	(0,6, 0,8)	1,52(1,99)	3,19(3,39)	1,11(1,38)	1,45(1,91)	2,29(2,77)
	(0,8, 0,8)	1,66(1,96)	3,64(3,27)	1,14(1,36)	1,52(1,83)	2,43(2,59)
Pour $\{m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0,2, 0,2)	0,88(2,59)	0,89(5,66)	0,87(2,26)	0,89(4,55)	1,21(8,69)
	(0,2, 0,4)	0,93(2,40)	1,27(4,86)	0,93(1,78)	1,00(3,08)	1,50(5,29)
	(0,2, 0,6)	1,09(2,11)	1,93(3,97)	0,98(1,40)	1,15(2,29)	1,85(3,61)
	(0,2, 0,8)	1,24(2,13)	2,36(3,90)	1,02(1,48)	1,27(2,18)	2,06(3,19)
	(0,4, 0,4)	1,03(2,18)	1,81(4,30)	0,96(1,60)	1,12(2,62)	1,85(4,39)
	(0,4, 0,6)	1,23(2,28)	2,62(4,28)	0,99(1,48)	1,29(2,42)	2,28(3,80)
	(0,4, 0,8)	1,42(2,05)	3,26(3,70)	1,07(1,42)	1,44(2,07)	2,53(3,09)
	(0,6, 0,6)	1,33(2,07)	3,22(3,95)	0,99(1,36)	1,36(2,14)	2,54(3,43)
	(0,6, 0,8)	1,65(2,09)	4,14(3,74)	1,13(1,43)	1,61(2,07)	2,98(3,13)
	(0,8, 0,8)	1,91(2,02)	4,48(3,50)	1,16(1,39)	1,66(1,93)	3,03(2,83)

Park et Brown (1994) ont utilisé $NON2_2^{BE}$ pour estimer les effectifs attendus par case dans un tableau à simple entrée incomplet sous un modèle de non-réponse non ignorable. Ils ont montré par des études en simulation que $NON2_2^{BE}$ avait une EQM plus petite que l'estimation du MV bien que son biais soit plus important. Cependant, les valeurs supérieures à 1 pour $NON2_2^{BE}$ dans les tableaux 5 et 6 indiquent qu'il n'en est pas ainsi dans un tableau à double entrée incomplet, indépendamment de la solution limite, et que les méthodes bayésiennes ne donnent pas systématiquement de meilleurs résultats que le MV, même en cas d'une solution limite. L'une des raisons pour lesquelles nos résultats de simulation diffèrent de ceux de Park et Brown (1994) quand a lieu une solution limite tient au choix de $(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$, Park et Brown n'ayant exécuté leur simulation que dans les conditions où $\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0,34$. Comme le montre le tableau 5, $NON2_2^{BE}$ est meilleure que le MV quand $\beta^{11}_{X_1R_1} \le 0,4$ et $\beta^{11}_{X_2R_2} \le 0,4$, tandis que $NON2_2^{BE}$ est pire que le MV quand les profils de réponse des répondants et des non-répondants sont fort différents (c'est-à-dire $\beta^{11}_{X_1R_1} \ge 0,6$ ou $\beta^{11}_{X_2R_2} \ge 0,6$).

Le tableau 7 donne la moyenne des écarts-types et des probabilités de couverture à 95 % pour $\beta^{11}_{X_1R_1}$. Ici, nous avons utilisé la formule de variance donnée dans (9) pour calculer les écarts-types, et les probabilités de couverture à 95 % sont les taux de couverture pour les intervalles de confiance à 95 % nominaux. Quand survient une solution limite, même si la probabilité de couverture de l'estimation du MV est celle qui s'approche le plus du niveau de couverture nominal de 95 %, l'estimation du MV possède un écart-type trop grand pour pouvoir être utilisé en pratique. Ces grands écarts-types sont dus au problème des solutions limites de l'estimation du MV. Parmi les estimations bayésiennes, les probabilités de couverture de $NON2_3^{BE}$ sont celles qui s'approchent le plus du niveau de couverture nominal de 95 %, tandis que celles des autres estimations sont généralement inférieures à ce niveau. Cela signifie qu'à part $NON2_3^{BE}$, les estimations bayésiennes sous-estiment les écarts-types.

Quand aucune solution limite ne se présente (deuxième partie du tableau 7), les écarts-types de l'estimation du MV sont nettement plus stables que ceux obtenus dans le cas d'une solution limite. La probabilité de couverture diminue à mesure que $\beta^{11}_{X_1R_1}$ et $\beta^{11}_{X_2R_2}$ augmentent. En particulier, les probabilités de couverture de NON_1^{BE}, NON_2^{BE} et NON_5^{BE} sont considérablement plus faibles que le niveau de couverture nominal de 95 % quand les profils de réponse des répondants et des électeurs indécis sont fort différents (c'est-à-dire $\beta^{11}_{X_1R_1} \ge 0,6$ et $\beta^{11}_{X_2R_2} \ge 0,6$).

Tableau 7
Moyenne des écarts-types et probabilités de couverture à 95 % (entre parenthèses) pour $\beta^{11}_{X_1R_1}$

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2_{ML}$	$NON2^{BE}_1$	$NON2^{BE}_2$	$NON2^{BE}_3$	$NON2^{BE}_4$	$NON2^{BE}_5$
Solution limite	(0,2, 0,2)	89,5(0,974)	0,082(0,978)	0,064(0,978)	0,093(0,973)	0,071(0,972)	0,060(0,957)
	(0,2, 0,4)	158,3(0,959)	0,072(0,963)	0,096(0,963)	0,079(0,958)	0,066(0,958)	0,058(0,940)
	(0,2, 0,6)	135,3(0,940)	0,065(0,941)	0,057(0,941)	0,071(0,941)	0,062(0,939)	0,056(0,922)
	(0,2, 0,8)	57,4(0,930)	0,070(0,938)	0,061(0,935)	0,076(0,928)	0,066(0,928)	0,060(0,908)
	(0,4, 0,4)	153,4(0,961)	0,079(0,920)	0,061(0,913)	0,096(0,956)	0,072(0,949)	0,060(0,911)
	(0,4, 0,6)	82,2(0,955)	0,072(0,893)	0,059(0,883)	0,086(0,951)	0,069(0,940)	0,058(0,874)
	(0,4, 0,8)	51,2(0,933)	0,071(0,862)	0,059(0,849)	0,084(0,926)	0,068(0,917)	0,059(0,846)
	(0,6, 0,6)	175,5(0,946)	0,077(0,820)	0,060(0,781)	0,101(0,943)	0,074(0,921)	0,061(0,823)
	(0,6, 0,8)	159,6(0,924)	0,071(0,728)	0,057(0,657)	0,089(0,913)	0,069(0,880)	0,058(0,737)
	(0,8, 0,8)	72,8(0,920)	0,070(0,572)	0,056(0,330)	0,093(0,900)	0,070(0,842)	0,058(0,607)
Pas de solution limite	(0,2, 0,2)	0,068(0,949)	0,060(0,959)	0,056(0,959)	0,062(0,937)	0,058(0,935)	0,055(0,922)
	(0,2, 0,4)	0,066(0,960)	0,060(0,970)	0,056(0,970)	0,061(0,935)	0,058(0,931)	0,055(0,951)
	(0,2, 0,6)	0,064(0,940)	0,058(0,945)	0,055(0,945)	0,059(0,959)	0,057(0,919)	0,054(0,909)
	(0,2, 0,8)	0,069(0,933)	0,063(0,944)	0,059(0,941)	0,065(0,926)	0,062(0,925)	0,058(0,920)
	(0,4, 0,4)	0,074(0,910)	0,061(0,836)	0,055(0,828)	0,064(0,899)	0,059(0,884)	0,055(0,824)
	(0,4, 0,6)	0,074(0,915)	0,060(0,815)	0,055(0,806)	0,064(0,922)	0,059(0,879)	0,055(0,792)
	(0,4, 0,8)	0,073(0,891)	0,061(0,786)	0,056(0,771)	0,064(0,873)	0,060(0,852)	0,056(0,763)
	(0,6, 0,6)	0,078(0,859)	0,061(0,567)	0,055(0,470)	0,067(0,853)	0,061(0,795)	0,056(0,572)
	(0,6, 0,8)	0,076(0,843)	0,060(0,515)	0,054(0,402)	0,065(0,817)	0,060(0,767)	0,055(0,556)
	(0,8, 0,8)	0,080(0,755)	0,059(0,110)	0,053(0,017)	0,065(0,728)	0,059(0,607)	0,055(0,158)

5. Conclusion

Nous avons étudié l'application de l'analyse bayésienne à des tableaux de contingence à double entrée incomplets en présence de non-réponse non ignorable. Dans cette situation, les estimations du MV se situent souvent sur la solution limite. Ces solutions limites peuvent donner $G^2 > 0$, même dans le cas d'un modèle saturé (Baker et coll. 1992 ; Park et Brown 1994). Autrement dit, G^2 peut ne pas être appropriée comme statistique pour la spécification des modèles. Pour contourner le problème des solutions limites et obtenir une statistique telle que le facteur de Bayes pour spécifier le modèle indépendamment de l'existence d'une solution limite, nous avons proposé des méthodes d'estimation bayésiennes s'appuyant sur cinq lois a priori différentes. Deux d'entre eux sont nouveaux et les trois autres ont été utilisés antérieurement pour analyser un tableau à simple entrée incomplet. Les deux nouvelles lois a priori tiennent compte des profils de réponse différents entre les répondants et les non-répondants.

L'analyse des données révèle que ces deux nouvelles lois a priori sont plus raisonnables en ce sens qu'ils tiennent mieux compte du mécanisme de non-réponse non ignorable et produisent des estimations proches des résultats réels. En outre, dans le cas des trois lois a priori utilisées antérieurement, notre étude par simulation montre que les estimations bayésiennes peuvent avoir une EQM plus grande que celle des estimations du MV pour un tableau de contingence sans solution limite ainsi qu'avec une solution limite, contrairement aux études antérieures. Cependant, quand a lieu une solution limite, les deux nouvelles lois a priori donnent de meilleures résultats que les trois lois a priori antérieures et que les estimations du MV en ce sens qu'ils sont généralement caractérisés par des EQM plus faibles, des biais plus petits et des probabilités de couverture plus proches du niveau de couverture nominal.

Nous avons discuté brièvement des questions de pondération à la section 2.2. Cependant, elles requièrent une discussion plus rigoureuse que celle présentée dans cette section. Notre discussion pourrait être approfondie de manière à inclure non seulement divers facteurs de pondération, mais aussi des biais de réponse et d'autres sources de biais et de variations. Ces questions pourront être développées avec soin ultérieurement dans un article élargi.

Remerciements

La présente étude a été financée par une subvention de l'Université de la Corée (K0822301).

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis.* 2ième Édition. New York : John Wiley & Sons, Inc.

Baker, S.G., et Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

Baker, S.G., Rosenberger, W.F. et Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11, 643-657.

Chen, T. (1972). Mixed-up frequencies and missing data in contingency tables. Mémoire de doctorat non publié, University of Chicago, Dept. of Statistics.

Chen, Q.L., et Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Rapport technique*, Department of Statistics, The Ohio State University.

Clarke, P.S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response. *Biometrical Journal*, 44, 701-717.

Clogg, C.C., Rubin, D.B., Schenker, N. et Schultz, B. (1991). Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

De Heer W. (1999). International response trends of an international survey. *Journal of Official Statistics*, 15, 129-142.

Dempster, A.P., Laird, N.M. et Rubin, D.M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Séries B, 39, 1-38.

Flannelly, K.J., Flannelly, L.T. et McLeod, M.S. Jr. (2000). Reducing undecided voters and other sources of error in election surveys. *International Journal of Market Research*, 42, 231-237.

Fenwick, I, Wiseman, F, Becker, J.F. et Heiman, J.R. (1982). Classifying undecided voters in pre-election polls. *Public Opinion Quarterly*, 46, 383-391.

Forster, J.J., et Smith, R.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society B*, 60, 57-70.

Gelman, A., Carlin, J.P., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Analysis*. 2ième Édition. New York : Chapman and Hall/CRC.

Groves, R.M., et Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.

Kim, T. (1995). Discriminant analysis as a prediction tool for uncommitted voters in pre-election polls. *International Journal of Public Opinion Research*, 7, 110-127.

Lau, R.R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential elections. *Public Opinion Quarterly*, 59, 589-605.

Lavrakas, P.J. (1993). *Telephone Survey Method*: *Sampling, selection, and supervision*. 2ième Édition. Newbury Park, Calif. : Sage.

Little, J.A., et Rubin, D.B. (2002). *Statistical analysis with missing data*. 2ième Édition. New York : John Wiley & Sons, Inc.

Martin, E.A., Traugott, M.W. et Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *The Public Opinion Quarterly*, 69, 342-369.

Michiels, B., et Molenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom drop-out. *Communications in Statistics*: *Theory and Methods*, 26, 65-94.

Molenberghs, G., Kenward, M.G. et Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Applied Statistics*, 50, 15-29.

Monterola, C., Lim, M., Garcia, F. et Saloma, C. (2001). Feasibility of a neural network as classifier of undecided respondents in a public opinion survey. *International Journal of Public Opinion Research*, 14, 222-299.

Myers, D.J., et O'Connor, R.E. (1983). The undecided respondents in mandatory voting settings: A Venezuelan exploration. *The Western Political Quarterly*, 36, 420-433.

Park, T., et Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, 89, 44-52.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54, 1579-1690.

Perry, P. (1979). Certain problem in election survey methodology. *Public Opinion Quarterly*, 43, 312-325.

Potthoff, R.F. (1994). Telephone sampling in epidemiologic research: To reap the benefits, avoid the pitfalls. *American Journal of Epidemiology*, 139, 967-978.

Rubin, D.B., Stern, H.S. et Vehovar, V. (1995). Handling "Don't Know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.

Smith, P.W.F., Skinner, C.J. et Clarke, P.S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intention data. *Applied Statistics*, 48, 563-577.

Smith, T.W. (1984). Non attitudes: A review and evaluation. Dans *Surveying Subjective Phenomena*, (Éds. C.F. Turner et E. Martin), New York : Russell Sage Foundation, 2, 215-255.

Stasny, E.A. (1986). Estimating gross flow using panel data with nonresponse: An example from the Canadian Labor Force survey. *Journal of the American Statistical Association*, 81, 42-47.

Stasny, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating Gross Labor-Force flows. *Journal of Business and Economic Statistics*, 6, 207-219.

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 57-70
Statistique Canada, N° 12-001-X au catalogue

Estimations pour petits domaines bayésiennes hiérarchiques et empiriques de la proportion de personnes sans assurance-maladie dans les groupes de population minoritaires

Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff et Van L. Parsons [1]

Résumé

Le présent article traite de l'estimation pour petits domaines de la proportion de personnes sans assurance-maladie dans divers groupes minoritaires. Les petits domaines sont définis par le croisement de l'âge, du sexe et d'autres caractéristiques démographiques. Des méthodes d'estimation bayésiennes hiérarchiques ainsi qu'empiriques sont appliquées. En outre, des approximations exactes jusqu'à l'ordre deux des erreurs quadratiques moyennes des estimateurs bayésiens empiriques et des estimateurs corrigés du biais de ces erreurs quadratiques moyennes sont fournies. La méthodologie générale est illustrée au moyen d'estimations de la proportion de personnes non assurées pour plusieurs petits domaines de la sous-population asiatique.

Mots-clés : Asiatique ; corrigé du biais ; erreur quadratique moyenne ; exact jusqu'à l'ordre deux.

1. Introduction

La motivation principale de l'étude était l'estimation pour petits domaines de la proportion de personnes sans assurance-maladie dans divers groupes de population minoritaires. Les petits domaines sont construits par croisement de l'âge, du sexe, de la race et de la région auxquels les personnes appartiennent. Les données de la National Health Interview Survey (NHIS) fournissent la réponse binaire au niveau de l'individu (c'est-à-dire possède ou non une assurance-maladie), ainsi que les covariables à ce niveau. Les données peuvent être obtenues à l'adresse http://www.cdc.gov/nchs/nhis.htm. La conception de la NHIS est décrite dans Botman, Moore, Moriarity et Parsons (2000).

Durant une année typique, on sélectionne pour la NHIS un échantillon d'unités de logement. L'ensemble des membres de chaque unité est considéré comme un ménage, et ceux entre lesquels existe une « forte » relation sont considérés comme une famille. (Les unités structurelles sont définies de manière plus explicite au chapitre 5.2 du document sur le recensement à l'adresse www.census.gov/prod/2002pubs/tp63rv.pdf). Chaque année, les données de la NHIS sont recueillies auprès d'environ 40 000 ménages, dont plus de 98 % sont des ménages unifamiliaux, ce qui représente environ 100 000 personnes. Pour les questions de type « familial », par exemple celles sur la couverture par une assurance, tous les adultes à domicile sont invités à participer à l'interview, mais la réponse par personne interposée est également permise. Pour les enfants, les réponses sont fournies par un adulte.

Pour toute année de référence, l'enquête originale contient des données sur plus de 100 000 personnes et sur plus de 800 variables. Pour ces personnes, nous possédons des renseignements sur la variable de réponse principale, c'est-à-dire le fait qu'une personne possède une assurance-maladie ou non. En plus, nous disposons d'information sur les caractéristiques démographiques, telles que l'âge, le sexe, la race, la région, le niveau d'études, la catégorie de revenu, les problèmes de santé, les incapacités (s'il en existe) et de nombreux autres facteurs socioéconomiques.

Pour l'ensemble de la population américaine, les estimations directes pour ces domaines, c'est-à-dire les proportions d'échantillon pondérées, sont assez fiables, parce que la taille d'échantillon est raisonnablement grande pour chaque domaine. Par contre, cela n'est pas nécessairement le cas quand l'analyse est axée sur des sous-populations particulières, tels que les Hispaniques, les Asiatiques et des groupes minoritaires similaires de la collectivité.

Pour une sous-population minoritaire visée, la taille d'échantillon dans les domaines n'est pas toujours très grande. Donc, les estimations directes risquent de ne pas être très fiables, et d'être assorties de grandes erreurs-types et de grands coefficients de variation. Il faut alors recourir à des méthodes d'estimation pour petits domaines, qui produisent des estimations indirectes de ces domaines en se basant sur des modèles implicites ou explicites. Ces modèles établissent un lien entre les domaines et produisent donc habituellement des estimations par domaine d'une plus grande précision par emprunt de données à d'autres domaines.

Nous nous servons de méthodes hiérarchiques bayésiennes (HB) et empiriques bayésiennes (EB) pour obtenir les estimations pour petits domaines, ainsi que les mesures de précision connexes. L'analyse est fondée sur un

1. Malay Ghosh, University of Florida ; Dalho Kim, Kyungpook National University ; Karabi Sinha, University of California at Los Angeles ; Tapabrata Maiti, Michigan State University ; Myron Katzoff, National Center for Health Statistics ; Van L. Parsons, National Center for Health Statistics.

analogue HB du modèle linéaire mixte généralisé (GLMM pour *generalized linear mixed model*) pour obtenir les moyennes et les erreurs-types a posteriori des proportions pour petits domaines de population. La méthode a été proposée par Ghosh, Natarajan, Stroud et Carlin (1998). L'approche EB est fondée sur la théorie des fonctions d'estimation optimales. Nous obtenons les estimateurs EB et les estimateurs approximatifs de l'erreur quadratique moyenne correspondant par une méthode asymptotique analogue à celles de Prasad et Rao (1990) et de Ghosh et Maiti (2004). Alors que celle de Ghosh et Maiti (2004) est fondée sur des données au niveau de la région, la présente approche s'appuie sur des données au niveau de l'unité. Par conséquent, nous devons apporter certaines modifications à la procédure de Ghosh et Maiti (2004) pour élaborer les estimateurs. En outre, comme celle de ces auteurs, la méthodologie générale n'est pas limitée aux données binaires. Elle est applicable à la famille exponentielle naturelle à fonctions de variance quadratiques (Morris 1982, 1983). Le calcul des erreurs quadratiques moyennes des estimations sous le modèle proposé est un peu plus simple que celui de Ghosh et Maiti (2004) pour le cas binaire. En outre, comme ces auteurs, nous utilisons dans notre analyse les poids de sondage ainsi que le modèle pour calculer les estimations pour petits domaines. Donc, dans un certain sens, notre méthode peut être considérée comme une estimation fondée sur un modèle et assistée par le plan de sondage.

Les poids de sondage liés aux unités d'échantillonnage individuelles sont habituellement proportionnels à l'inverse des probabilités de sélection. Ils sont souvent utilisés pour produire des estimations sans biais par rapport au plan. L'exemple classique est le fameux estimateur d'Horvitz-Thompson. Cependant, alors que les estimateurs de ce genre protègent contre l'échec du modèle, ils peuvent entraîner une perte d'efficacité si le modèle hypothétique est vérifié. Par exemple, dans des conditions bayésiennes simples, si les $y_i \mid \theta_i$ sont indépendamment distribués de loi $\mathrm{N}(\theta_i, 1)$, tandis que les θ_i sont indépendants et identiquement distribués de loi $\mathrm{N}(\mu, A)$, $(i = 1, \ldots, n)$, alors l'estimateur de Bayes (moyenne a posteriori) de $\bar{\theta} = n^{-1}\sum_{i=1}^{n}\theta_i$ est $n^{-1}\sum_{i=1}^{n}[(1-B)y_i + B\mu] = (1-B)\bar{y} + B\mu$, où $B = (1+A)^{-1}$. Cet estimateur possède le risque de Bayes $n^{-1}(1-B)$ sous le modèle supposé. Par ailleurs, l'estimateur $\sum_{i=1}^{n} w_i y_i$ de $\bar{\theta}$, avec $\sum_{i=1}^{n} w_i = 1$ possède le risque de Bayes $n^{-1}(1-B) + E[(1-B)\bar{y} + B\mu - \sum_{i=1}^{n} w_i y_i]^2$. Cependant, si le modèle supposé n'est pas vérifié, par exemple, si les θ_i sont indépendants et identiquement distribués de loi $\mathrm{N}(\mu, A)$, $(i = 1, \ldots, n)$, où A s'écarte fortement de A_0, alors l'estimateur $(1-B)\bar{y} + B\mu$ de $\bar{\theta}$ possède un risque de Bayes $n^{-1}(1-B_0) + (B-B_0)^2(\bar{y}-\mu)^2$, $B_0 = (1+A_0)^{-1}$, qui peut être assez bien supérieur au risque de Bayes correspondant de $\sum_{i=1}^{n} w_i y_i$, selon, évidemment, les valeurs de B_0, μ et les w_i.

Le présent article décrit la production des estimations par domaine de la proportion de personnes non couvertes par une assurance pour la population asiatique. Les estimations et les mesures de précision sont basées sur le modèle hiérarchique bayésien, ainsi que sur le modèle empirique bayésien. L'analyse a été effectuée pour chaque année de 1997 à 2000, mais par souci de concision, les résultats ne sont présentés que pour l'année 2000. Nous avons exécuté une analyse semblable pour la population hispanique. Dans ce cas, le nombre de petits domaines était égal à 336. Puisque la méthodologie était la même que pour la population asiatique, faute d'espace, nous n'incluons pas cette analyse dans l'article non plus.

Le groupe asiatique est formellement composé 1) des Chinois, 2) des Philippins, 3) des Indiens d'Asie et 4) d'autres, tels que les Coréens, les Vietnamiens, les Japonais, les Hawaïens, les Samoans, les Guyamiens, etc. Ces personnes sont affectées à des domaines particuliers selon leur âge, leur race, leur sexe et la région d'où elles proviennent. Il existe trois groupes d'âge (de 0 à 17 ans, de 18 à 64 ans et 65 ans et plus), deux sexes, quatre races et quatre régions selon la taille de la région statistique métropolitaine ($<$ 499 999 ; 500 000-999 999 ; 1 000 000-2 499 999 $>$ 2 500 000). Donc, le nombre total de domaines est égal à $3 \times 2 \times 4 \times 4 = 96$. Quand les personnes sont affectées à leurs domaines respectifs, de nombreux domaines ne contiennent que quelques observations. En effet, dans plusieurs d'entre eux, la taille de l'échantillon est égale à 1, tandis que dans un autre, elle est nulle.

Le plan des autres sections est le suivant. À la section 2, nous abordons la sélection des covariables pour la population asiatique. À la section 3, nous discutons de la méthodologie HB générale nécessaire pour obtenir les estimations pour petits domaines et les mesures de précision connexes. À la section 4, nous discutons de l'adéquation du modèle HB proposé. À la section 5, nous discutons d'une méthode EB alternative et nous trouvons des erreurs quadratiques moyennes (EQM) correctes jusqu'à l'ordre deux (à préciser plus loin) des estimateurs EB proposés, ainsi qu'une approximation correcte jusqu'à l'ordre deux de ces EQM. À la section 6, nous trouvons les estimations pour petits domaines et les mesures de précision correspondantes pour la sous-population asiatique en 2000 en utilisant les méthodes HB et EB, et nous les comparons aux estimations directes. Enfin, à la section 7, nous tirons certaines conclusions.

2. Choix des covariables

Comme nous l'avons mentionné dans l'introduction, le nombre de covariables est supérieur à 800. Les inclure toutes dans le modèle initial est impossible et inutile. Nous sommes partis de ce que nous considérions comme un ensemble sensé de six covariables et avons suivi un processus de sélection entièrement séquentiel (avec un seuil de signification de 0,05) pour finalement arriver au meilleur modèle possible.

Les six covariables que nous avons considérées étaient 1) l'état matrimonial légal, 2) la taille de la famille, 3) le niveau d'études, 4) les gains totaux l'année précédente, 5) le revenu familial total et 6) la situation d'emploi à temps plein.

Après la procédure séquentielle, notre modèle final contenait, en plus du terme d'ordonnée à l'origine, la taille de la famille, le niveau d'études et le revenu familial total comme covariables.

Puisque la procédure SURVEYREG de SAS Version 8 ajuste des modèles de régression linéaire et produit des tests d'hypothèse et des estimations pour les données d'enquête, nous l'avons utilisée pour choisir nos covariables. À l'époque où nous avons effectué nos travaux de recherche, la régression logistique pour la sélection des covariables n'était pas disponible. Il convient de souligner que SURVEYREG tient compte de la mise en grappes et de la pondération inégale, et produit des erreurs-types qui tiennent compte correctement des plans de sondage complexes.

3. Analyse hiérarchique bayésienne

Un modèle général de famille exponentielle à un paramètre est donné par

$$f(y_{ij}|\theta_{ij}) = \exp[\xi_{ij}\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\}]h(y_{ij};\xi_{ij}), \qquad (3.1)$$

$j = 1, \ldots, n_i, i = 1, \ldots, k$. Ici, y_{ij} est la réponse de la j^{e} unité dans le i^{e} petit domaine, tandis que ξ_{ij}, les « dits » paramètres de surdispersion, sont supposés connus et pris égaux à 1 sans perte de généralité, parce que l'on peut autrement travailler avec les paramètres transformés, $\zeta_{ij} = \xi_{ij}\theta_{ij}$. La fonction h est une fonction positive qui dépend des y_{ij}, mais non des θ_{ij}. Si y_{ij} est binaire avec une probabilité de succès p_{ij}, alors $\theta_{ij} = \text{logit}(p_{ij})$. Dans notre exemple, y_{ij}, la réponse de la j^{e} personne dans le i^{e} petit domaine, est vaut 1 ou 0 selon que la personne ne possède pas ou possède une assurance-maladie. Nous souhaitons estimer $\bar{\mu}_{iw} = \sum_{j=1}^{n_i} w_{ij} p_{ij}$, les moyennes de domaine pondérées des proportions de population. Dans ce cas, l'estimateur direct de $\bar{\mu}_{iw}$ est $\sum_{j=1}^{n_i} w_{ij} y_{ij}$. Ces estimateurs directs sont habituellement assortis d'erreurs-types et de coefficients de variation importants. Nous supposons que les poids de sondage w_{ij} sont connus et nous les normalisons de manière que $\sum_{j=1}^{n_i} w_{ij} = 1$ pour tout $i = 1, \ldots, k$. Il faut toutefois reconnaître qu'en pratique, les w_{ij} ne sont souvent que des estimations, par exemple en tenant compte de la poststratification et de la non-réponse. Cependant, les utilisateurs secondaires des données ne connaissent pas les mécanismes réels utilisés pour produire ces poids et nous devons supposer que ceux-ci sont connus. Un autre exemple important, qui n'est pas considéré spécifiquement dans le présent article, est $y_{ij} \sim \text{Poisson}(\lambda_{ij})$, de sorte que $\theta_{ij} = \log(\lambda_{ij})$. Nous pouvons utiliser ici un modèle de Poisson fondé sur les dénombrements de personnes non assurées au niveau du domaine. La difficulté tient au fait que, dans le présent exemple, nous possédons des covariables au niveau individuel et *non* au niveau du domaine. La modélisation des dénombrements en se servant de covariables au niveau du domaine n'est pas possible dans cette situation.

À la présente section, nous discutons de la façon d'exécuter l'analyse pour le modèle hiérarchique bayésien général quand nous voulons estimer $\mu_{ij} = E(y_{ij}|\theta_{ij}) = \psi'(\theta_{ij})$. Puisque $\psi''(\theta_{ij}) = \text{var}(y_{ij}|\theta_{ij})$, μ_{ij} est une fonction bijective de θ_{ij}. En particulier, $\mu_{ij} = p_{ij}$ dans le cas binaire. Nous considérerons des applications particulières à la section 5.

L'étape suivante du modèle est

$$\theta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{b} + u_i \; ; \; j = 1, \ldots, n_i, i = 1, \ldots, k, \qquad (3.2)$$

où les $\boldsymbol{x}_{ij}$ sont les vecteurs de plan d'expérience, ou de manière équivalente les vecteurs de prédicteurs, $\boldsymbol{b}$ est le vecteur des paramètres de régression, et les u_i sont les effets aléatoires. Nous supposons que les u_i sont iid de loi $\text{N}(0, \sigma_u^2)$. En outre, soit $\boldsymbol{X}^T = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \ldots, \boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kn_k})$ et supposons que $\boldsymbol{X}$ est une matrice de plein rang.

Enfin, nous supposons que $\boldsymbol{b}$ et σ_u^2 sont mutuellement indépendants, où est appliquée à $\boldsymbol{b}$ la loi a priori uniforme incorrecte, R^P, et σ_u^2 suit une loi gamma inverse avec les paramètres $c/2.d/2$. c'est-à-dire $\pi(\sigma_u^2) \propto \exp(-c/2\sigma_u^2)(\sigma_u^2)^{-d/2-1}, c > 0$.

Soit $\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{k1}, \ldots, y_{kn_k})^T$, et $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1n_1}, \ldots, \theta_{k1}, \ldots, \theta_{kn_k})^T$. Alors, la loi a posteriori conjointe est donnée par

$$\pi(\boldsymbol{\theta}, \boldsymbol{b}, \sigma_u^2 | \boldsymbol{y}) \propto \prod_{i=1}^{k}\prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}) \times (\sigma_u^2)^{-k/2} \exp\left[-\frac{1}{2\sigma_u^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\theta_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{b})^2\right] \times (\sigma_u^2)^{-d/2-1}\exp\left(-\frac{c}{2\sigma_u^2}\right). \qquad (3.3)$$

Il s'agit d'une analyse bayésienne non conjuguée qui n'est pas exécutable de manière analytique. Au lieu de cela, nous utilisons la méthode d'intégration numérique de Monté-Carlo par chaînes de Markov (MCCM). En particulier, nous employons l'échantillonneur de Gibbs. La méthode MCCM générale est exposée dans de nombreuses publications. Une référence pratique est Tanner (1996, chapitre 6).

Pour exécuter l'échantillonneur de Gibbs, nous devons trouver les conditionnelles complètes de θ_{ij}, $\boldsymbol{b}$ et σ_u^2. Elles sont données par

$$\sigma_u^2 \mid \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y} \sim \mathrm{IG}\left(\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\theta_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{b})^2 + c}{2}, \frac{k+d}{2} \right);$$

$$\boldsymbol{b} \mid \boldsymbol{\theta}, \sigma_u^2, \boldsymbol{y} \sim N((\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}, \sigma_u^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1});$$

$$\theta_{ij} \mid \boldsymbol{b}, \sigma_u^2, \boldsymbol{y} \sim f(y_{ij} \mid \theta_{ij}) \exp\left[-\frac{1}{2\sigma_u^2} (\theta_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{b})^2 \right].$$

Notre analyse des données est fondée sur la production d'échantillons à partir des conditionnelles susmentionnées, formulées pour le cas binaire. La production d'échantillons à partir des lois conditionnelles de σ_u^2 et $\boldsymbol{b}$ est standard, mais elle ne l'est pas pour θ_{ij}, et requiert l'utilisation de l'algorithme de Métropolies-Hastings. Pour une discussion de cet algorithme, nous renvoyons de nouveau le lecteur à Tanner (1996).

Si $\hat{\mu}_{ij}^{(r)}$ désigne la valeur échantillonnée de μ_{ij} produite à partir du r^{e} tirage, et que le nombre de tirages est R, alors l'estimation de Monté-Carlo de $E(\mu_{ij} \mid \boldsymbol{y})$ est $R^{-1} \sum_{r=1}^{R} \hat{\mu}_{ij}^{(r)}$. De même, l'estimation de Monté-Carlo de $\mathrm{var}(\mu_{ij} \mid \boldsymbol{y})$ est $R^{-1} \sum_{r=1}^{R} (\hat{\mu}_{ij}^{(r)})^2 - (R^{-1} \sum_{r=1}^{R} \hat{\mu}_{ij}^{(r)})^2$. Enfin, l'estimation de Monté-Carlo de $\mathrm{cov}(\mu_{ij}, \mu_{i'j'} \mid \boldsymbol{y})$ est donnée par $R^{-1} \sum_{r=1}^{R} (\hat{\mu}_{ij}^{(r)} \hat{\mu}_{i'j'}^{(r)}) - (R^{-1} \sum_{r=1}^{R} \hat{\mu}_{ij}^{(r)})(R^{-1} \sum_{r=1}^{R} \hat{\mu}_{i'j'}^{(r)})$. D'après ces calculs, nous trouvons maintenant immédiatement $E[\bar{\mu}_{iw} \mid \boldsymbol{y}] = \sum_{j=1}^{n_i} w_{ij} E(\mu_{ij} \mid \boldsymbol{y})$ et $V[\bar{\mu}_{iw} \mid \boldsymbol{y}] = \sum_{j=1}^{n_i} w_{ij}^2 V(\mu_{ij} \mid \boldsymbol{y}) + \sum_{1 \le j \ne j' \le n_i} w_{ij} w_{ij'} \mathrm{Cov}(\mu_{ij}, \mu_{ij'} \mid \boldsymbol{y})$. Par contre, l'estimateur sans biais direct de $\bar{\mu}_{iw}$ est donné par $\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij}$. Cependant, comme nous l'avons mentionné plus haut, pour nombre de ces domaines, les tailles d'échantillon sont si petites que ces estimateurs sans biais sont assortis d'erreurs-types et de coefficients de variation de grande taille.

4. Estimation empirique bayésienne

De nouveau, désignons par y_{ij} la réponse de la j^{e} unité dans le i^{e} petit domaine $(j = 1, \ldots, n_i \, ; i = 1, \ldots, k)$. En outre, nous posons que le modèle de famille exponentielle pour les y_{ij} est donné par (3.1), mais nous supposons aussi que les y_{ij} possèdent une fonction de probabilité ou une densité de probabilité appartenant à la classe des familles exponentielles naturelles à fonctions de variance quadratiques (FEN-FVQ). Rappelons que $\mu_{ij} = E(y_{ij} \mid \theta_{ij}) = \psi'(\theta_{ij})$. Avec la structure de fonction de variance quadratique, $\mathrm{Var}(y_{ij} \mid \theta_{ij}) = Q(\mu_{ij}) = v_0 + v_1 \mu_{ij} + v_2 \mu_{ij}^2$, où v_0, v_1 et v_2 ne sont pas simultanément nuls. Morris (1982, 1983) a caractérisé les lois appartenant à la famille FEN-FVQ. Celle-ci comprend les six lois de base, à savoir i) Bernoulli, ii) Poisson, iii) normale de variance connue, iv) géométrique, v) exponentielle, vi) hyperbolique sécante, ainsi que leurs convolutions. De cette façon, les lois binomiale, binomiale négative et gamma appartiennent aussi à cette famille. Pour la loi de Bernoulli, $v_0 = 0, v_1 = 1$ et $v_2 = -1$. Pour la loi de Poisson, $v_0 = v_2 = 0$ et $v_1 = 1$. Pour la loi normale de variance connue, $\sigma^2, \xi_{ij} = \sigma^{-2}, v_0 = 1$ et $v_1 = v_2 = 0$. De nouveau, nous supposerons sans perte de généralité que $\xi_{ij} = 1$.

Nous proposons à la présente section des estimateurs EB des moyennes de petit domaine. Pour cela, nous partons de la famille FEN-FVQ de lois ainsi que d'un prior conjugué pour le paramètre canonique du modèle exponentiel. Ensemble, ils constituent une famille FEN-FVQ sur-dispersée des lois. En particulier, nous considérons le prior conjugué avec la densité de probabilité

$$\pi(\theta_{ij}) = \exp[\lambda \{ m_{ij} \theta_{ij} - \psi(\theta_{ij}) \}] C(\lambda, m_{ij}) \qquad (4.1)$$

pour θ_{ij}, où $m_{ij} = g(\boldsymbol{x}_{ij}^T \boldsymbol{b})$, $j = 1, \ldots, n_i$; $i = 1, \ldots, k$. Ici, $\boldsymbol{x}_{ij}$ est le vecteur de plan associé à la j^{e} unité dans le i^{e} petit domaine, et g est la fonction lien. Alors (Morris 1983),

$$E(\mu_{ij}) = m_{ij} \, ; \mathrm{var}(\mu_{ij}) = Q(m_{ij})/(\lambda - v_2), \qquad (4.2)$$

où nous supposons que $\lambda > \max(0, v_2)$. Puisque $\mathrm{var}(\mu_{ij})$ est strictement décroissante en λ, nous pouvons interpréter ce dernier comme étant le paramètre de précision.

Nous obtenons d'abord l'estimateur de Bayes de μ_{ij}. Il est donné par (Morris 1983)

$$E(\mu_{ij} \mid y_{ij}) = \frac{1}{\lambda + 1} y_{ij} + \frac{\lambda}{\lambda + 1} m_{ij}(\boldsymbol{b}).$$

Cette expression peut également être considérée comme le meilleur prédicteur linéaire sans biais (BLUP) de μ_{ij}. Pour le montrer, nous calculons

$$E(y_{ij}) = E(\mu_{ij}) = m_{ij} \, ; \mathrm{cov}(y_{ij}, \mu_{ij}) = \mathrm{var}(\mu_{ij})$$

$$= Q(m_{ij})/(\lambda - v_2) \, ; \mathrm{var}(y_{ij}) = \frac{\lambda + 1}{\lambda - v_2} Q(m_{ij}).$$

D'où le BLUP de μ_{ij} est donné par

$$m_{ij}(\boldsymbol{b}) + \frac{\mathrm{cov}(y_{ij}, \mu_{ij})}{\mathrm{var}(y_{ij})} (y_{ij} - m_{ij}(\boldsymbol{b}))$$

$$= \frac{1}{\lambda + 1} y_{ij} + \frac{\lambda}{\lambda + 1} m_{ij}(\boldsymbol{b}). \qquad (4.3)$$

Donc, l'estimateur de Bayes de $\bar{\mu}_{iw} = \sum_{j=1}^{n_i} w_{ij}\mu_{ij}$ est donné par $\sum_{j=1}^{n_i} w_{ij} E(\mu_{ij} \mid y_{ij})$.

Cependant, en pratique, $\boldsymbol{b}$ et λ sont inconnus et doivent être estimés d'après les lois marginales des y_{ij}. Malheureusement, à part la loi normale, ces lois marginales sont assez compliquées et trouver l'EMV pour les vraisemblances marginales peut devenir une tâche assez redoutable. Nous calculons donc plutôt les estimations basées sur certaines équations d'estimation optimales sans biais (Godambe et Thompson 1989) qui requièrent uniquement l'évaluation des quatre premiers moments de ces lois marginales. Pour cela, nous partons des fonctions d'estimation sans biais élémentaires $ig_{1ij} = y_{ij} - m_{ij}$ et $g_{2ij} = (y_{ij} - m_{ij})^2 - (\lambda+1)/(\lambda - v_2)V(m_{ij})$. Afin de construire les équations d'estimation optimales, posons que

$$\boldsymbol{D}_{ij}^T = \begin{bmatrix} -E\left(\dfrac{\partial g_{1ij}}{\partial \boldsymbol{b}}\right) & -E\left(\dfrac{\partial g_{2ij}}{\partial \boldsymbol{b}}\right) \\ -E\left(\dfrac{\partial g_{1ij}}{\partial \lambda}\right) & -E\left(\dfrac{\partial g_{2ij}}{\partial \lambda}\right) \end{bmatrix}.$$

En outre, posons que

$$\boldsymbol{\Sigma}_{ij} = \begin{bmatrix} \mu_{2ij} & \mu_{3ij} \\ \mu_{3ij} & \mu_{4ij} - \mu_{2ij}^2 \end{bmatrix},$$

où $\mu_{rij} = E(y_{ij} - m_{ij})^r$ est le r^{e} moment central de y_{ij} basé sur sa loi marginale. Les équations d'estimation optimales sont alors données par $\sum_{i=1}^{k}\sum_{j=1}^{n_i} \boldsymbol{D}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \boldsymbol{g}_{ij} = \boldsymbol{0}$, où $\boldsymbol{g}_{ij} = (g_{1ij}\ g_{2ij})^T$. Nous obtenons les estimations de $\boldsymbol{b}$ et λ (si elles existent) en résolvant ces équations. Nous trouvons les solutions de ces équations à l'aide de l'algorithme de Nelder-Meade.

Malheureusement, la méthode susmentionnée échoue pour des données binaires. Dans ce cas, $v_2 = -1$, de sorte que $\text{var}(y_{ij})$ ne dépend pas de λ. En effet, les lois bêta binaires marginales des y_{ij} sont non identifiables en λ. Un simple moyen de le vérifier est que, si $y \mid p \sim \text{Bin}(1, p)$, et $p \sim \text{Beta}(\lambda m, \lambda(1-m))$, alors $E(y) = E(p) = m$, et une loi binaire est entièrement caractérisé par sa moyenne. Le problème ne se pose pas pour une loi binomiale (n, p) avec $n \geq 2$, puisque, avec la même loi marginale pour p, la fonction génératrice des moments de la loi marginale de la binomiale y est $E[(p\exp(t) + 1 - p)^n]$ qui dépend de λ.

Pour y_{ij} binaire, $\partial g_{1ij}/\partial\lambda = \partial g_{2ij}/\partial\lambda = 0$ de sorte que le deuxième élément du vecteur $\sum_{i=1}^{k}\sum_{j=1}^{n_i} \boldsymbol{D}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \boldsymbol{g}_{ij}$ soit nul. Conséquemment, l'approche des équations d'estimation proposées ne permet pas d'estimer λ. Les données de base considérées dans notre application sont binaires, ce qui nécessite la modification de la procédure proposée.

Nous avons donc considéré la fonction d'estimation optimale (pour λ connu)

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} [(y_{ij} - m_{ij})/(\text{var}(y_{ij})]\frac{\partial m_{ij}}{\partial \boldsymbol{b}} = \boldsymbol{0},$$

puisque $\partial g_{1ij}/\partial \boldsymbol{b} = -\partial m_{ij}/\partial \boldsymbol{b}$. Il convient de souligner également que, dans ce cas, $\text{var}(y_{ij}) = V(m_{ij}) = m_{ij}(1-m_{ij})$. En outre, avec la représentation logistique, $m_{ij}(\boldsymbol{b}) = \exp(\boldsymbol{x}_{ij}^T\boldsymbol{b})/[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{b})]$, nous obtenons $\partial m_{ij}/\partial \boldsymbol{b} = -m_{ij}(1-m_{ij})\boldsymbol{x}_{ij}$. Donc $\boldsymbol{b}$ est estimé à partir des équations d'estimation $\sum_{i=1}^{k}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} m_{ij}$. Si nous désignons cet estimateur par $\hat{\boldsymbol{b}}$, l'estimateur EB de μ_{ij} est donné par

$$\hat{\mu}_{ij}^{\text{EB}} = \frac{1}{\lambda+1} y_{ij} + \frac{\lambda}{\lambda+1} m_{ij}(\hat{\boldsymbol{b}}). \tag{4.4}$$

Par conséquent, l'estimateur EB de $\bar{\mu}_{iw}$ est $\hat{\bar{\mu}}_{iw}^{\text{EB}} = \sum_{j=1}^{n_i} w_{ij}\hat{\mu}_{ij}^{\text{EB}}$.

La méthode décrite ci-dessus repose sur l'hypothèse que λ est connu. Nous pouvons trouver des estimations de μ_{ij} pour divers choix de λ. Dans le présent article, nous avons essayé $\lambda = 0{,}1,\ 0{,}5$ et 1, et nous avons comparé les estimations ainsi obtenues aux estimations HB correspondantes.

Ensuite, à la présente section, nous trouvons les erreurs quadratiques moyennes (EQM), ainsi que les EQM estimées de $\hat{\bar{\mu}}_{iw}^{\text{EB}}$ en supposant que λ est connu. Nous énonçons deux théorèmes dans la présente section. Mais avant cela, il faut que nous précisions certains éléments de notation. Soit $\boldsymbol{M} = \text{Diag}(m_{11},\ldots,m_{1n_1},\ldots,m_{k1},\ldots,m_{kn_k})$ et $\boldsymbol{\Sigma}(\boldsymbol{b}) = \boldsymbol{X}^T\boldsymbol{M}(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{X} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} m_{ij}(1-m_{ij})\,\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T$. En outre, soit $n_T = \sum_{i=1}^{k} n_i$. Nous supposons que $1 \leq n_i \leq C$ pour chaque i, de sorte que $n_T = O_e(k)$, où O_e désigne l'ordre exact. Les deux théorèmes sont donnés ci-après.

Théorème 1. Supposons que $\boldsymbol{\Sigma}(\boldsymbol{b}) = O_e(k)$, c'est-à-dire que chaque élément de $\boldsymbol{\Sigma}(\boldsymbol{b})$ a pour borne inférieure une certaine constante C_1, et pour borne supérieure, une certaine constante C_2, où $0 < C_1 < C_2 < \infty$. Alors, une expression appropriée pour $\text{EQM}(\hat{\bar{\mu}}_{iw}^{\text{EB}})$ correcte jusqu'à l'ordre $O(k^{-1})$ est donnée par

$$\begin{aligned}\text{EQM}(\hat{\bar{\mu}}_{iw}^{\text{EB}}) \doteq\ & \frac{\lambda}{(\lambda+1)^2}\sum_{j=1}^{n_i} w_{ij}^2 m_{ij}(b)(1-m_{ij}(b)) \\ & + \frac{\lambda^2}{(\lambda+1)^2}\left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(b)(1-m_{ij}(b))\boldsymbol{x}_{ij}\right]^T \\ & \times \boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(b)(1-m_{ij}(b))\boldsymbol{x}_{ij}\right]. \end{aligned} \tag{4.5}$$

Théorème 2. Supposons que $\boldsymbol{\Sigma}(\boldsymbol{b}) = O_e(k)$. Alors, l'approximation suivante de $\text{EQM}(\hat{\bar{\mu}}^{\text{EB}})$ est correcte jusqu'à l'ordre $O(k^{-1})$.

$$\frac{\lambda}{(1+\lambda)^2}\sum_{j=1}^{n_i}\Big[\; m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}})) - (1-2m_{ij}(\hat{\boldsymbol{b}}))m_{ij}(\hat{\boldsymbol{b}})$$
$$(1-m_{ij}(\hat{\boldsymbol{b}}))\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\begin{pmatrix} tr(\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\boldsymbol{K}_1(\hat{\boldsymbol{b}})) \\ \vdots \\ tr(\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\boldsymbol{K}_p(\hat{\boldsymbol{b}})) \end{pmatrix}$$
$$+ m_{ij}^2(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))^2\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\boldsymbol{x}_{ij}\;\Big]$$
$$+\frac{\lambda^2}{(\lambda+1)^2}\left[\sum_{j=1}^{n_i} w_{ij}m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))\boldsymbol{x}_{ij}\right]^T$$
$$\times\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\left[\sum_{j=1}^{n_i} w_{ij}m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))\boldsymbol{x}_{ij}\right]. \qquad (4.6)$$

Les preuves de ces théorèmes sont présentées en annexe. À la section suivante, nous appliquerons ces résultats pour trouver des estimations approximatives des EQM des estimateurs EB. Cependant, avant cela, il convient de souligner le point suivant.

Si nous désignons le coefficient de $\lambda/(1+\lambda)^2$ par $B_i(\hat{b})$ et le coefficient de $\lambda^2/(1+\lambda)^2$ par $C_i(\hat{b})$ dans le théorème 2, alors en remarquant que $B_i(\hat{b}) = O(1)$ et $C_i(\hat{b}) = O(k^{-1})$, pour k grand, EQM($\hat{\mu}$) est maximisé à $\hat{\lambda} = (B_i(\hat{b}))/(B_i(\hat{b}) - 2C_i(\hat{b}))$, qui est habituellement très proche de 1. La distribution a priori résultante avec $\hat{\lambda}$ remplaçant λ est la distribution a priori approximative adaptable aux données le moins favorable. Dans l'exemple que nous considérerons, ce λ estimé s'avère être égal à 1,003, ce qui confirme l'observation qui précède.

5. Estimations pour petits domaines pour les Asiatiques

Nous commençons par décrire comment les petits domaines sont construits. Considérons le quadruplet (k_1, k_2, k_3, k_4), où $k_1 = 1, 2, 3$ ou 4 selon que la personne est chinoise, philippine, indienne d'Asie ou insulaire. Puis, $k_2 = 1$ ou 2, selon que la personne est un homme ou une femme. Puis, $k_3 = 1, 2$ ou 3, selon que la personne appartient au groupe des 0 à 17 ans, des 18 à 64 ans ou des 65 ans et plus. Enfin, $k_4 = 1, 2, 3$ ou 4, selon que la personne appartient à une région statistique métropolitaine (RSM) de taille [499 999, 500 000-999 999, 1 000 000-2 499 999 ou ≥ 2 500 000. Un petit domaine est maintenant numéroté par la formule $24(k_1-1)+12(k_2-1)+4(k_3-1)+k_4$ correspondant au quadruplet (k_1, k_2, k_3, k_4). Par exemple, le petit domaine constitué de femmes philippines de 18 à 64 ans vivant dans une RSM de taille 500 000 à 999 999 porte le numéro 42.

Les données de base sont $y_{ij} = 1$ ou 0 si la j^{e} personne dans le i^{e} petit domaine ne possède pas (possède) une assurance-maladie ;

$\tilde{w}_{ij}$ = le poids d'échantillonnage appliqué à la j^{e} unité dans le i^{e} petit domaine ;

w_{ij} = $\tilde{w}_{ij}/\sum_{j=1}^{n_i}\tilde{w}_{ij}$ de sorte que $\sum_{j=1}^{n_i} w_{ij} = 1$ pour chaque i ;

x_{ij1} = la taille de la famille de la j^{e} unité dans le i^{e} petit domaine ;

x_{ij2} = le niveau d'études de la j^{e} unité dans le i^{e} petit domaine ;

x_{ij3} = le revenu familial total de la j^{e} unité dans le i^{e} petit domaine.

Soit $p_{ij} = E(y_{ij})$. Pour l'analyse HB, nous modélisons

$$\theta_{ij} = \text{logit}(p_{ij}) = b_0 + b_1x_{ij1} + b_2x_{ij2} + b_3x_{ij3} + u_i,$$
$$j = 1, \ldots, n_i,\ i = 1, \ldots, 96.$$

Les estimations par domaine directes sont données par $\hat{p}_{iw} = \sum_{j=1}^{n_i} w_{ij}y_{ij}$. Les estimations hiérarchiques bayésiennes correspondantes sont données par $\hat{p}_{iw}^{\text{HB}} = \sum_{j=1}^{n_i} w_{ij}E(p_{ij} \mid \boldsymbol{y})$. Nous utilisons la méthode MCMC décrite à la section 2 pour obtenir ces estimations. Dans le tableau elles sont appelées HB. Les erreurs-types a posteriori connexes sont appelées é.-t. (HB). Dans notre hyperprior, nous considérons : $c = 0,2,\ 0,02,\ 0,002$; $d = 0,2, 0,02, 0,002$. Les résultats sont très insensibles au choix des hyperpriors et ne sont présentés que pour $c = d = 0,02$. En outre, nous avons des estimateurs EB pour différents choix du paramètre λ. Les résultats sont présentés pour $\lambda = 0,1,\ 0,5$ et 1.

Le tableau 1 donne les diverses estimations des proportions d'Asiatiques non assurés et les erreurs-types connexes pour les divers petits domaines pour l'année 2000. Le domaine 2 est exclu parce que la taille d'échantillon est nulle. Il correspond aux hommes philippins de 0 à 17 ans appartenant à une RSM de taille de 500 000 à 999 999. Les mesures de précision (é.-t. a posteriori) associées aux estimations HB sont désignées par é.-t. (HB) et sont données par la formule é.-t.^{2} (HB) = var($\sum_{j=1}^{n_i} w_{ij}p_{ij} \mid \boldsymbol{y}$). Un des avantages des estimations HB ou EB est que, pour les domaines dont la taille d'échantillon est très petite, les estimations directes de la proportion de personnes non assurées sont souvent nulles, tandis que les estimations bayésiennes sont faibles mais non nulles. Nous avons choisi de ne pas regrouper les estimations directes pour les domaines dont les tailles d'échantillon sont très faibles. Les covariables au niveau de l'unité étaient assez différentes et il n'y avait aucun moyen significatif de les combiner. Nous notons aussi que, quand $\lambda = 0,5$, c'est-à-dire quand les estimations directes et synthétiques ont un ratio de pondération de 1/2, les estimations EB et HB sont très proches.

Tableau 1
Estimations pour petits domaines des proportions d'Asiatiques non assurés : année 2000

Domaine	n_i	Directe	Moyenne 1997 à 1999	HB	é.-t. (HB)	EB $\lambda = 0,5$	EB $\lambda = 1$	é.-t. (EB) $\lambda = 0,5$	é.-t. (EB) $\lambda = 1$
1	10	0,126	0,034	0,133	0,043	0,148	0,158	0,057	0,060
2	0	-	0,085	-	-	-	-	-	-
3	24	0,063	0,016	0,074	0,025	0,076	0,082	0,037	0,039
4	28	0,146	0,105	0,150	0,027	0,163	0,171	0,041	0,043
5	20	0,138	0,265	0,143	0,032	0,153	0,160	0,043	0,046
6	17	0,112	0,124	0,120	0,032	0,134	0,144	0,019	0,021
7	78	0,097	0,100	0,104	0,015	0,107	0,112	0,022	0,024
8	66	0,274	0,229	0,253	0,023	0,240	0,224	0,072	0,076
9	5	0,173	0,000	0,164	0,061	0,160	0,154	0,078	0,082
10	6	0,000	0,000	0,033	0,051	0,082	0,123	0,070	0,074
11	7	0,000	0,084	0,032	0,047	0,090	0,134	0,054	0,057
12	11	0,335	0,000	0,302	0,056	0,275	0,245	0,060	0,064
13	7	0,134	0,061	0,134	0,045	0,130	0,128	0,103	0,110
14	2	0,000	0,151	0,020	0,064	0,026	0,039	0,031	0,033
15	27	0,000	0,104	0,023	0,023	0,035	0,052	0,032	0,034
16	29	0,113	0,191	0,119	0,024	0,123	0,127	0,033	0,035
17	27	0,120	0,223	0,127	0,025	0,141	0,152	0,044	0,047
18	14	0,000	0,106	0,024	0,030	0,041	0,062	0,019	0,021
19	77	0,131	0,111	0,133	0,015	0,133	0,134	0,021	0,023
20	75	0,223	0,222	0,213	0,018	0,207	0,200	0,089	0,095
21	3	0,000	0,000	0,022	0,056	0,028	0,043	0,070	0,074
22	6	0,000	0,184	0,026	0,045	0,052	0,079	0,071	0,075
23	8	0,000	0,022	0,037	0,050	0,108	0,162	0,063	0,067
24	9	0,000	0,000	0,029	0,042	0,062	0,093	0,052	0,055
25	10	0,000	0,083	0,023	0,034	0,031	0,046	0,061	0,065
26	6	0,000	0,018	0,020	0,039	0,029	0,044	0,031	0,033
27	32	0,098	0,041	0,105	0,023	0,108	0,114	0,035	0,037
28	23	0,000	0,092	0,024	0,025	0,037	0,055	0,032	0,034
29	25	0,187	0,211	0,173	0,030	0,151	0,134	0,035	0,037
30	23	0,227	0,076	0,210	0,032	0,188	0,169	0,021	0,022
31	71	0,118	0,059	0,123	0,016	0,125	0,128	0,024	0,026
32	50	0,109	0,156	0,113	0,019	0,112	0,113	0,113	0,120
33	2	0,000	0,000	0,024	0,071	0,037	0,055	0,115	0,122
34	2	0,000	0,000	0,026	0,073	0,047	0,070	0,058	0,061
35	8	0,108	0,000	0,113	0,042	0,112	0,114	0,067	0,071
36	7	0,000	0,000	0,030	0,045	0,065	0,098	0,051	0,054
37	9	0,062	0,197	0,069	0,035	0,062	0,063	0,036	0,038
38	17	0,000	0,019	0,019	0,024	0,023	0,034	0,037	0,040
39	24	0,117	0,022	0,124	0,028	0,134	0,142	0,040	0,043
40	20	0,000	0,070	0,028	0,029	0,052	0,078	0,025	0,027
41	50	0,163	0,145	0,160	0,020	0,156	0,153	0,027	0,029
42	38	0,141	0,114	0,139	0,021	0,133	0,130	0,020	0,022
43	76	0,104	0,090	0,112	0,016	0,120	0,128	0,020	0,022
44	73	0,142	0,149	0,142	0,016	0,139	0,137	0,119	0,127
45	2	0,000	0,000	0,027	0,076	0,051	0,076	0,090	0,095
46	3	0,000	0,052	0,021	0,056	0,023	0,035	0,052	0,055
47	10	0,000	0,072	0,024	0,034	0,044	0,066	0,068	0,072
48	7	0,000	0,172	0,029	0,045	0,068	0,102	0,051	0,054
49	10	0,087	0,364	0,095	0,037	0,099	0,105	0,078	0,083
50	5	0,000	0,000	0,027	0,050	0,053	0,080	0,032	0,034
51	23	0,038	0,092	0,053	0,023	0,056	0,066	0,037	0,039
52	21	0,243	0,195	0,223	0,037	0,198	0,176	0,030	0,032
53	31	0,114	0,184	0,120	0,022	0,121	0,124	0,040	0,042

Tableau 1 (suite)
Estimations pour petits domaines des proportions d'Asiatiques non assurés : année 2000

Domaine	n_i	Directe	Moyenne 1997-1999	HB	é.-t. (HB)	EB $\lambda = 0{,}5$	EB $\lambda = 1$	é.-t. (EB) $\lambda = 0{,}5$	é.-t. (EB) $\lambda = 1$
54	18	0,202	0,169	0,195	0,031	0,188	0,182	0,019	0,020
55	74	0,094	0,115	0,102	0,015	0,102	0,106	0,019	0,020
60	8	0,112	0,116	0,120	0,044	0,132	0,143	0,059	0,063
63	33	0,055	0,093	0,069	0,020	0,073	0,082	0,028	0,030
64	28	0,105	0,275	0,112	0,024	0,115	0,120	0,032	0,034
65	33	0,126	0,133	0,129	0,021	0,126	0,126	0,029	0,031
70	2	0,361	0,000	0,331	0,098	0,299	0,268	0,119	0,126
71	4	0,000	0,091	0,023	0,050	0,032	0,048	0,077	0,082
72	2	0,000	0,182	0,045	0,101	0,157	0,236	0,155	0,165
73	45	0,271	0,144	0,256	0,026	0,256	0,249	0,028	0,030
74	10	0,000	0,044	0,024	0,034	0,034	0,051	0,051	0,055
75	83	0,149	0,097	0,150	0,016	0,160	0,166	0,020	0,021
76	59	0,113	0,205	0,120	0,018	0,128	0,136	0,023	0,024
77	68	0,338	0,224	0,313	0,025	0,302	0,284	0,023	0,024
78	39	0,098	0,138	0,103	0,020	0,102	0,104	0,026	0,028
79	122	0,110	0,163	0,117	0,013	0,125	0,133	0,016	0,017
80	125	0,308	0,314	0,281	0,020	0,262	0,239	0,016	0,017
81	7	0,000	0,000	0,029	0,043	0,066	0,099	0,065	0,069
82	12	0,000	0,045	0,025	0,032	0,047	0,070	0,048	0,051
83	13	0,049	0,017	0,068	0,035	0,088	0,108	0,050	0,053
84	4	0,000	0,061	0,028	0,056	0,060	0,091	0,088	0,093
85	32	0,189	0,113	0,193	0,027	0,217	0,231	0,035	0,037
86	10	0,136	0,056	0,137	0,036	0,127	0,123	0,051	0,054
87	52	0,192	0,098	0,185	0,021	0,184	0,180	0,024	0,026
88	65	0,153	0,120	0,155	0,018	0,162	0,166	0,022	0,024
89	71	0,285	0,210	0,265	0,022	0,256	0,242	0,022	0,023
90	57	0,086	0,146	0,095	0,017	0,102	0,110	0,022	0,024
91	153	0,149	0,167	0,150	0,011	0,156	0,160	0,014	0,015
92	138	0,308	0,285	0,283	0,020	0,266	0,244	0,015	0,017
93	10	0,000	0,000	0,030	0,041	0,073	0,110	0,059	0,063
94	16	0,067	0,015	0,081	0,029	0,090	0,101	0,042	0,044
95	18	0,108	0,018	0,123	0,032	0,145	0,163	0,046	0,049
96	14	0,111	0,087	0,125	0,039	0,160	0,185	0,050	0,053

Les estimations HB de la proportion de personnes non assurées pour le groupe des Asiatiques varient de 2 % à 35 % pour les divers petits domaines, à l'exclusion du domaine 69. Il faut reconnaître que, pour ce dernier, les estimations EB et HB sont affectées défavorablement par la petite taille d'échantillon. Nous présentons également les erreurs-types associées aux estimations HB, ainsi que la racine des erreurs quadratiques moyennes approximatives estimées des estimations EB. L'approche proposée répond largement à la critique valide selon laquelle les estimations EB naïves des erreurs-types (qui ne tiennent pas compte du terme $O(k^{-1})$) sont habituellement des sous-estimations. Nous donnons également une colonne contenant la moyenne de trois ans des estimations directes pour 1997 à 1999. Celle-ci a principalement pour but d'examiner si les domaines pour lesquels l'estimation directe était nulle en 2000 possèdent la même caractéristique pour d'autres années, ainsi que de comparer les estimations EB et HB à ces estimations plutôt qu'aux estimations directes. Il s'avère qu'à très peu d'exceptions près, la moyenne pour 1997 à 1999 concorde peu aux estimations directes. Cependant, l'estimation directe demeure nulle pour le domaine 69.

Le tableau 2 donne un sommaire des proportions de personnes non assurées dans les trois groupes d'âge, à savoir 0 à 17 ans, 18 à 64 ans et 65 ans et plus, individuellement pour les Chinois (Asiatique 1), les Philippins (Asiatique 2), les Indiens d'Asie (Asiatique 3) et les autres Asiatiques (Asiatique 4). Il s'avère qu'à ce niveau plus élevé d'agrégation, les estimations EB et HB pour petits domaines sont assez proches des estimations directes, sauf peut-être pour le groupe des 65 ans et plus. Les résultats semblent assez satisfaisants, puisqu'à ce niveau d'agrégation, les estimations directes servent souvent de référence dans les comparaisons.

Tableau 2
Proportions n'ayant pas d'assurance-maladie, selon le groupe d'âge et le groupe d'Asiatiques en 2000

	Directes	HB	EB ($\lambda = 0,5$)	EB ($\lambda = 1$)
0 à 17 ans				
Total	0,120	0,126	0,131	0,137
Asiatique 1	0,087	0,097	0,105	0,114
Asiatique 2	0,046	0,062	0,071	0,083
Asiatique 3	0,113	0,117	0,114	0,114
Asiatique 4	0,165	0,165	0,171	0,175
18 à 64 ans				
Total	0,177	0,172	0,168	0,164
Asiatique 1	0,162	0,160	0,160	0,159
Asiatique 2	0,137	0,137	0,135	0,134
Asiatique 3	0,150	0,147	0,141	0,137
Asiatique 4	0,219	0,208	0,203	0,195
65 ans et plus				
Total	0,063	0,080	0,103	0,123
Asian 1	0,083	0,097	0,123	0,143
Asian 2	0,021	0,043	0,064	0,085
Asian 3	0,119	0,126	0,136	0,145
Asian 4	0,055	0,075	0,100	0,123

6. Diagnostics et mise en œuvre du modèle hiérarchique bayésien

Pour la mise en œuvre et les diagnostics de convergence de l'échantillonneur de Gibbs, nous nous sommes inspirés de Gelman et Rubin (1992). En particulier, nous nous sommes servis de cinq chaînes, chacune de taille 1 000, avec une période d'apprentissage initiale de 1 000 itérations. Nous avons vérifié la convergence des facteurs de réduction d'échelle possibles et elle semblait être très proche de l'unité (= 1 à la convergence) pour chacun des paramètres. Plusieurs autres critères diagnostiques décrits dans la littérature ont été appliqués au moyen du logiciel CODA. Une sortie partielle est présentée à la figure 1. Le côté gauche montre le chevauchement des cinq chaînes parallèles et le côté droit, l'inférence a posteriori pour chaque paramètre et la déviance (-2 log vraisemblance). Pour une description détaillée du logiciel que nous avons utilisé, nous renvoyons le lecteur à l'annexe C de Gelman, Carlin, Stern et Rubin (2004).

Un moyen bayésien de vérifier la qualité de l'ajustement d'un modèle aux données consiste à tirer des valeurs simulées de la loi prédictive a posteriori des données répliquées et de comparer ces échantillons aux données observées. Un écart important entre les données générées et les données observées indique un manque d'ajustement du modèle. À l'instar de Gelman et coll. (2004), nous avons calculé les valeurs *p* bayésiennes pour vérifier l'adéquation des modèles bayésiens proposés. La justification générale de ce genre de calcul est la suivante. Soit y le vecteur de données observées, ξ le vecteur de paramètres inconnus, $f(y \mid \xi)$ la densité de y sachant ξ et $\Pi(\xi \mid y)$, la densité a posteriori de ξ sachant y. Supposons que nous avons tiré les échantillons $\xi^{(1)}, \ldots, \xi^{(R)}$ de cette loi a posteriori en utilisant une simulation MCMC. Simulons maintenant R répliques hypothétiques des données, disons $y^{(1)}, \ldots, y^{(R)}$, où $y^{(l)}, (l = 1, \ldots, R)$ est tiré de la loi conditionnelle de y sachant le $\xi^{(l)}$ simulé. Si le modèle est raisonnablement précis, ces répliques hypothétiques devraient être similaires aux données observées y. Formellement, nous faisons cela en commençant par choisir une variable de divergence, disons $d(y, \xi)$, qui aura une valeur extrême si les données y sont entièrement en désaccord avec le modèle donné. Puis, nous estimons une valeur p par la proportion de cas dans lesquels la variable de divergence simulée excède la valeur réalisée de cette variable. Donc, la valeur p estimée (habituellement appelée valeur p prédictive a posteriori) est égale à $R^{-1}\sum_{l=1}^{R} I_{[d(y^{(l)}, \xi^{(l)}) \geq d(y, \xi^{(l)})]}$, où I est la fonction indicatrice habituelle. Pour vérifier la qualité de l'ajustement du modèle, nous pouvons utiliser un diagramme de dispersion des valeurs réalisées $d(y, \xi^{(l)})$ en fonction des valeurs prédictives $d(y^{(l)}, \xi^{(l)})$ sur la même échelle. Si la moitié environ des points du diagramme de dispersion se trouvent au-dessus de la droite à 45 degrés et l'autre moitié, sous celle-ci, l'ajustement est bon. Autrement dit, pour les grands échantillons, la valeur p estimée s'écartera peu d'une moitié. Naturellement, nous pouvons également exécuter une analyse graphique en utilisant différents diagrammes pour divers sous-groupes, ce qui permettrait de visualiser un échec local éventuel du modèle qui, autrement, pourrait être obscurci dans le diagramme de dispersion agrégé.

Plusieurs choix sont possibles pour la variable de divergence d. Dans le cas qui nous occupe, nous en avons considéré une en particulier. Notant que $E(Y_{ij} \mid p_{ij}) = p_{ij} = \exp(\theta_{ij})/(1 + \exp(\theta_{ij}))$, nous pouvons considérer le carré des résidus standardisés $((y_{ij} - p_{ij}^{(l)})^2)/(p_{ij}^{(l)}(1 - p_{ij}^{(l)}))$, où $p_{ij}^{(l)} = \exp(\theta_{ij}^{(l)})/(1 + \exp(\theta_{ij}^{(l)}))$ sont les valeurs générées des p_{ij} à partir de la l^{e} itération. Alors, la variable de divergence d est

$$d(y, p^{(l)}) = \sum_{i=1}^{95} \sum_{j=1}^{n_i} \frac{(y_{ij} - p_{ij}^{(l)})^2}{p_{ij}^{(l)}(1 - p_{ij}^{(l)})}$$

$$d(y^{(l)}, p^{(l)}) = \sum_{i=1}^{95} \sum_{j=1}^{n_i} \frac{(y_{ij}^{(l)} - p_{ij}^{(l)})^2}{p_{ij}^{(l)}(1 - p_{ij}^{(l)})}.$$

Manifestement, il existe d'autres choix possibles de d. Gelfand et Ghosh (1998) ont proposé un certain nombre de mesures de divergence et étudié leurs propriétés.

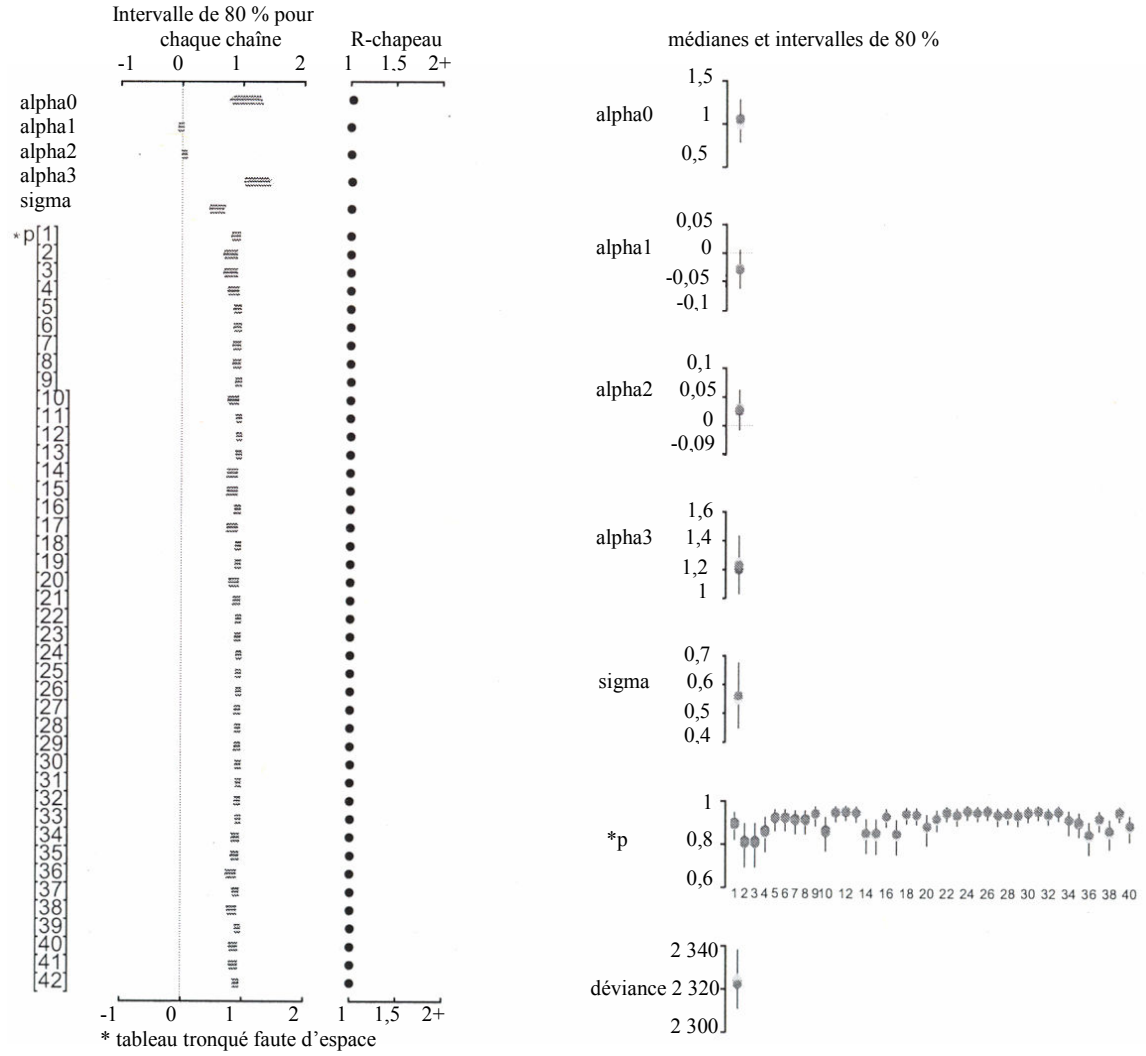


Figure 1 Modèle Bugs à « asian_model.bug », 5 chaînes, chacune avec 1 000 itérations

Pour le modèle de régression logistique hiérarchique bayésien, la valeur p estimée est 0,4216 pour $(c, d) = (0,02,\ 0,02)$. Les autres choix de (c, d) produisent des valeurs semblables. Une valeur p supérieure à 0,3 est habituellement traitée comme un bon ajustement. Donc, la méthode HB proposée semble donner de bons résultats dans la situation décrite ici.

Nous avons également calculé la valeur p_D = var(déviance)/2 et le *critère d'information de déviance* (DIC pour *deviance information criterion*) ou la déviance prédictive estimée. La valeur p_D peut être vue comme le nombre de paramètres non contraints dans le modèle, où un paramètre prend la valeur 1 s'il fait partie du modèle original (distribution des données) et la valeur 0 s'il est associé à toute loi a priori. Le DIC est estimé sous la forme

$$\text{DIC} = 2\hat{D}(y, \theta^{(l)}) - \hat{D}(y, \hat{\theta})$$

où $\hat{D}(y, \hat{\theta})$ est la déviance calculée aux valeurs estimées des paramètres et $\hat{D}(y, \theta^{(l)})$ est la déviance estimée en utilisant la simulation a posteriori. Pour des détails, voir Gelman et coll. (2004).

Pour notre analyse HB, $p_D = 56,75$ et DIC = 2 414,41. Habituellement, les valeurs p_D et DIC sont utilisées comme critères d'ajustement du modèle et pour choisir le modèle ayant le meilleur pouvoir prédictif. Dons, nous ajustons aussi le modèle de régression logistique simple (modèle courant sans aucun effet aléatoire), ce qui signifie qu'il n'y a pas de groupement de données, et les valeurs estimées de p_D et DIC sont 22,60 et 2 379,55, respectivement. La valeur p correspondante est 0,3848. Donc le modèle proposé semble être raisonnablement bien ajusté aux données.

7. Résumé, futurs travaux et discussion

L'estimation de la proportion de personnes non assurées, particulièrement parmi les groupes minoritaires, est définitivement un problème très important, susceptible d'avoir une incidence sur l'élaboration des politiques des organismes fédéraux et des États. Nous venons de commencer à aborder cette question très importante et fournissons des estimations pour petits domaines bayésiennes empiriques et hiérarchiques pour la sous-population asiatique, subdivisée par croisement de l'âge, du sexe et d'autres caractéristiques démographiques. Nous discutons également de l'adéquation de l'ajustement de notre modèle au moyen de la valeur *p* prédictive a posteriori. Néanmoins, beaucoup de travaux restent à faire. En particulier, nous voulons étendre les présents résultats à l'analyse de données binaires bivariées et multivariées.

Comme l'a souligné un examinateur, la présente analyse ne tient pas compte de la mise en grappes des ménages dans le calcul de la vraisemblance, puisque l'enquête originale est une enquête-ménage et que, très catégoriquement, la couverture par une assurance est corrélée à l'intérieur des ménages. Cependant, notre modèle hypothétique n'est que conditionnellement indépendant, sachant les covariables et les effets aléatoires. Une fois que nous avons attribué des distributions aux effets aléatoires et, subséquemment, des distributions aux coefficients de régression et aux composantes de variance, la dépendance est intégrée automatiquement dans le modèle final, au niveau de l'unité ainsi que du domaine. En outre, comme nous l'avons mentionné plus haut, nous avons testé l'adéquation du modèle hiérarchique bayésien au moyen des valeurs *p* prédictives a posteriori.

Enfin, il convient de souligner que ces travaux de recherche ne sont présentés ici qu'à titre illustratif. L'application de cette méthode dans un contexte d'élaboration de politiques nécessiterait un examen plus approfondi des méthodes et du respect des normes institutionnelles relatives à la diffusion de politiques officielles.

Remerciements

Nous remercions le rédacteur adjoint et deux examinateurs de leurs commentaires constructifs au sujet d'ébauches antérieures de l'article. Les travaux de recherche du premier auteur ont été financés en partie par les bourses SES-0317589 et SES-0631426 de la NSF. Ceux du quatrième auteur ont été financés en partie par la bourse SES-0318184 de la NSF. Les travaux ont également été financés partiellement par les NCHS/CDC sous le projet numéro 282286285 intitulé « Model-Based Subdomain Estimates ». Les constatations et les conclusions du présent article sont celles des auteurs et ne reflètent pas nécessairement les opinions du National Center for Health Statistics des Centers for Disease Control and Prevention.

Annexe

Preuve du théorème 1.

$$\mathrm{EQM}(\hat{\bar{\mu}}_{iw}^{\mathrm{EB}})$$

$$= E(\hat{\bar{\mu}}_{iw}^{\mathrm{EB}} - \hat{\bar{\mu}}_{iw})^2 = E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - p_{ij})\right)^2$$

$$= E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B} + \hat{p}_{ij}^{B} - p_{ij})\right)^2$$

$$= E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right)^2 + E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)^2$$

$$+ 2E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right)\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)\right].$$

En notant que $E(p_{ij} \mid \boldsymbol{y}) = \hat{p}_{ij}^{B}$,

$$E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right)\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)\right]$$

$$= E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right) \times E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij}) \mid \boldsymbol{y}\right)\right] = 0.$$

D'où,

$$\mathrm{EQM}(\hat{\bar{\mu}}_{iw}^{\mathrm{EB}}) = E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right)^2$$

$$+ E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)^2. \qquad \text{(A.1)}$$

Mais

$$E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)^2 = \sum_{j=1}^{n_i} w_{ij}^2 E(\hat{p}_{ij}^{B} - p_{ij})^2.$$

Ensuite, nous calculons

$$
\begin{aligned}
&E(\hat{p}_{ij}^{B} - p_{ij})^2 \\
&= E\left[\frac{1}{\lambda+1} y_{ij} + \frac{\lambda}{\lambda+1} m_{ij}(\boldsymbol{b}) - p_{ij}\right]^2 \\
&= E\left[\frac{1}{\lambda+1}(y_{ij} - p_{ij}) + \frac{\lambda}{\lambda+1}(m_{ij}(\boldsymbol{b}) - p_{ij})\right]^2 \\
&= \frac{1}{(\lambda+1)^2} E(y_{ij} - p_{ij})^2 + \frac{\lambda^2}{(\lambda+1)^2} E\left(m_{ij}(\boldsymbol{b}) - p_{ij}\right)^2 \\
&+ \frac{2\lambda}{(\lambda+1)^2} E(y_{ij} - p_{ij})(m_{ij}(\boldsymbol{b}) - p_{ij}) \\
&= \frac{1}{(\lambda+1)^2} E(p_{ij}(1-p_{ij})) + \frac{\lambda^2}{(\lambda+1)^2} V(p_{ij}) + 0 \\
&= \frac{1}{(\lambda+1)^2}\left(\frac{\lambda}{(\lambda+1)} m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\right) \\
&+ \frac{\lambda^2}{(\lambda+1)^2}\left(\frac{m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))}{\lambda+1}\right) \\
&= \frac{\lambda m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))}{(\lambda+1)^2},
\end{aligned}
$$

de sorte que

$$
E\left[\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right]^2 = \frac{\lambda}{(\lambda+1)^2} \sum_{j=1}^{n_i} w_{ij}^2 m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b})). \qquad \text{(A.2)}
$$

Enfin, nous calculons,

$$
\begin{aligned}
&E\left[\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right]^2 \\
&= \frac{\lambda^2}{(\lambda+1)^2} E\left[\sum_{j=1}^{n_i} w_{ij}(m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b}))\right]^2 \\
&= \frac{\lambda^2}{(\lambda+1)^2} E\left[\sum_{j=1}^{n_i} w_{ij}^2 (m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b}))^2\right. \\
&\qquad + \sum_{1 \le j \ne j' \le n_i} \sum w_{ij} w_{ij'} (m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b})) \\
&\qquad \left.(m_{ij'}(\hat{\boldsymbol{b}}) - m_{ij'}(\boldsymbol{b}))\right]. \qquad \text{(A.3)}
\end{aligned}
$$

Par développement en série de Taylor en deux étapes,

$$
\begin{aligned}
m_{ij}(\hat{\boldsymbol{b}}) &\doteq m_{ij}(\boldsymbol{b}) + \left(\frac{\partial m_{ij}(\boldsymbol{b})}{\partial \boldsymbol{b}}\right)^T \\
&(\hat{\boldsymbol{b}} - \boldsymbol{b}) + \frac{1}{2}(\hat{\boldsymbol{b}} - \boldsymbol{b})^T \frac{\partial^2 m_{ij}(\boldsymbol{b})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}(\hat{\boldsymbol{b}} - \boldsymbol{b}).
\end{aligned}
$$

Notant que $(\partial^2 m_{ij}(\boldsymbol{b}))/(\partial \boldsymbol{b} \partial \boldsymbol{b}^T) = (1 - 2m_{ij}(\boldsymbol{b}))m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T$, il s'ensuit que

$$
\begin{aligned}
&E[m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b})]^2 \\
&\doteq E\left[m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b}) + \frac{1}{2}(\hat{\boldsymbol{b}} - \boldsymbol{b})^T\right. \\
&\qquad \left. m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))(1 - 2m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})\right]^2 \\
&= m_{ij}^2(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))^2 E\left[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b}) + \frac{1}{2}(1 - 2m_{ij}(\boldsymbol{b}))\right. \\
&\qquad \left.(\hat{\boldsymbol{b}} - \boldsymbol{b})^T \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})\right]^2. \qquad \text{(A.4)}
\end{aligned}
$$

Le premier terme négligé est $O_p(\| \hat{\boldsymbol{b}} - \boldsymbol{b} \|^3)$. De Sarkar et Ghosh (1998), il découle que $\hat{\boldsymbol{b}} - \boldsymbol{b}$ est asymptotiquement de loi $\mathrm{N}(0, \boldsymbol{\Sigma}^{-1}(\boldsymbol{b}))$, où $\boldsymbol{\Sigma}(\boldsymbol{b})$ est défini avant le théorème 1. Sous l'hypothèse que $\boldsymbol{\Sigma}(\boldsymbol{b}) = O_e(k)$, il découle que $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b}) = O_e(k^{-1})$. Donc, $\| \hat{\boldsymbol{b}} - \boldsymbol{b} \| = O_p(k^{-1/2})$. D'où, le premier terme négligé est $O_p(k^{-3/2})$. Ensuite, nous observons que

$$
\begin{aligned}
E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})]^2 &= E[(\hat{\boldsymbol{b}} - \boldsymbol{b})^T \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})] \\
&= tr[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T E(\hat{\boldsymbol{b}} - \boldsymbol{b})(\hat{\boldsymbol{b}} - \boldsymbol{b})^T]. \qquad \text{(A.5)}
\end{aligned}
$$

Afin de trouver $E[(\hat{\boldsymbol{b}} - \boldsymbol{b})(\hat{\boldsymbol{b}} - \boldsymbol{b})^T]$, nous procédons comme il suit.

Soit $\boldsymbol{T}(\boldsymbol{b}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}$ de sorte que $\boldsymbol{T}(\hat{\boldsymbol{b}}) = \boldsymbol{0}$.

Par développement en série de Taylor en une étape, $\boldsymbol{0} = \boldsymbol{T}(\hat{\boldsymbol{b}}) = \boldsymbol{T}(\boldsymbol{b}) + [\nabla \boldsymbol{T}(\boldsymbol{b})]^T (\hat{\boldsymbol{b}} - \boldsymbol{b}) + O_p(n_T^{-1})$, où

$$
\begin{aligned}
\nabla \boldsymbol{T}(\boldsymbol{b}) &= -\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{\partial m_{ij}(\boldsymbol{b})}{\partial \boldsymbol{b}}\right) \boldsymbol{x}_{ij}^T \\
&= -\sum_{i=1}^{k} \sum_{j=1}^{n_i} m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T \\
&= -\boldsymbol{X}^T \boldsymbol{M}(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{X} \\
&= -\boldsymbol{\Sigma}(\boldsymbol{b}). \qquad \text{(A.6)}
\end{aligned}
$$

Donc, $\hat{\boldsymbol{b}}-\boldsymbol{b}=\boldsymbol{\Sigma}^{-1}\boldsymbol{T}(\boldsymbol{b})+O_p(n_T^{-1})$. Puisque $V(y_{ij})=m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))$, $V(T(\boldsymbol{b}))=\boldsymbol{\Sigma}(\boldsymbol{b})$. D'où $E[(\hat{\boldsymbol{b}}-\boldsymbol{b})(\hat{\boldsymbol{b}}-\boldsymbol{b})^T]=\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})+O(n_T^{-3/2})$. En outre, dans (8.5), nous avons $E(\boldsymbol{x}_{ij}^T)(\hat{\boldsymbol{b}}-\boldsymbol{b})^2=tr(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b}))+O_p(n_T^{-1})$. Conséquemment, par (A.4) et (A.5), nous obtenons l'approximation

$$E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]^2=m_{ij}^2(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))^2\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij} \quad \text{(A.7)}$$

qui est correcte jusqu'à l'ordre $O(n_T^{-1})$ en vertu de notre hypothèse.

Soulignons que le terme négligé

$$E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})(1-2m_{ij}(\boldsymbol{b}))(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]$$
$$=O(n_T^{-3/2})$$

puisque

$$E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})(1-2m_{ij}(\boldsymbol{b}))(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]$$
$$=(1-2m_{ij}(\boldsymbol{b}))E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\boldsymbol{x}_{ij}]$$
$$=(1-2m_{ij}(\boldsymbol{b}))E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]^3$$
$$=O(n_T^{-3/2}).$$

De même, notons que $E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]^4=O(n_T^{-2})$ et

$$E[(m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b}))(m_{ij'}(\hat{\boldsymbol{b}})-m_{ij'}(\boldsymbol{b}))]$$
$$=m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))m_{ij'}(\boldsymbol{b})(1-m_{ij'}(\boldsymbol{b}))\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij'}$$
$$+O(n_T^{-3/2}). \quad \text{(A.8)}$$

Cela nous mène à

$$E\left[\sum_{j=1}^{n_i}w_{ij}(\hat{p}_{ij}^{\mathrm{EB}}-\hat{p}_{ij}^{B})\right]^2$$
$$=\frac{\lambda^2}{(\lambda+1)^2}\left[\sum_{j=1}^{n_i}w_{ij}^2m_{ij}^2(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))^2\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij}\right.$$
$$+\sum_{1\le j\ne j'\le n_i}\sum w_{ij}w_{ij'}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))m_{ij'}(\boldsymbol{b})$$
$$\left.(1-m_{ij'}(\boldsymbol{b}))\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij'}\right]+O(n_T^{-3/2})$$
$$=\frac{\lambda^2}{(\lambda+1)^2}\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\right]^T$$
$$\times\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\right]+O(n_T^{-3/2}). \quad \text{(A.9)}$$

Puisque $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})=O(k^{-1})$, et $n_T=O_e(k)$ en vertu de notre hypothèse, le théorème découle de (A.2) et (A.9).

Preuve du théorème 2. Nous commençons par noter que $\hat{\boldsymbol{b}}=\boldsymbol{b}+O_p(n_T^{-1/2})=\boldsymbol{b}+O_p(K^{-1})$ et $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})=O(k^{-1})$. D'où, le deuxième terme du deuxième membre de (8.7) est approximé par

$$c\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))\boldsymbol{x}_{ij}\right]^T$$
$$\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))\boldsymbol{x}_{ij}\right] \quad \text{(A.10)}$$

($c=\lambda^2/(1+\lambda)^2$) qui est exact jusqu'à l'ordre $O(k^{-1})$. Cependant, si nous estimons $m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))$ simplement par $m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))$, nous ignorerons le terme d'ordre $O(k^{-1})$. Donc, nous avons besoin d'une approximation prudente du biais $E(\hat{\boldsymbol{b}}-\boldsymbol{b})$ pour obtenir l'approximation souhaitée. Pour cela, nous nous inspirons de Cox et Snell (1968).

Nous partons de l'identité

$$E[m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))]$$
$$=E[(m_{ij}(\boldsymbol{b})+m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b}))(1-m_{ij}(\boldsymbol{b})+m_{ij}(\boldsymbol{b})-m_{ij}(\hat{\boldsymbol{b}}))]$$
$$=m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))+(1-2m_{ij}(\boldsymbol{b}))E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]$$
$$-E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]^2.$$

Maintenant, de nouveau par développement en série de Taylor en deux étapes, nous avons

$$E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]=\left[\frac{\partial m_{ij}(\boldsymbol{b})}{\partial\boldsymbol{b}}\right]^TE(\hat{\boldsymbol{b}}-\boldsymbol{b})$$
$$+\frac{1}{2}E\left[(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\frac{\partial^2m_{ij}(\boldsymbol{b})}{\partial\boldsymbol{b}\partial\boldsymbol{b}^T}(\hat{\boldsymbol{b}}-\boldsymbol{b})\right]+O(n_T^{-3/2}).$$

Afin de trouver $E(\hat{\boldsymbol{b}}-\boldsymbol{b})$, nous procédons comme il suit. Nous commençons par le développement en série de Taylor de deuxième ordre

$$0=T_r(\hat{\boldsymbol{b}})=T_r(\boldsymbol{b})+\sum_{s=1}^{p}(\hat{b}_s-b_s)\frac{\partial T_r(\boldsymbol{b})}{\partial b_s}$$
$$+\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}(\hat{b}_s-b_s)(\hat{b}_t-b_t)\frac{\partial^2T_r(\boldsymbol{b})}{\partial b_s\partial b_t}+O(n_T^{-3/2}).$$

En prenant les espérances et en suivant Cox et Snell (1968),

$$0 = E(T_r(\hat{\boldsymbol{b}}))$$

$$= \sum_{s=1}^{p}\left[E(\hat{b}_s - b_s)E\left(\frac{\partial T_r(\boldsymbol{b})}{\partial b_s}\right) + \mathrm{Cov}\left(\hat{b}_s - b_s, \frac{\partial T_r(\boldsymbol{b})}{\partial b_s}\right)\right]$$

$$+ \frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}(E(\hat{b}_s - b_s)(\hat{b}_t - b_t))\left(\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right)$$

$$+ \frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}\mathrm{Cov}\left[(\hat{b}_s - b_s)(\hat{b}_t - b_t), \left(\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right)\right]$$

$$+ O(n_T^{-3/2}) = -\sum_{s=1}^{p} E(\hat{b}_s - b_s)\sigma_{rs}$$

$$+ \sum_{s=1}^{p}\sum_{u=1}^{p}\mathrm{Cov}\left[\sigma^{su}(\boldsymbol{b})T_u(\boldsymbol{b}), \frac{\partial T_r(\boldsymbol{b})}{\partial b_s}\right]$$

$$+ \frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}\sigma^{st}E\left[\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right] + O(n_T^{-3/2}). \qquad \text{(A.11)}$$

Notons que $\mathrm{Cov}[\sigma^{su}(\boldsymbol{b})T_u(\boldsymbol{b}), \partial T_r(\boldsymbol{b})/\partial b_s] = 0$, puisque $\partial T_r(\boldsymbol{b})/\partial b_s$ est une constante indépendante des y_{ij}. De même,

$$\mathrm{Cov}\left[(\hat{b}_s - b_s)(\hat{b}_t - b_t), \left(\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right)\right] = 0.$$

En outre, posons que

$$K_{rst} = E\left[\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right]$$

$$= \frac{\partial}{\partial b_t}\sum_{i=1}^{k}\sum_{j=1}^{n_i} - m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))x_{ijr}\, x_{ijs}$$

$$= -\sum_{i=1}^{k}\sum_{j=1}^{n_i}(1 - 2m_{ij}(\boldsymbol{b}))m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))x_{ijr}\, x_{ijs}\, x_{ijt}. \qquad \text{(A.12)}$$

Donc, nous avons

$$\sum_{s=1}^{k}\sigma_{rs}E(\hat{b}_s - b_s) \doteq \sum_{s=1}^{k}\sum_{t=1}^{p}\sigma^{su}K_{rst},\ r = 1, ..., p.$$

En notation matricielle, nous obtenons

$$\boldsymbol{\Sigma}\, E(\hat{\boldsymbol{b}} - \boldsymbol{b}) = \frac{1}{2}\begin{pmatrix} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_1) \\ . \\ . \\ . \\ tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_p) \end{pmatrix}$$

où $\boldsymbol{K}_r = ((K_{rst}))$.

D'où,

$$E(\hat{\boldsymbol{b}} - \boldsymbol{b}) \doteq \frac{1}{2}\boldsymbol{\Sigma}^{-1}\begin{pmatrix} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_1) \\ . \\ . \\ . \\ tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_p) \end{pmatrix} + O(n_T^{-3/2}).$$

Puisque $n_T = O_e(k)$, le théorème s'ensuit.

Bibliographie

Botman, S.L., Moore, T.F., Moriarity, C.L. et Parsons, V.L. (2000). Design and estimation for the National Health Interview Survey, 1995-2004. *Vital and Health Statistics*, 2, 130.

Cox, D.R., et Snell, E.J. (1968). A general distribution of residuals (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 30, 248-275.

Ghosh, M., et Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, 91, 95-112.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Gelman, A., Carlin, J., Stern, H. et Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A., et Rubin, D. (1992). Inference from iterative simulation using multiple sequences (avec discussion). *Statistical Science*, 457-511.

Godambe, V.P., et Thompson, M.E. (1989). An extension of quasi-likelihood estimation (avec discussion). *Journal of Statistiscal Planning and Infererence*, 22, 137-152.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10, 65-80.

Morris, C. (1983). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 11, 515-529.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Sarkar, S., et Ghosh, M. (1998). Empirical Bayes estimation of local area means for NEF-QVF superpopulations. *Sankhyā*, Séries B, 60, 464-487.

Tanner, M.A. (1996). *Tools for Statistical Inference*. New York : Springer.

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 71-90
Statistique Canada, N° 12-001-X au catalogue

Un test non paramétrique pour la saisonnalité résiduelle

Tucker McElroy et Scott Holan [1]

Résumé

La présence de pics dans le spectre d'un processus stationnaire signale l'existence de phénomènes périodiques stochastiques, tels que l'effet saisonnier. Nous proposons une mesure de ces pics spectraux et un test de détection de leur présence qui s'appuient sur l'évaluation de leur pente et de leur convexité agrégées. Notre méthode est élaborée de manière non paramétrique et peut donc être utile durant l'analyse préliminaire d'une série. Elle peut aussi servir à détecter la présence d'une saisonnalité résiduelle dans les données désaisonnalisées. Nous étudions le test diagnostique au moyen d'une simulation et d'une étude de cas à grande échelle portant sur des données provenant du U.S. Census Bureau et de l'Organisation de coopération et de développement économiques (OCDE).

Mots clés : Tests multiples ; estimation non paramétrique de la densité ; désaisonnalisation ; densité spectrale.

1. Introduction

La présence d'un pic dans le spectre d'un processus stationnaire est un indice de comportement périodique, tel que la saisonnalité ou un effet de jour ouvrable. Dans les domaines du génie et de l'économétrie, de nombreux auteurs s'intéressent à la détection de ces pics, car un nœud spectral prononcé exerce une forte influence sur la dynamique du processus stochastique. Un pic est indicatif d'une gamme de fréquences dont la contribution à la variance globale du processus stochastique est relativement importante. Si la force du pic, évaluée par sa hauteur et sa largeur relativement aux valeurs voisines, est suffisamment prononcée, tout modèle de la dynamique du processus ne tenant pas compte des périodicités correspondantes sera spécifié incorrectement. Les ingénieurs ainsi que les économétristes pourraient vouloir extraire ou prévoir le signal, tâche qui, dans l'un et l'autre cas, est sensible à la présence de pics spectraux.

Par pic spectral, nous entendons une région de la densité spectrale dont la masse spectrale est supérieure à celle de ses voisines directes ; une définition plus précise est présentée plus loin. Dans le contexte des applications qui nous intéressent, les pics ont une hauteur finie et correspondent donc à des effets périodiques stochastiques dans un processus stationnaire. Par conséquent, notre principale préoccupation n'est pas la détection des effets périodiques fixes (déterministes) ni celle des phénomènes périodiques non stationnaires (quoique nous fassions certaines extensions à ce cas à la section 3.4) qui, dans les deux cas, correspondent à des pics spectraux de hauteur infinie. L'abondante littérature traitant de la détection des effets fixes est discutée dans Priestley (1981) ; dans nos applications, au lieu d'être fixes, les aspects périodiques évoluent au cours du temps.

Le présent article est axé sur l'application à la désaisonnalisation. Plus précisément, nous nous intéressons aux pics dits saisonniers, qui peuvent survenir aux fréquences saisonnières $\pi/6$, $2\pi/6$, $3\pi/6$, $4\pi/6$, $5\pi/6$ et $6\pi/6$ (en supposant que l'intervalle d'échantillonnage est mensuel). Dans le cas des statistiques fédérales, la détection de la saisonnalité et de la saisonnalité résiduelle pose un problème pratique important et le spectre est un outil logique pour l'effectuer. L'approche basée sur le domaine fréquentiel pour détecter et analyser la saisonnalité est très répandue, parce qu'elle offre un moyen tout naturel de visualiser un comportement quasi périodique. En fait, la saisonnalité est, informellement parlant, caractérisée par la présence d'au moins un pic saisonnier dans le spectre (Nerlove 1964). Les méthodes axées sur le domaine fréquentiel sont maintenant employées dans X-12-ARIMA (Findley, Monsell, Bell, Otto et Chen 1998) et font partie de TRAMO-SEATS (Maravall et Caporello 2004), qui sont les deux programmes de désaisonnalisation les plus répandus mis à la disposition du public. Notons que les méthodes du domaine fréquentiel peuvent être appliquées selon une approche paramétrique (c'est-à-dire fondée sur un modèle) ou non paramétrique. Nous mettons au point un diagnostic non paramétrique appelable pour déterminer l'efficacité de toute procédure de désaisonnalisation, qu'elle soit fondée sur un modèle ou non paramétrique. Comme l'ont souligné Findley, Monsell, Bell, Otto et Chen (1998), l'utilisation de fonctions périodiques fixes seulement pour modéliser la saisonnalité est habituellement inadéquate pour les données économiques (voir aussi la discussion dans Bell et Hillmer 1984).

1. Tucker McElroy, Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100. Courriel : tucker.s.mcelroy@census.gov; Scott Holan, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100. Courriel : holans@missouri.edu.

La présence de pics spectraux à des fréquences saisonnières dans une série désaisonnalisée peut signifier que les filtres saisonniers sont inadéquats – voir Soukup et Findley (1999) pour une discussion. Au minimum, les filtres de désaisonnalisation devraient éliminer la saisonnalité *non stationnaire* et tout effet périodique fixe, c'est-à-dire les phénomènes qui, dans la série observée, contribuent à la formation d'un pôle saisonnier dans le spectre. Cependant, la grande majorité des désaisonnaliseurs estiment qu'il est également souhaitable d'éliminer certains aspects de la saisonnalité *stationnaire* – d'où l'explosion des travaux de développement de filtres de désaisonnalisation fondés sur un modèle (Bell et Hillmer 1984).

La publication antérieure la plus importante à ce sujet est celle de Soukup et Findley (1999), qui proposent d'utiliser un spectre autorégressif pour trouver des pics « visuellement significatifs » – essentiellement, la valeur du spectre à chaque fréquence saisonnière (ou fréquence de jour ouvrable) est comparée aux valeurs voisines les plus proches, et est considérée comme un pic si l'écart est suffisamment grand. Cette méthode est appliquée à l'heure actuelle dans le programme X-12-ARIMA du U.S. Census Bureau (2002). L'une de ses limites tient au fait qu'elle ne comporte vraiment aucune composante statistique : la signification est non statistique – c'est-à-dire qu'elle n'est pas associée à un test d'hypothèse – et les seuils pour déterminer la « signification visuelle » sont établis de manière *ad hoc*. Dans le présent article, nous proposons un test de signification statistique pour la détection des pics qui peut donc être utilisé pour offrir des preuves statistiques supplémentaires de la présence d'un pic.

Un autre article apparenté est celui de Newton et Pagano (1983), qui élaborent des estimateurs convergents pour les maximiseurs locaux du spectre. Notre approche est légèrement différente, en ce sens que nous connaissons déjà les fréquences d'intérêt (les six fréquences saisonnières), mais que nous cherchons à tester la présence d'un pic statistiquement significatif. Si nous considérons la vraie densité spectrale f comme une fonction lisse (qui peut être quantifiée au moyen d'une décroissance suffisamment rapide de la fonction d'autocovariance), un pic est une fréquence λ_0 telle que

$$\dot{f}(\lambda_0) = 0 \qquad \ddot{f}(\lambda_0) < 0, \tag{1}$$

où $\dot{f}$ et $\ddot{f}$ désignent les dérivées première et seconde. Manifestement, la dérivée seconde doit être négative *avec une certaine signification* pour que le concept ait un sens. Tout bien réfléchi, il semble naïf d'examiner la géométrie infinitésimale de f au point unique λ_0, puisque n'importe quel petit pic dans une fonction monotone peut satisfaire (1), tout en étant dissocié des notions plus intuitives de ce qui constitue un pic. Par conséquent, il faut que la convexité soit négative dans un voisinage raisonnablement grand de λ_0. Cette réflexion aboutit au diagnostic proposé dans le présent article, c'est-à-dire des mesures agrégées de la pente et de la convexité de la densité spectrale, normalisées comme il convient. Mathématiquement, elles prennent la forme d'estimations d'un périodogramme lissé par la méthode du noyau, mais sans que la largeur de bande dépende de la taille de l'échantillon.

À la section 2, nous développons les notions mathématiques de cette méthode et les illustrons au moyen de deux noyaux soigneusement choisis. À la section 3, nous montrons comment formuler les estimateurs statistiques et comment tester les hypothèses statistiques concernant les pics. À la section 4, nous mettons la méthodologie à l'essai ; des simulations fournissent une description, en échantillon fini, de la taille et de la puissance de notre test. Nous démontrons en outre l'utilité de nos méthodes au moyen d'une étude de cas à grande échelle portant sur 130 séries chronologiques provenant du U.S. Census Bureau et de l'Organisation de coopération et de développement économiques (OCDE). Nous appliquons certains concepts tirés de la littérature sur les tests multiples (Hochberg 1988) pour combiner en un seul diagnostic des tests basés sur les fréquences individuelles. À la section 5, nous présentons nos conclusions, et à l'annexe, tous les théorèmes et preuves.

2. Mesure de la géométrie locale du spectre

En premier lieu, nous discutons de la géométrie de la densité spectrale (ou spectre) de la série chronologique étudiée. Comme point de départ, nous considérons des mesures de la pente et de la convexité du spectre qui sont entièrement déterministes (voir l'approche de Newton et Pagano 1983) ; plus tard, à la section 3, nous envisagerons des mesures statistiques. À la section 2.1, nous présentons les concepts des mesures de la pente et de la convexité. À la section 2.2, nous discutons de la pertinence de ces mesures pour la détection des pics, tandis qu'à la section 2.3, nous présentons, à titre d'exemples explicites, deux noyaux simples.

Supposons qu'après des transformations appropriées et une différentiation au besoin, $X_1, X_2, ..., X_n$ est un échantillon issu d'un processus stochastique stationnaire de moyenne nulle. Nous utilisons la notation $X = (X_1, X_2, ..., X_n)'$. La densité spectrale $f(\lambda)$, qui est bien définie à condition que la fonction d'autocovariance $\gamma_f(h)$ soit absolument sommable, est donnée par

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_f(h) e^{-ih\lambda} \tag{2}$$

avec $i = \sqrt{-1}$ et $\lambda \in [-\pi, \pi]$. Il s'ensuit que la transformée de Fourier inverse donne

$$\gamma_f(h) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\lambda)e^{ih\lambda}d\lambda, \qquad (3)$$

une relation que nous utiliserons à plusieurs reprises dans la suite. Naturellement, cette relation entre γ_g et g est vérifiée pour toute fonction intégrable g et non pas uniquement pour une densité spectrale. En outre, si nous désignons la matrice de Toeplitz associée à γ_g par $\Sigma(g)$, il s'ensuit que

$$\Sigma_{jk}(g) = \gamma_g(j-k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\lambda)e^{i(j-k)\lambda}d\lambda.$$

Or, d'après (2), f est une fonction d fois continûment dérivable si $\sum_{h=-\infty}^{\infty}|h|^d|\gamma_f(h)|<\infty$. Dans la suite de l'article, nous supposons que f est deux fois continûment dérivable (nous représenterons en abrégé cet espace de fonctions par C^2).

2.1 Mesures de la pente et de la convexité

La géométrie locale d'une fonction dans C^2 peut être décrite au moyen de ses dérivées première et seconde ; une mesure agrégée de ces dérivés s'obtient par intégration sur une bande de fréquences. Alternativement, on peut intégrer contre une fonction A qui possède un support compact sur cette bande, à condition que A fournisse une approximation appropriée pour l'intégration sur la bande. Nous désignons cette intégrale au moyen du dispositif général d'une fonctionnelle θ_A, où

$$\theta_A(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} A(\lambda)f(\lambda)d\lambda. \qquad (4)$$

La fonction A sera appelée le « noyau » de cette fonctionnelle. Les mesures agrégées de la pente et de la convexité sont donc définies par $\theta_A(\dot{f})$ et $\theta_A(\ddot{f})$, où chaque point désigne une seule dérivée. Ces fonctionnelles donnent une mesure sommaire de la pente et de la convexité de f sur la bande $[\mu-\beta/2,\ \mu+\beta/2]\subset[0,\ \pi]$, et les noyaux correspondants seront par conséquent désignés par $A_{\beta,\,\mu}$. Nous considérons les noyaux ayant les propriétés suivantes : i) $A_{\beta,\,\mu}$ est une fonction dans C^2 sur $[-\pi,\ \pi]$; ii) $A_{\beta,\,\mu}$ est nul en dehors de la bande $[\mu-\beta/2,\ \mu+\beta/2]$; iii) $A_{\beta,\,\mu}$ est symétrique autour de μ sur cette bande ; iv) $\dot{A}_{\beta,\,\mu}(\mu\pm\beta/2)=0$. La condition (iii) garantit que l'emplacement du pic dans f ne sera pas modifié par l'emploi du noyau $A_{\beta,\,\mu}$. Il convient de souligner que nous n'imposons pas que l'intégrale totale de $A_{\beta,\,\mu}$ soit égale à l'unité, parce que plus tard, nous emploierons une normalisation qui tiendra compte automatiquement de la masse totale du noyau. Maintenant, au moyen de (iv) et d'une intégration par partie dans (4), nous obtenons

$$\begin{aligned}\theta_{A_{\beta,\,\mu}}(\dot{f}) &= -\theta_{\dot{A}_{\beta,\,\mu}}(f)\\ \theta_{A_{\beta,\,\mu}}(\ddot{f}) &= \theta_{\ddot{A}_{\beta,\,\mu}}(f).\end{aligned} \qquad (5)$$

Ces formules sont commodes, parce qu'elles ne requièrent que la connaissance de f, et non celle de ses dérivées (en supposant que nous puissions calculer $\dot{A}_{\beta,\,\mu}$ et $\ddot{A}_{\beta,\,\mu}$). En nous inspirant de la littérature abondante sur les noyaux dans la régression non paramétrique et l'estimation de la densité spectrale, nous pouvons commencer par utiliser un noyau pair A défini sur la bande $[-\pi,\ \pi]$ qui satisfait (i) et $\dot{A}(\pm\pi)=0$. Alors, $A_{\beta,\,\mu}$ est défini par

$$A_{\beta,\,\mu}(\lambda) = \frac{2\pi}{\beta}A\left(\frac{2\pi}{\beta}(\lambda-\mu)\right),$$

et est nul en dehors de la bande de fréquence $[\mu-\beta/2,\ \mu+\beta/2]$. Manifestement, nous devons imposer que $\beta\le 2\mu$ et $\beta\le 2(\pi-\mu)$, pour que $[\mu-\beta/2,\ \mu+\beta/2]\subset[0,\ \pi]$; et le noyau $A_{\beta,\,\mu}$ satisfait les conditions (i) à (iv). Notons que nous ne pouvons pas construire ces types de mesures si μ est égal à 0 ou π. En modifiant les variables, nous voyons que

$$\gamma_{A_{\beta,\,\mu}}(h) = \exp\{ih\mu\}\gamma_A(h\beta/2\pi), \qquad (6)$$

de sorte que les effets de β et de μ sont, en un certain sens, séparables. Notons que nous évaluons habituellement γ_A à des valeurs non entières, de sorte que ces relations sont obtenues en étendant (3) à des arguments non entiers. Il peut être troublant de constater que $\gamma_{A_{\beta,\,\mu}}$ est une fonction à valeurs complexes, mais en réalité, seule sa partie réelle entrera dans nos estimateurs statistiques. Évidemment, en dernière analyse, nous nous intéressons à $\dot{A}_{\beta,\,\mu}$ et $\ddot{A}_{\beta,\,\mu}$, qui sont donnés par

$$\dot{A}_{\beta,\,\mu}(\lambda) = \frac{4\pi^2}{\beta^2}\dot{A}\left(\frac{2\pi}{\beta}(\lambda-\mu)\right),$$

et

$$\ddot{A}_{\beta,\,\mu}(\lambda) = \frac{8\pi^3}{\beta^3}\ddot{A}\left(\frac{2\pi}{\beta}(\lambda-\mu)\right).$$

Plus tard, nous considérerons les carrés de ces noyaux, et leurs transformées de Fourier inverses correspondantes. Donc, en supposant que $[\mu-\beta/2,\ \mu+\beta/2]\subset[0,\ \pi]$, les carrés sont donnés par

$$\dot{A}^2_{\beta,\,\mu}(\lambda) = \frac{16\pi^4}{\beta^4}\dot{A}^2\left(\frac{2\pi}{\beta}(\lambda-\mu)\right)$$

et

$$\ddot{A}^2_{\beta,\,\mu}(\lambda) \;=\; \frac{64\pi^6}{\beta^6}\,\ddot{A}^2\left(\frac{2\pi}{\beta}\,(\lambda \,-\, \mu)\right).$$

Enfin, nous notons d'après (4) que nous pouvons réécrire $\theta_A(f)$ sous la forme

$$\theta_A(f) \;=\; \sum_{h=-\infty}^{\infty} \gamma_A(h)\gamma_f(h). \qquad (7)$$

Donc, il pourrait être avantageux de déterminer la séquence $\gamma_A(h)$ à partir du noyau A. En prenant la transformée de Fourier inverse des noyaux susmentionnés de la pente et de la convexité, nous pouvons construire $\Sigma(\dot{A}_{\beta,\,\mu})$, $\Sigma(\ddot{A}_{\beta,\,\mu})$, $\Sigma(\dot{A}^2_{\beta,\,\mu})$ et $\Sigma(\ddot{A}^2_{\beta,\,\mu})$ comme il suit :

$$\begin{aligned}
\gamma_{\dot{A}_{\beta,\,\mu}}(h) \;&=\; \frac{2\pi}{\beta}\exp\{ih\mu\}\,\gamma_{\dot{A}}(h\beta/2\pi),\\
\gamma_{\ddot{A}_{\beta,\,\mu}}(h) \;&=\; \frac{4\pi^2}{\beta^2}\exp\{ih\mu\}\,\gamma_{\ddot{A}}(h\beta/2\pi),\\
\gamma_{\dot{A}^2_{\beta,\,\mu}}(h) \;&=\; \frac{8\pi^3}{\beta^3}\exp\{ih\mu\}\,\gamma_{\dot{A}^2}(h\beta/2\pi),\\
\gamma_{\ddot{A}^2_{\beta,\,\mu}}(h) \;&=\; \frac{32\pi^5}{\beta^5}\exp\{ih\mu\}\,\gamma_{\ddot{A}^2}(h\beta/2\pi). \qquad (8)
\end{aligned}$$

Donc, si nous possédons l'information relative au domaine temporel $\gamma_f(h)$ pour le processus $\{X_t\}$, nous pouvons calculer les mesures de la pente et de la convexité en utilisant (7), sachant la séquence des transformées de Fourier inverses des noyaux appropriés. Puisque $\gamma_f(h)$ est une séquence symétrique, nous ne devons considérer que la partie réelle de $\gamma_A(h)$ s'il se fait que sa valeur est complexe.

2.2 Creux et pics

Les mesures agrégées de la pente et de la convexité spectrales décrites antérieurement représentent les composantes élémentaires des déterminants de la géométrie spectrale locale. Notre objectif général est de déterminer si un intervalle donné du spectre correspond à un pic ou à un creux (ou est monotone). Dans la géométrie de deuxième ordre de l'analyse infinitésimale, un maximum local possède la propriété déterminante que sa dérivée première est nulle et sa dérivée seconde, strictement négative. Il faut donc manifestement examiner, séquentiellement, une mesure de pente et une mesure de convexité définies sur la même bande de fréquences.

Afin de vérifier la présence d'un pic, l'approche séquentielle peut être considérée comme équivalant à formuler des énoncés inférentiels au sujet de $\theta_{A_{\beta,\,\mu}}(\dot{f})$ et $\theta_{A_{\beta,\,\mu}}(\ddot{f})$. Soulignons qu'en formulant ces énoncés inférentiels, nous choisissons la valeur de μ d'avance, en fonction de l'endroit dans le spectre où nous souhaitons détecter un pic (ou un creux) ; β est choisi en fonction des fréquences que nous désirons exclure, décision qui dépend de la mesure dans laquelle nous souhaitons que notre point de vue du spectre soit local. Puis, nous disons que μ est un pic β-agrégé (par rapport à A) du spectre si

$$\theta_A(\dot{f}) \;=\; 0 \qquad \text{et} \qquad \theta_A(\ddot{f}) \;<\; 0.$$

L'aspect séquentiel découle de l'idée que, en général, nous commençons par déterminer si $\theta_A(\dot{f}) \;=\; 0$, puis que nous déterminons la convexité ; cela deviendra plus évident quand nous examinerons les tests statistiques à la section 3.2. De manière comparable, nous définissons un creux β-agrégé quand $\theta_A(\ddot{f}) > 0$. En ce qui concerne le test d'hypothèse pour un pic, nous avons

$$\begin{aligned}
H_0^{(1)} &: \theta_A(\dot{f}) \;=\; 0 \qquad vs \qquad H_a^{(1)} : \theta_A(\dot{f}) \;\neq\; 0\\
H_0^{(2)} &: \theta_A(\ddot{f}) \;=\; 0 \qquad vs \qquad H_a^{(2)} : \theta_A(\ddot{f}) \;<\; 0.
\end{aligned}$$

L'aspect inhabituel de ce test d'hypothèse est que nous voulons d'abord ne pas rejeter $H_0^{(1)}$, puis, conditionnellement à ce test, nous voulons rejeter $H_0^{(2)}$ en faveur de l'alternative $H_a^{(2)}$.

2.3 Exemples de noyaux

Il existe une foule de noyaux qui satisfont les conditions (i) à (iv) ; il nous suffit de les emprunter à la littérature sur l'estimation non paramétrique de la densité. Par exemple, les noyaux de Parzen et de Tukey-Hanning (TH) (discutés dans Priestley 1981) conviennent, tandis que ceux de Bartlett et Daniell ne sont pas appropriés, puisque (iv) n'est pas vérifiée. En général, il suffit d'utiliser (8) pour déterminer les transformées de Fourier inverses. À la présente section, nous examinons deux exemples, les noyaux quartique et TH. Leur avantage est qu'il est facile de calculer leurs dérivées première et seconde, et que l'on peut obtenir une forme explicite de leurs transformées de Fourier inverses.

Exemple 1 : *Noyau quartique*

Nous commençons par considérer un noyau polynomial de degré quatre, c'est-à-dire quartique. L'imposition de toutes les contraintes (i) à (iv) donne la forme suivante :

$$\begin{aligned}
A(\lambda) \;&=\; \frac{15}{8\pi^4}\,(\lambda^4 \,-\, 2\pi^2\lambda^2 \,+\, \pi^4),\\
\dot{A}(\lambda) \;&=\; \frac{15}{8\pi^4}\,(4\pi^3 \,-\, 4\pi^2\lambda), \text{ et}\\
\ddot{A}(\lambda) \;&=\; \frac{15}{8\pi^4}\,(12\pi^2 \,-\, 4\pi^2).
\end{aligned}$$

En prenant la transformée de Fourier inverse des noyaux de la pente et de la convexité (et leurs carrés), nous obtenons

$$\gamma_{\dot{A}}(h) = \frac{15i}{\pi^5}\left(\frac{\pi^2 \sin \pi h}{h^2} + \frac{3\pi \cos \pi h}{h^3} - \frac{3 \sin \pi h}{h^4}\right),$$

$$\gamma_{\ddot{A}}(h) = \frac{15}{\pi^5}\left(\frac{\pi^2 \sin \pi h}{h} + \frac{3\pi \cos \pi h}{h^2} - \frac{3 \sin \pi h}{h^3}\right),$$

$$\gamma_{\dot{A}^2}(h) = \frac{225}{\pi^9}\left(-\frac{2\pi^4 \sin \pi h}{h^3} - \frac{18\pi^3 \cos \pi h}{h^4} + \frac{78\pi^2 \sin \pi h}{h^5} + \frac{180\pi \cos \pi h}{h^6} - \frac{18 \sin \pi h}{h^7}\right),$$

$$\gamma_{\ddot{A}^2}(h) = \frac{225}{4\pi^{11}}\left(\frac{\pi^4 \sin \pi h}{h} + \frac{6\pi^3 \cos \pi h}{h^2} - \frac{24\pi^2 \sin \pi h}{h^3} - \frac{54\pi \cos \pi h}{h^4} + \frac{54 \sin \pi h}{h^5}\right),$$

auxquels nous appliquons (8) et obtenons

$$\gamma_{\dot{A}_{\beta,\mu}}(h) = \frac{30i}{\beta}\exp\{ih\mu\}\left(\frac{\sin k}{k^2} + \frac{3\cos k}{k^3} - \frac{3\sin k}{k^4}\right),$$

$$\gamma_{\ddot{A}_{\beta,\mu}}(h) = \frac{30}{\beta^2\pi}\exp\{ih\mu\}\left(\frac{\sin k}{k} + \frac{3\cos k}{k^2} - \frac{3\sin k}{k^3}\right),$$

$$\gamma_{\dot{A}^2_{\beta,\mu}}(h) = \frac{1\,800\pi}{\beta^3}\exp\{ih\mu\}\left(\frac{2\sin k}{k^3} + \frac{18\cos k}{k^4} - \frac{78\sin k}{k^5} - \frac{180\cos k}{k^6} + \frac{180\sin k}{k^7}\right), \text{ et}$$

$$\gamma_{\ddot{A}^2_{\beta,\mu}}(h) = \frac{1\,800}{\beta^5\pi}\exp\{ih\mu\}\left(\frac{\sin k}{k} + \frac{6\cos k}{k^2} - \frac{24\sin k}{k^3} - \frac{54\cos k}{k^4} + \frac{54\sin k}{k^5}\right),$$

où $k = h\beta/2$. Notons que $\gamma_{\dot{A}^2_{\beta,\mu}}(0) = 240\pi/(7\beta^3)$ et $\gamma_{\ddot{A}^2_{\beta,\mu}}(0) = 360/(\beta^5\pi)$ découlent de l'application de la règle de L'Hôpital. Ces formules nous permettent de construire les matrices de Toeplitz appropriées pour le diagnostic (comme il est exposé plus bas, à la section 3.1, il suffit de considérer la partie réelle de ces séquences).

Exemple 2 : *Noyau TH*

Nous pouvons obtenir une forme semblable au noyau quartique en utilisant une fonction cosinus. Le choix qui suit satisfait toutes les conditions énoncées sur un noyau :

$$A(\lambda) = \frac{1}{2\pi}(1 + \cos\lambda),$$

$$\dot{A}(\lambda) = \frac{1}{2\pi}(-\sin\lambda), \text{ et}$$

$$\ddot{A}(\lambda) = \frac{1}{2\pi}(-\cos\lambda).$$

Cette fonction est identique à la fenêtre des décalages de Tukey-Hanning, quoique ici, nous l'appliquons comme une fenêtre spectrale (voir Priestley 1981). Dans la suite de l'exposé, nous la désignerons par noyau TH. En prenant la transformée de Fourier inverse des noyaux de la pente et de la convexité (et leurs carrés), nous obtenons

$$\gamma_{\dot{A}}(h) = \frac{i}{4\pi^2}\left(\frac{\sin\pi(h+1)}{h+1} - \frac{\sin\pi(h-1)}{h-1}\right),$$

$$\gamma_{\ddot{A}}(h) = -\frac{1}{4\pi^2}\left(\frac{\sin\pi(h+1)}{h+1} + \frac{\sin\pi(h-1)}{h-1}\right),$$

$$\gamma_{\dot{A}^2}(h) = \frac{1}{16\pi^3}\left(\frac{2\sin\pi h}{h} - \frac{\sin\pi(h+2)}{h+2} - \frac{\sin\pi(h-2)}{h-2}\right), \text{ et}$$

$$\gamma_{\ddot{A}^2}(h) = \frac{1}{16\pi^3}\left(\frac{2\sin\pi h}{h} + \frac{\sin\pi(h+2)}{h+2} + \frac{\sin\pi(h-2)}{h-2}\right).$$

L'application de (8) donne alors

$$\gamma_{\dot{A}_{\beta,\mu}}(h) = \frac{i}{2\beta}\exp\{ih\mu\}\left(\frac{\sin(k+\pi)}{k+\pi} - \frac{\sin(k-\pi)}{k-\pi}\right),$$

$$\gamma_{\ddot{A}_{\beta,\mu}}(h) = -\frac{\pi}{\beta^2}\exp\{ih\mu\}\left(\frac{\sin(k+\pi)}{k+\pi} + \frac{\sin(k-\pi)}{k-\pi}\right),$$

$$\gamma_{\dot{A}^2_{\beta,\mu}}(h) = \frac{\pi}{2\beta^3}\exp\{ih\mu\}\left(\frac{2\sin k}{k} - \frac{\sin(k+2\pi)}{k+2\pi} - \frac{\sin(k-2\pi)}{k-2\pi}\right), \text{ et}$$

$$\gamma_{\ddot{A}^2_{\beta,\mu}}(h) = \frac{2\pi^3}{\beta^5}\exp\{ih\mu\}\left(\frac{2\sin k}{k} - \frac{\sin(k+2\pi)}{k+2\pi} + \frac{\sin(k-2\pi)}{k-2\pi}\right),$$

où $k = h\beta/2$. Notons que $\gamma_{\dot{A}^2_{\beta,\mu}}(0) = \pi/\beta^3$ et $\gamma_{\ddot{A}^2_{\beta,\mu}}(0) = 4\pi^3/\beta^5$ découlent de l'application de la règle de L'Hôpital (en utilisant la convention que $\sin(0)/0 = 1$). Ces formules nous permettent de construire les matrices de Toeplitz appropriées pour le diagnostic (de nouveau, comme nous l'exposons plus bas à la section 3.1, il suffit de considérer la partie réelle de ces séquences).

3. Méthodologie statistique

Évidemment, d'habitude, nous ne connaissons pas le spectre f si bien qu'il faut généralement produire des estimations d'après les données. À la présente section, nous décrivons des estimations statistiques des mesures de la pente et de la convexité qui sont convergentes et simples à calculer dans le domaine temporel. Sous certaines hypothèses supplémentaires faibles, ces estimations sont asymptotiquement normales, ce qui sera avantageux quand nous effectuerons les tests d'hypothèse. À la section 3.1, nous définissons les estimations statistiques et discutons de leurs

propriétés asymptotiques. À la section 3.2, nous discutons de l'application à la détection individuelle des pics et à la section 3.3, nous donnons une extension à leur détection conjointe, qui facilite une application importante dans la désaisonnalisation. À la section 3.4, nous discutons des extensions à des données non stationnaires avec tendance.

3.1 Estimateurs de la pente et de la convexité

Nous commençons par noter que la forme quadratique (pour toute fonction intégrable g) est

$$\frac{1}{n} X' \Sigma(g) X = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda)\, I(\lambda)\, d\lambda,$$

où I désigne le périodogramme. Bien que ce dernier soit habituellement défini par les fréquences de Fourier ($2\pi j/n$; $j = 1, \ldots, \lfloor n/2 \rfloor$), nous le définissons comme une bande continue de fréquences de la façon suivante

$$I(\lambda) = \frac{1}{n} \left| \sum_{t=1}^{n} X_t\, e^{-it\lambda} \right|^2 = \sum_{h=1-n}^{n-1} R(h)\, e^{-it\lambda},\ \lambda \in [-\pi, \pi] \qquad (9)$$

avec $R(h)$ égal à la fonction d'autocovariance (non centrée) d'échantillon. Cela donne un moyen élégant de passer du domaine temporel au domaine fréquentiel qui est bien décrit dans la littérature sur les séries chronologiques (voir Taniguchi et Kakizawa 2000). En outre, ce genre d'intégrales du périodogramme sont généralement convergentes, c'est-à-dire que $\theta_g(I) \xrightarrow{p.s.} \theta_g(f)$ quand $n \to \infty$, sous les conditions faibles discutées plus bas (notons que la non-convergence du périodogramme est résolue par l'agrégation spectrale contre la fonction g, comme il est montré en annexe). Par conséquent, nous obtenons des estimations statistiques des mesures de la pente et de la convexité f en utilisant une approche de substitution, c'est-à-dire en remplaçant simplement f par I dans $\theta_{A_{\beta,\mu}}$. En particulier,

$$\hat{\theta}_{A_{\beta,\mu}}(\dot{f}) = -\theta_{\dot{A}_{\beta,\mu}}(I) = -\frac{1}{n} X' \Sigma(\dot{A}_{\beta,\mu}) X, \text{ et}$$

$$\hat{\theta}_{A_{\beta,\mu}}(\ddot{f}) = \theta_{\ddot{A}_{\beta,\mu}}(I) = \frac{1}{n} X' \Sigma(\ddot{A}_{\beta,\mu}) X. \qquad (10)$$

Cette définition comprend l'utilisation de (5), qui explique le signe négatif de la mesure de la pente. Afin de calculer l'estimation, nous utilisons la représentation dans le domaine temporel (exprimée sous une forme quadratique). Cette représentation est commode parce que nous devons uniquement déterminer une longueur appropriée pour les séquences $\gamma_{\dot{A}_{\beta,\mu}}(h)$ et $\gamma_{\ddot{A}_{\beta,\mu}}(h)$, former les matrices de Toeplitz $\Sigma(\dot{A}_{\beta,\mu})$ et $\Sigma(\ddot{A}_{\beta,\mu})$, puis calculer les formes quadratiques. Notons que les transformées de Fourier inverses de $\dot{A}_{\beta,\mu}$ et $\ddot{A}_{\beta,\mu}$ ne doivent être déterminées qu'une seule fois (voir la section 2.3 pour des exemples explicites) et peuvent être produites d'avance, puis appliquées de manière répétée à de nombreuses séries chronologiques différentes.

Afin de calculer la représentation dans le domaine temporel des mesures de la pente et de la convexité données par (10), nous utilisons (8) ; par exemple voir les formules dans les exemples 1 et 2. Évidemment, il résultera en général que $\Sigma(\dot{A}_{\beta,\mu})$ et $\Sigma(\ddot{A}_{\beta,\mu})$ sont complexes. Cependant, même si $\Sigma(g)$ (où g peut être $A_{\beta,\mu}$, $\dot{A}_{\beta,\mu}$ ou $\ddot{A}_{\beta,\mu}$) est une matrice de Toeplitz complexe, $X'\Sigma(g)X$ sera toujours réel. Partant de (8), il est facile de voir que $\Sigma(g) = M + iN$ où M est réelle, symétrique et Toeplitz, et que N est réelle, antisymétrique et Toeplitz. D'où $X'NX = 0$ pour tout vecteur X, de sorte que $X'\Sigma(g)X = X'MX$. Par conséquent, pour calculer les mesures statistiques de la pente et de la convexité, nous pouvons prendre la partie réelle de $\gamma_A(h)$ dans (8).

Ces estimations statistiques sont non seulement convergentes, mais aussi asymptotiquement normales sous certaines conditions supplémentaires (discutées à l'annexe). Cependant, pour construire une normalisation appropriée, il sera nécessaire d'estimer leur variation. La variance asymptotique de $\theta_g(I)$ est $\theta_{g^2}(f^2)$ (si g est appuyée sur $[0, \pi]$), qui peut être estimée de manière convergente par la voie de $\theta_{g^2}(I^2)/2$. (Le facteur de 2 est requis, puisque l'intégrale de I^2 tend vers l'intégrale correspondante de $2f^2$ – voir Chiu (1988)). Nous pouvons représenter cela dans le domaine temporel de la façon suivante. Soit $R = \{R(1-n), \ldots, R(0), \ldots, R(n-1)\}'$ un vecteur de dimension $2n-1$ des autocovariances d'échantillon et soit $\Sigma(g^2)$ de dimension $2n-1$ dans la formule suivante : $R'\Sigma(g^2)R/2 = \theta_{g^2}(I^2)/2$. Cette relation peut être vérifiée facilement en utilisant (9). Donc, nous normaliserons $\theta_g(I)$ par la racine carrée de $\theta_{g^2}(I^2)/2$. D'où une mesure statistique normalisée de la pente et de la convexité sont données par

$$-\psi_{\dot{A}_{\beta,\mu}}(I) = -\frac{\theta_{\dot{A}_{\beta,\mu}}(I)}{\sqrt{\theta_{\dot{A}^2_{\beta,\mu}}(I^2)/2}} = -\frac{1}{n} \frac{X' \Sigma(\dot{A}_{\beta,\mu}) X}{\sqrt{R' \Sigma(\dot{A}^2_{\beta,\mu}) R/2}}$$

et

$$\psi_{\ddot{A}_{\beta,\mu}}(I) = -\frac{\theta_{\ddot{A}_{\beta,\mu}}(I)}{\sqrt{\theta_{\ddot{A}^2_{\beta,\mu}}(I^2)/2}} = \frac{1}{n} \frac{X' \Sigma(\ddot{A}_{\beta,\mu}) X}{\sqrt{R' \Sigma(\ddot{A}^2_{\beta,\mu}) R/2}},$$

où les dimensions des matrices Σ sont soit n soit $2n-1$ comme il est approprié. Les propriétés asymptotiques de $\psi_{\dot{A}_{\beta,\mu}}(I)$ et de $\psi_{\ddot{A}_{\beta,\mu}}(I)$ sont discutées à l'annexe. En résumé, $-\sqrt{n}\,\psi_{\dot{A}_{\beta,\mu}}(I)$ et $\sqrt{n}\,\psi_{\ddot{A}_{\beta,\mu}}(I)$ sont toutes deux marginalement asymptotiquement de loi $N(0, 1)$ sous $H_0^{(1)}$ et $H_0^{(2)}$, respectivement, et sous les hypothèses discutées à l'annexe. Les simulations indiquent que la normalisation de la variance ne converge que lentement et que sa corrélation avec le numérateur cause un certain

degré de non-normalité dans les petits échantillons. D'après l'histogramme de la distribution simulée sous une hypothèse nulle de bruit blanc gaussien avec n = 360 et 10 000 répétitions (figure 1), la concordance avec la loi normale est proche, sauf aux extrémités dans les queues. À la section 4, nous examinons ce comportement plus en détail au moyen d'études par simulation.

3.2 Applications au test de la présence d'un seul pic

Nous considérons maintenant l'application au test de la présence d'un pic. Rappelons que nous avons une hypothèse nulle initiale $H_0^{(1)}$ qui ne doit pas être rejetée pour pouvoir poursuivre. Cela peut s'interpréter comme signifiant qu'il n'existe pas de preuve suffisamment convaincante pour conclure que la dérivée première (pente) de la densité spectrale diffère de manière significative de zéro. Or, nous savons que $-\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$ est asymptotiquement de loi $N(0,\,1)$ sous $H_0^{(1)}$ ainsi que les hypothèses discutées à l'annexe. Si nous supposons en outre qu'une valeur suffisamment petite x est obtenue pour la statistique de test, nous ne pourrons rejeter $H_0^{(1)}$ avec aucun degré de confiance. Dans ce cas, nous pouvons considérer l'hypothèse $H_0^{(2)}$, que nous cherchons à rejeter, ce que nous testons au moyen de $\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$. Bien que $-\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$ et $\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$ soient asymptotiquement corrélées (voir le théorème 1 de l'annexe), nous considérerons les tests de la pente et de la convexité comme s'ils étaient exécutés séparément. (Cette corrélation peut être estimée et utilisée pour déterminer la distribution du diagnostic de convexité conditionnellement au diagnostic de pente ; cependant, l'interprétation des valeurs p devient confuse. Pour simplifier, nous traitons les tests séparément, un à la fois, et nous ne tenons pas compte explicitement de la corrélation.) Notre procédure de test s'exécute alors de la façon suivante :

1. Faire le test bilatéral de $H_0^{(1)}$ en utilisant $-\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$.
2. Poser que p est la valeur p associée avec la première valeur de la statistique de test $x = -\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$, avec x et p reliées par $p = 2\Phi(-|x|)$.
3. Si $p > 0{,}05$ (ou un autre seuil de tolérance préétabli), poursuivre ; sinon, conclure qu'aucun pic n'est présent.
4. Exécuter le test unilatéral inférieur de $H_0^{(2)}$ en utilisant $\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I)$.
5. Rejeter $H_0^{(2)}$ et conclure qu'il existe un pic si $\sqrt{n}\,\psi_{\ddot{A}_{\beta,\,\mu}}(I) < \Phi^{-1}(\alpha)$, où α est le niveau du test de convexité.

3.3 Test conjoint de la présence de plusieurs pics : application à la désaisonnalisation

Considérons maintenant la situation où nous voulons tester simultanément la présence de plusieurs pics spectraux. Il est clair que nous pourrions concevoir un noyau comportant plusieurs nœuds, un à chaque pic, mais il serait simplement la somme de plusieurs diagnostics de pic spectral individuels. L'inconvénient serait qu'un pic spectral significatif à un endroit pourrait annuler un creux spectral significatif ailleurs. Par conséquent, nous préférons un test permettant d'examiner un ensemble de diagnostics spectraux dans un paradigme de tests multiples.

Par exemple, considérons le contexte de la détection de la présence de pics spectraux dans des données désaisonnalisées. Six pics saisonniers présentent un intérêt, mais nous devons nous limiter à cinq d'entre eux à cause de problèmes d'alias, c'est-à-dire de repliement du spectre (le pic de fréquence π ne peut pas être identifié). Si un ou plusieurs pics spectraux sont significatifs, nous devons rejeter notre procédure de désaisonnalisation (puisqu'elle n'a pas réussi à éliminer tous les pics) ; par conséquent, nous sommes dans des conditions de tests multiples et nous utiliserons une méthode permettant de contrôler le taux d'erreur de type I global (FWER pour *familywise error rate*) proposé par Hochberg (1988) et décrit dans Benjamini et Hochberg (page 294, 1995). En nous limitant à la question de la convexité, nous avons les hypothèses nulles $H_0^{(2)}$ pour chacune des cinq fréquences saisonnières. Dans nos conditions, la procédure de Hochberg (1988) consiste à calculer les valeurs p pour le test de convexité à chacune de ces cinq fréquences saisonnières et à les classer selon $p_{(1)} \le p_{(2)} \le p_{(3)} \le p_{(4)} \le p_{(5)}$, avec les hypothèses nulles correspondantes désignées par $H_{(i)}$. Pour un taux d'erreur de type I global spécifié de niveau α (par exemple α = 0,05), soit k la valeur de i la plus grande pour laquelle $p_{(i)} \le i/(6 - i)\,\alpha$; alors, nous rejetons toutes les $H_{(i)}$ pour $i \le k$.

En utilisant cette procédure, nous devrions commettre des erreurs de type I – c'est-à-dire déterminer la présence d'au moins une fréquence saisonnière ayant une convexité négative alors qu'il n'en existe aucune – une proportion de fois environ égale à α (si nous nous limitions à examiner $H_0^{(2)}$, l'hypothèse de convexité). L'avantage de l'approche du taux d'erreur de type I global FWER de Hochberg (H-FWER) est qu'elle améliore spectaculairement la puissance statistique comparativement aux autres méthodes. La validité de cette méthode requiert l'indépendance des statistiques de test prises en considération et, pour cette raison, nous prenons cinq noyaux A_1, ..., A_5 – centrés sur les fréquences saisonnières $\pi/6$, ..., $5\pi/6$ respectivement – qui ont un support disjoint. Alors, le théorème 1 peut être généralisé pour obtenir l'indépendance asymptotique de cinq

statistiques de test de convexité (voir la discussion qui suit le théorème 1 à l'annexe). Naturellement, nous exécutons aussi cinq tests distincts de la pente à chaque fréquence saisonnière, où il faut que nous n'aboutissions à un rejet dans aucun cas pour pouvoir poursuivre

Enfin, nous notons qu'en pratique, une désaisonnalisation est rarement rejetée sur la base d'une masse spectrale significative à la cinquième fréquence saisonnière de $5\pi/6$ (Findley 2006). Cela tient en partie à la difficulté d'attribuer une interprétation à cette fréquence. Par conséquent, le praticien pourrait être plus intéressé par un « test à quatre pics » qui est axé sur les quatre premières fréquences saisonnières ; ce test s'obtient par une modification évidente de la procédure H-FWER décrite plus haut.

3.4 Extention à des données non stationnaires

La méthodologie que nous venons d'exposer s'appuie sur l'hypothèse que les données sont un échantillon d'un processus stationnaire. Cependant, dans le contexte de la désaisonnalisation, les données désaisonnalisées sont habituellement intégrées une fois ou deux fois. Dans ce cas, on différentierait les données désaisonnalisées une ou deux fois avant d'appliquer les diagnostics. Or, les opérateurs de différentiation $1 - B$ et $(1 - B)^2$ appliqués sont essentiellement des filtres passe-haut, qui en principe atténuent les pics spectraux résiduels proches de la fréquence zéro (en particulier, la première fréquence saisonnière à $\pi/6$). Donc, il serait peut-être souhaitable d'appliquer le diagnostic au pseudo-spectre, ce qui peut se faire si le support du noyau est borné de façon à ne pas contenir les pôles de la densité spectrale.

Supposons que les données observées sont maintenant $Y_{1-d}, ..., Y_n$ pour d l'ordre de différentiation de la tendance (donc habituellement $d = 1$ ou 2). Lorsque les données observées sont différentiées, nous obtenons l'échantillon X, qui est strictement stationnaire. La pseudo-densité spectrale du processus $\{Y_t\}$ est $g(\lambda) = f(\lambda)\,|1-e^{-i\lambda}|^{-2d}$, où f est le spectre de $\{X_t\}$. Ce pseudo-spectre pourrait être estimé au moyen de $\hat{g}(\lambda) = I(\lambda)\,|1-e^{-i\lambda}|^{-2d}$, où I est le périodogramme de X comme auparavant ; il s'agit de l'approche de recoloration de Nerlove (1964). Alors, $\theta_A(g)$ est bien défini à condition que $A(\lambda)\,|1-e^{-i\lambda}|^{-2d}$ soit une fonction intégrable ; essentiellement, nous devons nous assurer que la fréquence zéro est exclue du support du noyau A. Puisqu'en pratique A est centré autour des fréquences saisonnières, nous pouvons facilement trouver le moyen d'appliquer cette condition. L'estimateur correspondant est alors

$$\hat{\theta}_A(\ddot{g}) = \theta_{\ddot{A}b}(I),$$

où $b(\lambda) = |1 - e^{-i\lambda}|^{-2d}$. L'estimateur est bien défini si $\ddot{A}b$ est intégrable ; en outre, les propriétés asymptotiques discutées à l'annexe pour le cas stationnaire s'étendent à ce cas également, à condition que $\ddot{A}b$ soit borné.

Cette extension pourrait paraître plus séduisante à certains chercheurs. Cependant, son prix est que la transformée de Fourier inverse de $\ddot{A}b$ doit être déterminé, ce qui requiert un travail mathématique supplémentaire. Dans les études par simulation et les illustrations des données de la section 4, nous différentions les données désaisonnalisées, mais nous n'appliquons pas le facteur de correction b dans le noyau.

4. Études empiriques

Ayant développé les aspects théoriques du diagnostic spectral, nous nous tournons maintenant vers sa performance en pratique. Pour commencer, nous présentons certains résultats obtenus par simulation, qui donnent une idée des propriétés de taille et de puissance de la statistique de test dans les échantillons finis. Puis, nous étudions empiriquement la taille et la puissance en appliquant les diagnostics spectraux à une suite de 130 séries chronologiques (65 séries du U.S. Census Bureau et 65 séries de l'OCDE) ; nous considérons les séries originales ainsi que les séries désaisonnalisées et nous comparons les résultats à ceux du diagnostic de signification visuelle et aux diagnostics de contrôle de la qualité M7 et M8 de X-12-ARIMA (U.S. Census Bureau 2002). D'autres études empiriques sont décrites dans Evans, Holan et McElroy (2006).

4.1 Étude par simulation

Afin d'évaluer la performance de nos diagnostics, nous avons exécuté plusieurs simulations. Le premier ensemble de simulations a pour but d'examiner la taille (niveau) pour le diagnostic de la présence d'un pic unique. Pour cette simulation, nous considérons les diagnostics de pente et de convexité séparément. Même si, en pratique, quand nous considérons le diagnostic de pente, nous souhaitons ne pas rejeter l'hypothèse nulle $H_0^{(1)}$, ici nous voulons étudier empiriquement les propriétés distributionnelles et nous imposons donc la définition habituelle de taille pour cette étude. Donc, nous simulons un bruit blanc gaussien qui satisfait les hypothèses du théorème 1, ainsi que $\psi_{\ddot{A}}(I) = \psi_{\ddot{A}}(I) = 0$, de sorte que $H_0^{(1)}$ et $H_0^{(2)}$ sont vraies. Naturellement, il existe de nombreux processus pour lesquels $H_0^{(1)}$ et $H_0^{(2)}$ sont vraies simultanément – par exemple, tout processus dont la densité spectrale est localement uniforme ; cependant, en raison de considérations asymptotiques, il suffit de considérer le bruit blanc. Pour une (grande) taille d'échantillon de $n = 360$, en utilisant le noyau TH avec $\mu = \pi/6$ et $\beta = \pi/6$ (ce qui correspond à un noyau centré sur l'intervalle $[0, \pi/6]$), 10 000 répétitions

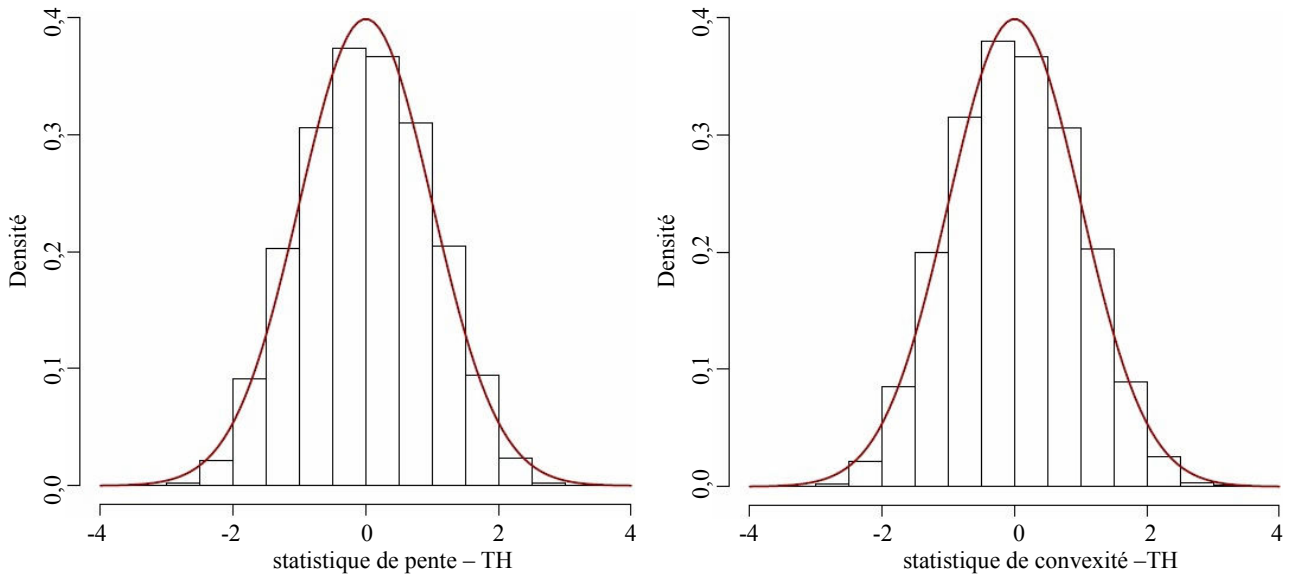
produisent une distribution empirique des diagnostics normalisés, $\psi_A(I)$ et $\psi_{\ddot{A}}(I)$, dont les histogrammes sont présentés à la figure 1. À partir de maintenant, représentons par δ et α les niveaux associés aux tests de pente et de convexité, respectivement. Notons que, dans ce cas, nous définissons le niveau comme étant la probabilité de rejeter $H_0^{(j)}$ $(j=1, 2)$ quand $H_0^{(j)}$ est vraie. Même si, en pratique, dans le cas de l'hypothèse de pente, nous souhaitons ne pas aboutir au rejet, nous suivons la définition stricte du niveau et supposons (pour les besoins de cette simulation) que l'hypothèse nulle $H_0^{(1)}$ pour la pente est vérifiée. De même, l'hypothèse nulle pour la convexité est $H_0^{(2)}$. Les hypothèses de pente et de convexité sont évaluées indépendamment. Le tableau 1 résume les résultats en utilisant les deux noyaux décrits à la section 2, pour diverses tailles d'échantillons ; les niveaux δ, α indiqués s'appliquent au seuil nominal de 5 %. En outre, d'autres choix de μ et β (non présentés ici) ont donné des résultats semblables. Comme l'illustre la présente étude, dans les échantillons de petite taille, nous observons une asymétrie de la distribution qui semble être due à la corrélation entre $\theta_A(I)$ et $\theta_{A^2}(I^2)$. En outre, notons que dans le cas du test de convexité, la taille est plus grande pour le noyau quartique que pour le noyau TH.

Ensuite, considérons la puissance empirique de notre diagnostic de la présence d'un seul pic. Dans ces conditions, nous évaluons la puissance basée sur un test conjoint de la pente et de la convexité. Plus précisément, nous souhaitons ne pas rejeter $H_0^{(1)}$, tout en rejetant simultanément $H_0^{(2)}$, au niveau $\delta = \alpha = 0{,}05$, et donc identifier correctement les pics spectraux. Puisque notre hypothèse nulle composite est qu'il n'existe pas de pic, l'hypothèse alternative inclut les processus tels que le processus $AR(2)$ donné par

$$(1 - 2\rho\cos\omega B + \rho^2 B^2) X_t = \varepsilon_t \qquad (11)$$

avec la variance de bruit blanc σ^2, associée à une fréquence fixe donnée $\omega \in [0, \pi]$. Le spectre associé au processus (11) est donné par $f(\lambda) = \sigma^2 |1 - 2\rho\cos\omega e^{-i\lambda} + \rho^2 e^{-2i\lambda}|^{-2}$, qui est maximisé à $\lambda_0 = \cos^{-1}(\cos\omega(1 + \rho^2)/2\rho)$. Par conséquent, nous pouvons explorer la puissance d'une procédure de dépistage de pic en procédant à une simulation selon (11) avec divers choix de ρ, ω et σ. Le tableau 2 donne les résultats de 10 000 simulations, pour diverses tailles d'échantillons, d'après le modèle cyclique $AR(2)$ donné en (11) avec un pic à $\mu = \pi/6$ et la largeur de bande fixée à $\beta = \pi/6$. La force du pic est paramétrisée au moyen de ρ, que nous faisons varier de 0,85 à 0,95 ; il est clair que $H_0^{(1)}$ et $H_a^{(2)}$ sont toutes deux vraies pour ce modèle. Autrement dit, il existe des pics spectraux, de différentes hauteurs, à $\lambda = \pi/6$. Ce modèle AR a été choisi parce qu'il donne une paramétrisation commode de la localisation et de la forme du pic spectral. De plus, ce choix de β est compatible avec les conditions de désaisonnalisation, car il fournit la largeur de fenêtre maximale tout en évitant le chevauchement des pics spectraux. Comme prévu, la puissance de notre diagnostic augmente avec la taille de l'échantillon et la pointicité, variant de 0,227 (noyau quartique) dans les petits échantillons possédant un faible pic spectral à $\approx 0{,}95$ (noyau TH) dans les échantillons plus grands ayant un pic spectral plus prononcé (voir le tableau 2). Notons que, dans cette procédure, la valeur de la variance d'innovation est fixée à 1, mais qu'elle est négligeable en raison de la normalisation du diagnostic. En résumé, les noyaux quartique et TH possèdent tous deux des propriétés de taille et de puissance convenables. Généralement, le noyau quartique semble avoir une taille et une puissance supérieures, si bien qu'on le préférerait pour des spectres de cette forme (notons que la puissance plus faible du noyau TH est due en partie au fait que sa taille est trop petite). En outre, il semble que les valeurs plus faibles de β (résultats non présentés) requièrent une plus grande taille d'échantillon ; un β plus petit correspond à des « conditions d'observation » plus fines du pic spectral, qui nécessiterait une plus grande quantité de données pour traiter la résolution.

Bien que le scénario du test de la présence de pics individuels constitue le fondement de notre cadre de test conjoint, comme nous l'avons mentionné, ce cadre représente une méthodologie importante pour les applications aux statistiques fédérales. L'application vraiment importante est l'exploration de la saisonnalité résiduelle en vue de déterminer si la désaisonnalisation est efficace. Il est donc particulièrement intéressant de savoir comment notre approche de tests multiples se comporte dans la simulation. Par conséquent, pour étudier la taille et la puissance associées à notre test conjoint, nous avons simulé 10 000 répétitions à partir d'un processus de bruit blanc gaussien et d'un modèle $AR(25)$ obtenu par un ajustement à la série sur l'emploi courant (hommes occupés, de 16 à 19 ans). L'objectif de notre étude de puissance était de construire un processus $AR(p)$ (à cause de la facilité de son utilisation dans les simulations et de ses propriétés théoriques désirables en tant qu'estimateur spectral paramétrique – voir Parzen 1983) avec des pics spectraux (stationnaires) qui sont réalistes, ou qui s'approchent de ce qu'on pourrait trouver en pratique. Donc, nous obtenons notre modèle $AR(25)$ – ajusté par la méthode du maximum de vraisemblance en utilisant le critère AIC – qui possède la même dynamique saisonnière (comportement spectral local) que la série sur l'emploi courant (CES pour *Current Employment Series*).



(a) (b)

Figure 1 Histogramme de la distribution de $-\sqrt{n}\,\psi_{\hat{A}_{\beta,\mu}}(I)$ (a) et de $\sqrt{n}\,\psi_{\hat{A}_{\beta,\mu}}(I)$ (b) sous une hypothèse nulle de bruit blanc gaussien en utilisant le noyau TH. La taille d'échantillon est $n = 360$ avec 10 000 répétitions.

Tableau 1
Résultats de la simulation de la taille pour le diagnostic d'un pic unique. Ici $\mu = \beta = \pi/6$ et 10 000 répétitions sont utilisées. Les diagnostics de pente et de convexité ont été étudiés séparément pour les noyaux quartique et TH

Taille pour un pic unique $\mu = \beta = \pi/6$												
	Pente						Convexité					
	Noyau quartique			Noyau TH			Noyau quartique			Noyau TH		
n	Moyenne	Écart-type	Niveau δ	Moyenne	Écart-type	Niveau δ	Moyenne	Écart-type	Niveau α	Moyenne	Écart-type	Niveau α
120	0,003	0,903	0,007	-0,011	0,903	0,008	-0,065	0,852	0,032	0,025	0,888	0,018
144	-0,004	0,920	0,014	-0,011	0,927	0,015	-0,077	0,882	0,042	0,006	0,892	0,025
180	-0,003	0,942	0,022	0,002	0,920	0,017	-0,071	0,892	0,043	0,005	0,902	0,028
288	0,003	0,954	0,027	-0,002	0,950	0,025	-0,072	0,921	0,051	-0,006	0,926	0,033
360	0,003	0,962	0,032	-0,009	0,954	0,031	-0,056	0,922	0,051	0,006	0,951	0,040

Tableau 2
Résultats de la simulation de la puissance pour le diagnostic d'un pic unique. Ici $\mu = \beta = \pi/6$ et 10 000 répétitions sont utilisées. L'hypothèse alternative est donnée par le modèle $AR(2)$ défini par (11). Les diagnostics de pente et de convexité ont été étudiés simultanément pour les noyaux quartique et TH en utilisant $\delta = \alpha = 0{,}05$ pour les deux tests (voir la section 4.1)

Puissance pour un pic unique $\mu = \beta = \pi/6$ - $(\delta, \alpha) = (0{,}05, 0{,}05)$						
	Noyau quartique			Noyau TH		
n	$\rho = 0{,}85$	$\rho = 0{,}90$	$\rho = 0{,}95$	$\rho = 0{,}85$	$\rho = 0{,}90$	$\rho = 0{,}95$
120	0,227	0,438	0,758	0,147	0,335	0,670
144	0,287	0,532	0,856	0,208	0,431	0,799
180	0,354	0,643	0,923	0,272	0,567	0,901
288	0,447	0,755	0,949	0,372	0,706	0,950
360	0,601	0,872	0,937	0,537	0,859	0,948

Pour évaluer la taille, nous avons considéré un test basé sur la convexité seulement ainsi qu'un test basé simultanément sur la pente et la convexité. Les tests portant sur la convexité seulement (C) sont exécutés aux seuils de signification nominaux α de 0,05 et 0,10, en utilisant la méthode H-FWER pour contrôler le taux d'erreur de type I global. Les tests fondés simultanément sur la pente et la convexité (S, C) ont été exécutés de la façon suivante :

1. Exécuter les tests multiples de convexité, $H_0^{(2)}$, en utilisant la méthode H-FWER pour contrôler le taux d'erreur global au niveau α (qui est égal à 0,05 ou 0,10).
2. Pour tout pic significatif découvert à l'étape 1, exécuter le test de pente individuel, $H_0^{(1)}$, au niveau δ (qui est égal à 0,10 ou 0,25). Notons qu'ici, nous souhaitons ne pas rejeter $H_0^{(1)}$ afin de déclarer n'importe quel « pic » comme étant statistiquement significatif.
3. Déclarer qu'il existe un pic statistiquement significatif si l'étape 1 permet de découvrir une fréquence saisonnière dont la convexité agrégée dans le spectre est significative et que l'étape 2 n'aboutit pas simultanément à la découverte d'une pente agrégée significative pour la fréquence saisonnière correspondante.

Les résultats de cette simulation sont résumés au tableau 3. Un aspect de cette procédure qui nécessite une explication plus approfondie est l'étape 2, où δ (le niveau pour le test de pente) est pris égal à 0,10 et à 0,25. Bien que la partie test de pente de la procédure soit exécutée pour des pics individuels, il semble raisonnable de vouloir être prudent. La question est que, même si certains tests de pente individuels sont rejetés, nous pouvons encore poursuivre dans d'autres cas. Donc la situation observée ici diffère de l'hypothèse classique d'« absence de pics » qui peut être rejetée si un seul pic est découvert. Évidemment, puisque nous effectuons chaque test d'hypothèse de pente sur la base de pics individuels, tout niveau δ supérieur à 0,05 sera considéré comme plus modéré.

Si, dans le cas de la procédure combinée (S, C), nous ne pouvons pas nous attendre à ce que la taille s'approche de la valeur nominale (parce que l'utilisation du test de pente déséquilibre le taux d'erreur de type I), cette taille n'est pas extrêmement exacte non plus dans le cas de l'utilisation du test de convexité seulement (C), comme le montre l'examen du cas $\alpha = 0{,}10$ avec $n = 288, 360$. Ici, pour le noyau quartique, la convexité est de trop grande taille, tandis que dans le cas d'un pic unique, le test de convexité possède une taille exacte (tableau 1) pour ces tailles d'échantillon. Notons que la méthode H-FWER produit uniquement une procédure approximativement correctement dimensionnée ; un autre facteur est que les tests pour les cinq pics ne sont qu'asymptotiquement indépendants. Toutes ces raisons font que la taille empirique observée au tableau 3 diffère quelque peu des niveaux nominaux.

Pour étudier la puissance, nous avons considéré la même procédure en trois étapes décrite plus haut. Cependant, pour cette simulation, nous n'avons envisagé que le test conjoint pente-convexité et examiné quatre paires de niveaux (δ, α), à savoir $(\delta, \alpha) = (0{,}10, 0{,}05), (0{,}25, 0{,}05), (0{,}10, 0{,}10)$ et $(0{,}25, 0{,}10)$. Les résultats de cette simulation (tableau 4) révèlent une puissance énorme, même pour des tailles d'échantillon aussi petites que $n = 120$. Il s'agit d'une propriété extrêmement importante, car $n = 120$ est représentatif des tailles d'échantillon observées en pratique lorsque l'on procède à la désaisonnalisation (par exemple données mensuelles couvrant 10 années). Pour des échantillons de taille $n = 144$, une puissance supérieure à 90 % est réalisée.

Tableau 3
Résultats de la simulation de la taille pour le diagnostic de pics multiples. Ici, 10 000 répétitions ont été utilisées. Le test de convexité a été étudié séparément avec le taux d'erreur de type I global contrôlé à $\alpha = 0{,}05$ et à $\alpha = 0{,}10$ en utilisant la méthode H-FWER. En outre, les diagnostics de pente et de convexité ont été étudiés simultanément en utilisant la méthode H-FWER pour la convexité en contrôlant le taux d'erreur de type I global à $\alpha = 0{,}05$ et à $\alpha = 0{,}10$, tandis que la pente a été évaluée à $\delta = 0{,}5$ et $\delta = 0{,}10$. Les noyaux quartique et TH ont tous deux été utilisés pour les deux tests (voir la section 4.1)

	Taille pour H-FWER pour pics multiples											
	C = 0,05		(S, C) = (0,10, 0,05)		(S, C) = (0,25, 0,5)		C = 0,10		(S, C) = (0,10, 0,10)		(S, C) = (0,25, 0,10)	
n	Quartique	TH	Quartique	TH	Quartique	TH	Quartique	TH	Quartique	TH	Quartique	TH
120	0,006	0,002	0,006	0,002	0,005	0,002	0,076	0,047	0,070	0,044	0,076	0,046
144	0,009	0,002	0,011	0,002	0,008	0,003	0,087	0,053	0,090	0,051	0,086	0,050
180	0,019	0,005	0,020	0,006	0,017	0,005	0,107	0,062	0,097	0,059	0,093	0,057
288	0,031	0,009	0,026	0,008	0,025	0,008	0,117	0,069	0,116	0,069	0,112	0,068
360	0,042	0,012	0,045	0,019	0,035	0,015	0,140	0,087	0,133	0,084	0,129	0,087

Tableau 4
Résultats de la simulation de puissance pour le diagnostic de pics multiples. Ici, 10 000 répétitions ont été utilisées. Les diagnostics de pente et de convexité ont été étudiés simultanément en utilisant la méthode H-FWER pour la convexité en contrôlant le taux global d'erreur de type I à $\alpha = 0{,}05$ et à $\alpha = 0{,}10$. La pente a été évaluée à $\delta = 0{,}25$ et $\delta = 0{,}10$. Les noyaux quartique et TH ont tous deux été utilisés pour les deux tests (voir la section 4.1)

	Puissance pour H-FWER pour pics multiples							
	(0,10, 0,05)		(0,25, 0,05)		(0,10, 0,10)		(0,25, 0,10)	
n	Quartique	TH	Quartique	TH	Quartique	TH	Quartique	TH
120	0,877	0,897	0,860	0,881	0,997	0,997	0,996	0,998
144	0,943	0,952	0,942	0,950	0,999	1,00	0,999	1,00
180	0,989	0,992	0,989	0,992	1,00	1,00	0,999	1,00
288	1,00	1,00	0,999	0,999	1,00	1,00	0,999	1,00
360	1,00	1,00	0,998	0,999	1,00	1,00	0,998	0,999

4.2 Études de cas

Nous avons également examiné 130 séries chronologiques, 65 provenant du U.S. Census Bureau et 65, de l'OCDE. Ces séries comprennent 35 séries américaines sur la fabrication, 10 séries américaines sur le logement, 10 séries américaines sur les importations et exportations et 10 séries américaines sur le commerce de détail, ainsi que 22 séries allemandes, 15 séries pour la zone euro, 11 séries françaises et 17 séries britanniques provenant de l'OCDE et portant sur les secteurs de la fabrication, du commerce de détail, du commerce de gros, du commerce extérieur, du chômage et de l'industrie. Pour chaque série, nous avons calculé les tests de pics saisonniers pour les données brutes (log-transformées) et les données désaisonnalisées (log-transformées) – en utilisant la spécification x11 de X-12-ARIMA – avec le noyau quartique ainsi que le noyau TH. Nous avons appliqué la procédure H-FWER en contrôlant le taux d'erreur de type I global à $\alpha = 0{,}05$ et 0,10 et où le seuil pour les tests de pente était $\delta = 0{,}25$ et $\delta = 0{,}10$ à chaque pic (voir la section 4.1). Notons que $\delta = 0{,}25$ et $\delta = 0{,}10$ produisent des résultats semblables, si bien que, par souci de concision, nous ne présentons ici que les résultats pour $\delta = 0{,}10$. Les résultats pour $\delta = 0{,}25$ peuvent être obtenus sur demande auprès du premier auteur. Pour les données brutes ainsi que désaisonnalisées, nous n'avons effectué qu'une seule différentiation de la tendance (comme cela est le cas pour le diagnostic de signification visuelle décrit plus bas) avant d'appliquer les tests de dépistage des pics saisonniers.

En outre, nous présentons les statistiques M7 et M8, ainsi que les résultats du diagnostic de signification visuelle (SV) avant et après la désaisonnalisation. La statistique de contrôle de la qualité M7 mesure la quantité de saisonnalité stable relativement à la saisonnalité mobile dans la série originale, les valeurs supérieures à 1 indiquant que la saisonnalité dans la série n'est pas identifiable (Lothian et Morry 1978) ; de même, la statistique M8 mesure la taille des fluctuations dans la composante saisonnière, avec une interprétation similaire. Nous avons également considéré le test non paramétrique robuste de Kruskal-Wallis (U.S. Census Bureau en 2002) pour la présence de saisonnalité sous l'hypothèse de stabilité. Le test SV est fondé sur une estimation spectrale $AR(30)$ des séries brute et désaisonnalisée, et est décrit dans Soukup et Findley (1999). Pour les tableaux 5 à 10, chaque entrée de cellule énumère les fréquences saisonnières pour lesquelles un pic significatif a été découvert, avec j correspondant à $\pi j/6$ pour $j = 1, 2, 3, 4, 5$; une entrée $\emptyset$ indique qu'aucun pic n'a été détecté. Pour les diagnostics M7 et M8, seule la valeur est consignée, puisqu'il n'existe aucune valeur p connexe (et qu'ils ne s'appliquent qu'à la série brute).

Les résultats de cette étude empirique sont présentés dans les tableaux 5 à 10. Toutes les valeurs de la statistique de Kruskal-Wallis étaient significatives avec $p = 0{,}000$, si bien qu'elles ne sont pas présentées dans les tableaux. L'ensemble de colonnes qui correspond à l'en-tête « Données originales » peut être considéré comme donnant la puissance empirique (pour chaque sous-ensemble de séries), en supposant que chaque série est effectivement saisonnière et qu'elle présente des pics spectraux saisonniers. Autrement dit, le total de résultats « corrects » donne la proportion de fois que chaque méthode détecte correctement la saisonnalité et cette proportion est donc une approximation grossière de la puissance empirique. Nous présentons aussi le nombre moyen de pics qui ont été détectés, ce qui constitue une mesure empirique de l'efficacité des méthodes (la méthode est d'autant meilleure que le nombre de pics détectés correctement est grand). L'ensemble de colonnes sous l'en-tête « Données désaisonnalisées » donne une taille empirique (pour chaque sous-ensemble de séries), en supposant que la désaisonnalisation a effectivement éliminé les pics spectraux. Il s'agit de considérations grossières, puisque nous ne savons pas vraiment a priori si les données désaisonnalisées ont été corrigées adéquatement.

Le test SV indique que toutes les séries brutes sont saisonnières et que la plupart des séries désaisonnalisées n'ont aucun pic spectral saisonnier ; les diagnostics M7 et M8 donnent des résultats comparables, bien qu'évidemment, ils n'indiquent pas quels pics saisonniers sont présents dans les données brutes. Notre procédure révèle quelques cas (quand $\alpha = 0{,}10$ pour les tests de convexité) où la correction peut être inadéquate, mais ces cas sont compris dans la fourchette de la proportion attendue d'erreurs de type I. Pour les séries brutes, la puissance empirique (c'est-à-dire la proportion totale de résultats corrects) pour notre méthode varie de 0,66 à 0,89, avec une puissance plus élevée pour le niveau $\alpha = 0{,}10$, comme il fallait s'y attendre. Dans de nombreux cas, les pics indiqués sont les mêmes que ceux détectés par le test SV, mais parfois ils sont assez différents. Notons que le nombre moyen de pics détectés pour les séries brutes est habituellement beaucoup plus élevé dans le cas de notre procédure que dans celui de la méthode SV, qui produit souvent un nombre moyen de l'ordre de 3,2. Lorsque le seuil α passe de 0,05 à 0,10, notre méthode produit naturellement un nombre moyen plus élevé de pics détectés ; le test SV ne peut pas être ajusté de cette façon. Inversement, pour les données désaisonnalisées, le nombre moyen de pics détectés a tendance à être inférieur à 1 dans le cas de notre méthode (sauf pour les séries allemandes).

Les résultats sont assez semblables pour les noyaux quartique et TH. Bien que les diagnostics M7, M8 et SV donnent des résultats un peu meilleurs que notre procédure de détection des pics spectraux pour $\alpha = 0{,}10$, il est important de souligner que notre méthode fournit un niveau de détail que les tests M7 et M8 ne peuvent pas reproduire, tandis que le diagnostic SV ne fournit de valeur p pour aucun des pics (M7 ou M8 ne le font pas non plus). Dans l'ensemble, nous pensons que les résultats sont très encourageants et informatifs.

5. Conclusion

Le présent article décrit une approche novatrice de détection statistique des pics spectraux. Le diagnostic de convexité correspond au calcul d'une moyenne du périodogramme pondérée par la dérivée seconde d'un noyau typique, tel que la fenêtre des décalages de Tukey-Hanning. Implicitement, ce type de statistique comprend une comparaison d'une moyenne du périodogramme près d'une fréquence donnée à sa moyenne un peu plus en dehors ; cela découle de la forme générale de $\dot{A}_{\beta,\,\mu}$. Le diagnostic de pente aide à dépister les cas où la convexité est négative, mais où il existe aussi une augmentation/diminution importante dans le spectre. Le fait que la méthode fonctionne effectivement comme prévu est prouvé par les simulations et les résultats d'analyse présentés dans les tableaux 1 à 10.

Pour le scénario de test de détection de pics multiples, nous employons des résultats connus extraits de la littérature sur les tests multiples (c'est-à-dire les applications pour contrôler le taux global d'erreur de type I) pour combiner les valeurs p issues des cinq fréquences saisonnières de manière à accroître considérablement la puissance statistique, comme il est démontré au tableau 4. Malgré un certain écart dans la taille (tableau 3) pour le test de détection de pics multiples, les résultats demeurent relativement utilisables. Sur un lot type de séries saisonnières, le nombre d'erreurs de type I est celui auquel on s'attendrait et la puissance est assez décente (tableaux 5 à 10). Bien que notre méthode se compare assez favorablement aux diagnostics SV, M7 et M8, aucun des diagnostics ne fournit une valeur p et seul le premier permet de discerner quels pics spectraux contribuent au comportement saisonnier. Il s'agit d'un aspect important pour le désaisonnaliseur qui veut savoir non seulement s'il existe une saisonnalité résiduelle, mais aussi à quelles fréquences saisonnières, afin de pouvoir prendre les mesures appropriées pour modifier les filtres de désaisonnalisation (ce qui peut se faire en lissant sur des années supplémentaires, en changeant les filtres saisonniers dans X-11-ARIMA ; ou bien, on pourrait envisager de raccourcir la série. Pour une description des travaux de recherche courants sur une approche fondée sur un modèle pour concevoir des filtres de désaisonnalisation ciblés sur des fréquences saisonnières particulières, voir Aston, Findley, McElroy, Wills et Martin (2007)).

Le choix du noyau a certainement une incidence sur les résultats, quoiqu'en pratique, nous ne constatons que peu de différence entre les noyaux quartique et TH ; ce dernier pourrait être marginalement plus puissant. Évidemment, de nombreux autres noyaux d'usage répandu pourraient aussi être utilisés par un praticien et nous en avons seulement choisi deux qui semblaient intuitifs et simples à mettre en œuvre. Le choix de l'emplacement μ est clairement dicté par la caractérisation de la saisonnalité. Puisque la puissance statistique diminue généralement avec β, nous recommandons systématiquement de prendre la valeur maximale de β de sorte que les supports de noyaux soient disjoints, ce qui garantit la propriété d'indépendance asymptotique des divers diagnostics qui est un élément crucial de notre méthode de tests multiples.

Enfin, les résultats asymptotiques requièrent que les données soient différentiées jusqu'à la stationnarité. Comme les séries chronologiques économiques sont habituellement non stationnaires, il est souhaitable de différentier la tendance des données désaisonnalisées avant d'appliquer notre diagnostic. Cette différentiation pouvant atténuer la détection du premier pic saisonnier, les praticiens peuvent « recolorer » les données comme il est décrit à la section 3.4.

Tableau 5
Analyses des données pour 35 séries sur la fabrication (U.S. Census Bureau) comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

	Analyses des données – Séries sur la fabrication											
	Données originales							Données désaisonnalisées				
	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
Série	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
M_1	2345	1235	2345	1235	12	0,24	0,39	ø	ø	ø	ø	ø
M_2	12345	12345	12345	12345	1235	0,20	0,32	ø	ø	ø	ø	ø
M_3	12345	12345	12345	12345	1235	0,28	0,46	ø	ø	ø	ø	ø
M_4	ø	ø	12345	ø	12	0,28	0,44	ø	ø	ø	ø	ø
M_5	12345	12345	12345	12345	12345	0,27	0,47	ø	ø	12345	12345	ø
M_6	ø	ø	123	123	12	0,28	0,49	ø	ø	ø	ø	ø
M_7	ø	ø	123	123	24	0,50	0,79	ø	ø	ø	ø	ø
M_8	12345	12345	12345	12345	12345	0,18	0,37	ø	ø	ø	ø	ø
M_9	12345	12345	12345	12345	124	0,42	0,73	ø	ø	ø	ø	ø
M_{10}	ø	ø	ø	ø	1	0,38	0,72	ø	ø	ø	ø	ø
M_{11}	ø	ø	12345	1234	123	0,15	0,27	ø	ø	ø	ø	ø
M_{12}	1234	1234	12345	12345	1234	0,30	0,54	ø	ø	ø	ø	ø
M_{14}	ø	ø	1234	1234	1234	0,24	0,39	ø	ø	ø	ø	ø
M_{15}	12345	12345	12345	12345	12345	0,23	0,43	ø	ø	ø	ø	ø
M_{16}	1234	1234	1234	1234	1234	0,23	0,40	ø	ø	ø	ø	ø
M_{17}	ø	ø	1234	12345	12	0,64	0,66	ø	ø	ø	ø	ø
M_{18}	12345	12345	12345	12345	245	0,20	0,37	ø	ø	ø	ø	ø
M_{19}	ø	ø	ø	ø	4	0,86	1,00	ø	ø	ø	ø	ø
M_{20}	ø	ø	ø	12345	4	0,56	0,84	ø	ø	ø	ø	ø
M_{21}	12345	12345	12345	12345	1234	0,37	0,58	ø	ø	ø	ø	ø
M_{22}	12345	12345	12345	12345	1234	0,26	0,45	ø	ø	ø	ø	ø
M_{23}	12345	12345	12345	12345	1234	0,20	0,47	ø	ø	ø	ø	ø
M_{24}	12345	12345	12345	12345	2345	0,26	0,43	ø	ø	ø	ø	ø
M_{25}	12345	12345	12345	12345	12345	0,27	0,42	ø	ø	ø	ø	ø
M_{26}	12345	12345	12345	12345	1235	0,37	0,62	ø	ø	ø	ø	ø
M_{27}	1345	1234	1345	1234	2345	0,25	0,22	ø	ø	ø	ø	ø
M_{28}	ø	ø	ø	ø	24	0,57	0,44	ø	ø	ø	ø	ø
M_{29}	ø	12345	12345	12345	24	0,78	1,13	ø	ø	ø	ø	ø
M_{30}	123	1234	12345	12345	245	0,45	0,65	ø	ø	ø	ø	ø
M_{31}	ø	ø	123	123	4	0,64	0,46	ø	ø	1234	1234	ø
M_{32}	1235	12345	1235	12345	12345	0,21	0,37	ø	ø	ø	ø	ø
M_{33}	12345	12345	12345	1234	1234	0,24	0,38	ø	ø	ø	ø	ø
M_{34}	12345	12345	12345	12345	234	0,46	0,85	ø	ø	ø	ø	ø
M_{35}	12345	12345	12345	12345	2345	0,25	0,66	ø	ø	ø	ø	ø
M_{36}	12345	12345	12345	12345	123	1.32	1,56	ø	ø	ø	ø	ø
Total de résultats « corrects »	23/35	24/35	31/35	31/35	35/35	34/35	33/35	35/35	35/35	33/35	33/35	35/35
Nombre moyen	3,09	3,29	4,09	4,09	3,23			0	0	0,26	0,26	0

Tableau 6
Analyses des données pour 30 séries du U.S. Census Bureau (10 sur le logement, 10 sur les importations/exportations et 10 sur les ventes au détail) comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

Analyses des données – Séries sur la fabrication												
	Données originales							Données désaisonnalisées				
	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
Série	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
MW1Fam	12345	12345	12345	12345	12	0,13	0,25	ø	ø	ø	ø	ø
NWTot	12345	12345	12345	12345	12	0,18	0,31	ø	ø	ø	ø	ø
NE1Fam	12345	12345	12345	12345	12	0,16	0,33	ø	ø	ø	ø	ø
NETot	12345	12345	12345	12345	123	0,25	0,27	ø	ø	ø	ø	ø
S1Fam	12345	12345	12345	12345	125	0,22	0,47	ø	ø	ø	ø	ø
STot	124	124	1245	1245	125	0,29	0,57	ø	ø	ø	ø	ø
US1Fam	12345	12345	12345	12345	125	0,17	0,39	ø	ø	ø	ø	ø
USTot	12345	12345	12345	12345	125	0,20	0,42	ø	ø	ø	ø	ø
W1Fam	1234	1234	12345	12345	125	0,21	0,44	ø	ø	ø	ø	ø
WTot	1234	1234	12345	1234	12	0,27	0,56	ø	ø	ø	ø	ø
Total de résultats « corrects »	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Séries sur les importations/exportations												
	Données originales							Données désaisonnalisées				
	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
Série	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
M00120	12345	ø	12345	12345	125	0,23	0,48	ø	ø	ø	ø	ø
M00190	12345	12345	12345	12345	1235	0,38	0,59	ø	ø	ø	ø	ø
M3000	12345	12345	12345	12345	234	0,48	0,95	ø	ø	ø	ø	ø
M3010	1234	1234	1234	12345	2345	0,52	0,88	ø	ø	ø	ø	ø
M12060	12345	12345	12345	12345	123	0,53	0,77	ø	ø	ø	ø	ø
X3	12345	12345	12345	12345	2345	0,57	0,94	ø	ø	ø	ø	ø
X00300	134	134	134	134	2	0,56	0,97	ø	ø	ø	ø	ø
X3020	12345	12345	12345	12345	12345	0,39	0,70	ø	ø	ø	ø	ø
X3022	12345	12345	12345	12345	23	0,69	1,04	ø	ø	ø	ø	ø
X10140	1234	1234	1234	1234	15	0,29	0,47	ø	ø	ø	ø	ø
Total de résultats « corrects »	10/10	9/10	10/10	10/10	10/10	10/10	9/10	10/10	10/10	10/10	10/10	10/10
Séries sur les ventes de détail												
	Données originales							Données désaisonnalisées				
	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
Série	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
s0b441x0	12345	12345	12345	12345	135	0,22	0,41	ø	ø	ø	ø	ø
s0b 44000	12345	12345	12345	12345	2345	0,12	0,26	ø	ø	ø	ø	ø
s0b 44100	12345	12345	12345	12345	135	0,21	0,40	ø	ø	ø	ø	ø
s0b 44130	12345	12345	12345	12345	1235	0,21	0,42	ø	ø	ø	ø	ø
s0b 44200	12345	12345	12345	12345	12345	0,13	0,27	ø	ø	ø	ø	ø
s0b 44300	1234	12345	1234	12345	12345	0,12	0,18	ø	ø	ø	ø	ø
s0b 44312	1234	1234	1234	1234	12345	0,31	0,48	ø	ø	ø	ø	ø
s0b 44400	12345	12345	12345	12345	1235	0,16	0,32	ø	ø	ø	ø	ø
s0b 44410	12345	12345	12345	12345	1235	0,14	0,32	ø	ø	ø	ø	ø
s0b 44500	12345	12345	12345	12345	235	0,14	0,23	ø	ø	ø	ø	ø
Total de résultats « corrects »	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Total général des résultats « corrects »	30/30	29/30	30/30	30/30	30/30	30/30	29/30	30/30	30/30	30/30	30/30	30/30
Nombre moyen	4,67	4,5	4,77	4,8	3,23			0	0	0	0	0

Tableau 7
Analyses des données pour 22 séries allemandes de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

Analyses des données – OCDE, Allemagne												
	Données originales							Données désaisonnalisées				
Série	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
ALLEMAGNE	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
PRMNCG03	12345	12345	12345	12345	12345	0,17	0,33	Ø	Ø	Ø	Ø	Ø
PRMNCS01	1234	1234	1234	1234	235	0,25	0,79	Ø	Ø	Ø	Ø	Ø
PRMNIG01	12345	12345	12345	12345	12345	0,28	0,48	Ø	Ø	Ø	Ø	Ø
PRMNTO01	134	134	134	134	12345	0,26	0,46	Ø	Ø	Ø	Ø	Ø
PRMNVG01	1235	1235	1235	1235	2345	0,29	0,46	Ø	Ø	Ø	Ø	Ø
SLMNCD01	1245	1235	1245	1235	23	0,22	0,40	Ø	Ø	Ø	Ø	Ø
SLMNCN01	1345	2345	1345	2345	123	0,37	0,72	Ø	Ø	Ø	Ø	Ø
SLMNDM01	2345	2345	2345	2345	123	0,32	0,63	Ø	Ø	Ø	Ø	Ø
SLMNEX01	12345	12345	12345	12345	12	0,32	0,51	1:5	Ø	12345	12345	Ø
SLMNIG01	2345	2345	2345	2345	123	0,21	0,65	Ø	Ø	Ø	Ø	Ø
SLMNTO01	245	345	245	345	23	0,20	0,66	Ø	Ø	234	Ø	Ø
SLRTCR01	1345	1345	1345	1345	1234	0,19	0,51	Ø	Ø	Ø	Ø	Ø
SLRTTO01	12345	12345	12345	12345	12345	0,12	0,25	Ø	Ø	Ø	Ø	Ø
SLRTTO02	12345	12345	12345	12345	12345	0,13	0,29	Ø	Ø	Ø	Ø	Ø
SLWHTO01	2345	2345	2345	2345	123	0,20	0,62	Ø	Ø	134	123	Ø
SLWHTO02	2345	2345	2345	2345	123	0,20	0,62	Ø	Ø	134	123	Ø
UNLVRG01	23	23	23	23	124	0,23	0,48	Ø	Ø	1	Ø	5
UNLVSUMA	345	345	345	345	12	0,30	0,53	12	Ø	1245	12345	Ø
UNLVSUTT	234	234	234	234	12	0,24	0,53	Ø	Ø	45	45	25
UNRTRG01	235	1245	235	1245	123	0,19	0,59	Ø	Ø	2345	2345	Ø
XTEXVA01	1234	1234	1234	1234	23	0,28	0,77	Ø	Ø	Ø	Ø	Ø
XTIMVA01	234	1234	2345	12345	23	0,31	0,95	Ø	Ø	Ø	Ø	Ø
Total de résultats « corrects »	22/22	22/22	22/22	22/22	22/22	22/22	22/22	20/22	22/22	14/22	16/22	20/22
Nombre moyen	3,86	3,95	3,91	4,00	3,23			0	0	1,14	1,00	0,14

Tableau 8
Analyses des données pour 15 séries sur la zone euro de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

Analyses des données – OCDE, Euro												
	Données originales							Données désaisonnalisées				
Série	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
EURO	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
PRCNTO01	345	1345	345	1345	12345	0,14	0,44	Ø	Ø	Ø	Ø	Ø
PRINTO01	12345	12345	12345	12345	12345	0,10	0,23	Ø	Ø	Ø	Ø	Ø
PRMNCG03	1245	1245	1245	1245	12345	0,12	0,32	1	1	1234	1234	Ø
PRMNCS01	1234	12345	1234	12345	1234	0,22	0,47	Ø	Ø	Ø	Ø	Ø
PRMNIG01	12345	12345	12345	12345	2345	0,15	0,27	Ø	Ø	Ø	Ø	Ø
PRMNTO01	2345	2345	2345	2345	2345	0,14	0,23	Ø	Ø	1	Ø	Ø
PRMNVG01	1234	12345	1234	12345	2345	0,13	0,23	Ø	Ø	Ø	Ø	Ø
SLMNCN02	12345	12345	12345	12345	123	0,31	0,57	Ø	Ø	Ø	Ø	Ø

Tableau 8 (suite)
Analyses des données pour 15 séries sur la zone euro de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

Analyses des données – OCDE, Euro												
	Données originales							Données désaisonnalisées				
Série	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
EURO	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
SLMNIG02	12345	12345	12345	12345	23	0,21	0,45	Ø	Ø	Ø	Ø	Ø
SLMNTO02	1345	2345	1345	2345	23	0,20	0,41	24	Ø	24	34	Ø
SLMNVG02	12345	12345	12345	12345	2345	0,17	0,30	1	1	1245	1345	Ø
SLRTTO01	12345	12345	12345	12345	12345	0,05	0,12	Ø	Ø	Ø	Ø	Ø
SLRTTO02	12345	12345	12345	12345	12345	0,05	0,11	Ø	Ø	Ø	Ø	Ø
XTEXVA01	1345	2345	1345	2345	23	0,31	0,57	Ø	Ø	12	12	Ø
XTIMVA01	2345	2345	2345	2345	23	0,40	0,72	Ø	Ø	Ø	Ø	Ø
Total de résultats « corrects »	15/15	15/15	15/15	15/15	15/15	15/15	15/15	12/15	13/15	10/15	11/15	15/15
Nombre moyen	4,40	4,60	4,40	4,60	3,73			0,20	0,13	0,87	0,80	0

Tableau 9
Analyses des données pour 11 séries françaises de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

Analyses des données – OCDE, France												
	Données originales							Données désaisonnalisées				
Série	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
FRANCE	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
PRAFAG01	12345	12345	12345	12345	123	0,13	0,29	Ø	Ø	Ø	Ø	Ø
PRNCTO01	2345	2345	2345	2345	235	0,14	0,44	Ø	Ø	Ø	Ø	Ø
PRMNCG01	12345	12345	12345	12345	234	0,15	0,38	1	1	1	1	Ø
PRMNCS01	12345	12345	12345	12345	234	0,25	0,58	Ø	Ø	Ø	Ø	Ø
PRMNIG01	12345	12345	12345	12345	12345	0,11	0,26	Ø	Ø	Ø	Ø	Ø
PRMNTO01	12345	12345	12345	12345	12345	0,16	0,29	123	123	123	1234	Ø
PRMNVE01	12345	12345	12345	12345	12345	0,24	0,34	Ø	Ø	1245	1245	Ø
SLRTCR01	1345	2345	1345	2345	123	0,27	0,71	Ø	Ø	Ø	Ø	Ø
SLRTTO02	12345	12345	12345	12345	12345	0,16	0,36	Ø	Ø	Ø	Ø	Ø
XTEXVA01	1345	1345	1345	1345	23	0,14	0,44	Ø	Ø	Ø	Ø	Ø
XTIMVA01	1245	1245	1245	1245	23	0,18	0,54	Ø	Ø	Ø	Ø	Ø
Total de résultats « corrects »	11/11	11/11	11/11	11/11	11/11	11/11	11/11	9/11	9/11	8/11	8/11	11/11
Nombre moyen	4,64	4,64	4,64	4,64	3,55			0,36	0,36	0,73	0,81	0

Tableau 10
Analyses des données pour 17 séries britanniques de l'OCDE comparant notre diagnostic de pics multiples avec les diagnostics SV, M7 et M8. Notre diagnostic de pics multiples s'appuie sur la méthode H-FWER pour contrôler le taux d'erreur de type I global à $\alpha = 0{,}05$ et $\alpha = 0{,}10$ et examine la pente à $\delta = 0{,}10$ (voir la section 4.2)

Analyses des données – OCDE, Grande-Bretagne												
	Données originales							Données désaisonnalisées				
Série	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV	M7	M8	H-FWER 0,10/0,05		H-FWER 0,10/0,10		SV
GRANDE-BRETAGNE	Quartique	TH	Quartique	TH				Quartique	TH	Quartique	TH	
PPIAMP01	1234	1234	1234	1234	24	0,56	1,58	ø	ø	ø	ø	ø
PPIAMP02	1	1	123	123	2	0,53	0,92	ø	ø	ø	ø	ø
PPIPFU01	1345	12345	1345	12345	12	0,64	0,59	ø	ø	ø	ø	ø
PRINTO01	1345	1345	1345	1345	23	0,16	0,40	ø	ø	ø	ø	ø
PRMNCG02	2345	2345	2345	2345	123	0,23	0,56	ø	ø	ø	ø	ø
PRMNCG03	12345	12345	12345	12345	123	0,20	0,49	ø	ø	ø	ø	ø
PRMNCS01	123	12	123	1234	12	0,68	1,31	ø	ø	ø	ø	ø
PRMNIG01	2345	2345	2345	2345	123	0,15	0,47	1	ø	ø	ø	ø
PRMNTO01	1345	1345	1345	1345	23	0,17	0,45	ø	ø	ø	ø	ø
PRMNVE02	12345	12345	12345	12345	12345	0,25	0,76	ø	ø	ø	ø	ø
PRMNVE03	1234	1234	1234	1234	234	0,29	0,91	ø	ø	ø	ø	ø
PRMNVG01	124	134	124	134	234	0,18	0,58	ø	ø	ø	ø	ø
SLRTCR03	1345	1345	1345	1345	124	0,42	0,74	124	123	124	123	ø
SLRTTO02	12345	12345	12345	12345	12345	0,05	0,15	1234	ø	1234	12345	ø
UNLVRG01	12345	12345	12345	12345	1245	0,63	0,61	ø	ø	ø	ø	ø
XTEXVA01	134	134	1345	1345	23	0,34	1,02	ø	ø	ø	ø	ø
XTIMVA01	23	23	23	23	23	0,31	0,90	ø	ø	ø	ø	ø
Total de résultats « corrects »	17/17	17/17	17/17	17/17	17/17	17/17	14/17	14/17	16/17	13/17	15/17	17/17
Nombre moyen	3,76	3,76	3,94	4,06	2,76			0,47	0,18	0,53	0,47	0

Remerciements

Le présent article est diffusé en vue de tenir les parties intéressées au courant de la recherche courante et de favoriser la discussion des travaux en cours. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau. Les travaux de Holan ont été financés par une bourse de recherche de l'ASA/NSF/BLS.

Annexe

Nous établissons ici les formules asymptotiques pour les mesures statistiques ψ_A de la pente et de la convexité. Ces résultats peuvent alors être appliqués dans le paradigme de test pour obtenir des valeurs critiques asymptotiques. Certaines conditions faibles sont requises sur les données pour la théorie asymptotique ; nous nous inspirons de Taniguchi et Kakizawa (2000, section 3.1.1). La condition (B), due à Brillinger (1981), énonce que le processus est strictement stationnaire et la condition (B1) de Taniguchi et Kakizawa (2000, page 55) est vérifiée. La condition (HT), due à Hosoya et Taniguchi (1982), énonce que le processus a une représentation linéaire et les conditions (H1) à (H6) de Taniguchi et Kakizawa (2000, pages 55-56) sont vérifiées. L'hypothèse 1 (8) de Chiu (1988) est une condition de sommabilité sur divers cumulants d'ordre élevé, qui est satisfaite, par exemple, par un processus gaussien dont la densité spectrale est dans C^2. Aucune de ces conditions n'est stricte ; par exemple, un processus causal linéaire avec quatre moments satisfait (HT). Le résultat principal est une convergence conjointe de toute paire de mesures $\psi_A(I)$; par exemple, il peut s'agir d'une mesure de pente et de convexité avec le même noyau A. Nous présentons le théorème général qui couvre ces deux cas.

Théorème 1 *Supposons que les cumulants de quatrième ordre de* $\{X_t\}$ *disparaissent* ; *que la condition* (*B*) *ou* (*HT*) *est vérifiée* ; *et que l'hypothèse* 1 (8) *de Chiu* (1988) *est vérifiée. Soit les noyaux* A *et* B *satisfaisant les conditions* (*i*) *à* (*iv*) *de la section* 2.1. *Alors*

$$\left\{ \sqrt{n}\,\frac{(\theta_A(I) - \theta_A(f))}{\sqrt{\theta_{A^2}(I^2)/2}}, \quad \sqrt{n}\,\frac{(\theta_B(I) - \theta_B(f))}{\sqrt{\theta_{B^2}(I^2)/2}} \right\} \overset{\mathcal{L}}{\Rightarrow} N(0, V)$$

quand $n \to \infty$. *Ici*, 0 *dénote le vecteur nul* $(0, 0)'$, *et* V *est une matrice de dimension* 2×2 *dont les entrées sont*

$$V_{11} = V_{22} = 1 \quad V_{12} = V_{21} = \frac{\theta_{AB}(f^2)}{\sqrt{\theta_{A^2}(f^2)\theta_{B^2}(f^2)}}.$$

Preuve. Pour commencer, établissons que $\theta_{A^2}(I^2) \xrightarrow{p.s.} 2\theta_{A^2}(f^2)$. Puisque le noyau A est continu dans un intervalle (tel que $[\mu - \beta/2, \mu + \beta/2]$), ce résultat découle directement du corollaire 1 de Chiu (1988), en notant qu'il traite l'approximation de la fonction intégrale par les sommes de Riemann (Chiu (1988) définit aussi le périodogramme avec un facteur 2π). Évidemment, les mêmes résultats tiennent avec B à la place de A. Deuxièmement, considérons la convergence conjointe de $\theta_A(I)$ et $\theta_B(I)$. Nous utilisons la technique de Cramér-Wold et appliquons le lemme 3.1.1 de Taniguchi et Kakizawa (2000), généralisé comme il convient pour inclure les fonctions impaires (voir le théorème 3 de Chiu (1988)). Donc, pour tout x, y réel,

$$\sqrt{n}\left(x\,\frac{(\theta_A(I) - \theta_A(f))}{\sqrt{\theta_{A^2}(f^2)}} + y\,\frac{(\theta_B(I) - \theta_B(f))}{\sqrt{\theta_{B^2}(f^2)}} \right) \overset{\mathcal{L}}{\Rightarrow} N\left(0, \frac{1}{2\pi}\int_{-\pi}^{\pi} (C(\lambda)C(-\lambda) + C^2(\lambda)) f^2(\lambda)\, d\lambda \right)$$

en utilisant le théorème de Slutsky (Bickel et Doksum 1977), où le noyau C est défini par

$$C(\lambda) = \frac{x}{\sqrt{\theta_{A^2}(f^2)}}\, A(\lambda) + \frac{y}{\sqrt{\theta_{B^2}(f^2)}}\, B(\lambda).$$

De toute évidence, $C(\lambda)\, C(-\lambda) = 0$ et

$$C^2(\lambda) = \frac{x^2}{\theta_{A^2}(f^2)}\, A^2(\lambda) + 2\frac{xy}{\sqrt{\theta_{A^2}(f^2)}\sqrt{\theta_{B^2}(f^2)}}\, A(\lambda) B(\lambda) + \frac{y^2}{\theta_{B^2}(f^2)} B^2(\lambda).$$

En prenant x et y égaux à zéro et un dans diverses combinaisons, nous déduisons la matrice de variance énoncée V.

Ensuite, nous discutons du scénario de test de détection de pics multiples. Supposons donc que nous avons une série finie de noyaux A_i pour $i = 1, 2, ..., d$, qui satisfont chacun les hypothèses de la section 2. Alors, nous pouvons facilement généraliser le théorème 1 pour passer de deux à d noyaux comme il suit. La matrice de covariance asymptotique V aura la ij^{e} entrée

$$\frac{\theta_{A_i A_j}(f^2)}{\sqrt{\theta_{A_i^2}(f^2)\theta_{A_j^2}(f^2)}}.$$

Donc, si le support pour n'importe quelle paire de noyaux est disjoint, nous obtenons l'indépendance asymptotique et nous pouvons par conséquent appeler la procédure de tests multiples H-FWER.

Bibliographie

Aston, J., Findley, D., McElroy, T., Wills, K. et Martin, D. (2007). New ARIMA Models for Seasonal Time Series and Their Application to Seasonal Adjustment and Forecasting. *SRD Research Report No. RRS 2007*-14, *U.S. Census Bureau.*

Bell, W., et Hillmer, S. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-320.

Bickel, P., et Doksum, K. (1977). *Mathematical Statistics*: *Basic Ideas and Selected Topics*. Englewood Cliffs, New Jersey : Prentice Hall.

Benjamini, Y., et Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, 57, 289-300.

Brillinger, D. (1981). *Time Series Data Analysis and Theory*. San Francisco : Holden-Day.

Chiu, S. (1988). Weighted least squares estimators on the frequency domain for the parameters of a time series. *The Annals of Statistics*, 16, 1315-1326.

Evans, T., Holan, S. et McElroy, T. (2006). Evaluating measures for assessing spectral peaks. *2006 Proceedings American Statistical Association*, [CD-ROM] : Alexandria, VA}.

Findley, D. (2006). Communication personnelle.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. et Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177 (avec discussion).

Hosoya, Y., et Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics*, 10, 132-153.

Lothian, J., et Morry, M. (1978). A test for identifiable seasonality when using the X-11-ARIMA program. Document de travail, Division de la recherche et analyse en séries chronologiques, Statistique Canada.

Maravall, A., et Caporello, G. (2004). Program TSW: Revised Reference Manual. Document de travail 2004, *Research Department*, *Bank of Spain*. http://www.bde.es.

Nerlove, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica*, 32, 241-286.

Newton, H., et Pagano, M. (1983). A method for determining periods in time series. *Journal of the American Statistical Association*, 78, 152-157.

Parzen, E. (1983). Autoregressive spectral estimation. *Handbook of Statistics III*, (Éds., D. Brillinger et P. Krishnaiah), Amsterdam : North Holland, 221-247.

Priestley, M. (1981). *Spectral Analysis and Time Series*. London : Academic Press.

Soukup, R.J., et Findley, D.F. (1999). On the spectrum diagnostics used by X-12-ARIMA to indicate the presence of trading day effects after modeling or adjustment. Aussi au www.census.gov/pub/ts/papers/rr9903s.pdf. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 144-149.

Taniguchi, M., et Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. New York City, New York : Springer-Verlag.

U.S. Census Bureau (2002). X-12 ARIMA Reference Manual (Version 0.2.10), Washington, DC.

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 91-106
Statistique Canada, N° 12-001-X au catalogue

De la définition et de l'interprétation de la variabilité d'intervieweur pour un plan d'échantillonnage complexe

Siegfried Gabler et Partha Lahiri [1]

Résumé

La variabilité d'intervieweur est une composante importante de la variabilité des statistiques produites par sondage. Diverses stratégies liées au format et à la formulation des questions, ainsi qu'à la formation, à la charge de travail, à l'expérience et à l'affectation des intervieweurs sont employées pour essayer de réduire la variabilité d'intervieweur. La formule classique de mesure de la variabilité d'intervieweur, souvent appelée effet d'intervieweur, est donnée par $ieff := deff_int = 1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$, où ρ_{int} et $\overline{n}_{\text{int}}$ sont, respectivement, la corrélation intra-intervieweur et la moyenne simple des charges de travail d'intervieweur. Dans le présent article, nous donnons une justification assistée par modèle de cette formule bien connue pour les méthodes d'échantillonnage avec probabilités égales (EPE) quand il n'existe pas de grappes spatiales dans l'échantillon et que les charges de travail des intervieweurs sont égales. Toutefois, les grappes spatiales ainsi que la pondération inégale sont très fréquentes dans les enquêtes à grande échelle. Dans le contexte d'un plan d'échantillonnage complexe, nous obtenons une formule appropriée de la variabilité d'intervieweur qui tient compte des probabilités inégales de sélection et des grappes spatiales. Notre formule fournit une évaluation plus exacte des effets d'intervieweur et permet donc d'affecter un budget plus raisonnable au contrôle de la variabilité d'intervieweur. Nous proposons aussi une décomposition de l'effet global en effets dus à la pondération, aux grappes spatiales et aux intervieweurs. Cette décomposition aide à comprendre différents moyens de réduire la variance totale.

Mots clés : Effet d'intervieweur ; charges de travail des intervieweurs ; corrélation intra-intervieweur ; grappe spatiale ; pondération inégale.

1. Introduction

L'intervieweur est une source importante d'erreurs de mesure dans les enquêtes. Ce fait avait déjà été reconnu par Rice en 1929 et l'a été plus tard par de nombreux spécialistes de la recherche sur les enquêtes. Des facteurs tels que la qualité de la conception du questionnaire et les caractéristiques de l'intervieweur peuvent influer sur les effets d'intervieweur dans les statistiques produites par sondage.

L'intervieweur peut introduire dans les données d'enquête une homogénéité qui réduit généralement la taille effective de l'échantillon et, par conséquent, augmente la variance totale d'un estimateur. Traditionnellement, l'homogénéité intra-intervieweur a été mesurée au moyen du coefficient de corrélation intra-intervieweur ρ_{int}. L'importance de ce type de corrélation a été étudiée par de nombreux chercheurs, principalement dans le contexte des enquêtes téléphoniques sans grappes spatiales (Kish 1962 ; Gray 1956 ; Hanson et Marks 1958 ; Tucker 1983 ; Groves et Magilavy 1986 ; Heeb et Gmel 2001, et d'autres). Ces chercheurs ont soutenu que la nature des questions d'enquête peut avoir une incidence sur la valeur de ρ_{int}. Les questions sur les attitudes et les questions factuelles complexes sont considérées comme étant plus sensibles à la corrélation intra-intervieweur que les questions factuelles simples (Collins et Butcher 1982 ; Feather 1973 ; Fellegi 1964 ; Gray 1956 ; Hansen, Hurwitz et Bershad 1961). Selon Groves (1989), des valeurs supérieures à 0,1 sont rarement observées. Pour une discussion plus approfondie de la question, voir Schnell et Kreuter (2005).

Comme l'ont souligné plusieurs chercheurs, la formule standard de l'effet d'intervieweur $1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$ donne à penser que, même si la corrélation intra-intervieweur est faible, l'effet d'intervieweur pourrait être important, simplement à cause d'une charge moyenne de travail de l'intervieweur élevée. Par exemple, si $\rho_{\text{int}} = 0,01$ et $\overline{n}_{\text{int}} = 70$, nous avons $ieff = 1,69$ (Schnell et Kreuter 2005). Il convient de souligner qu'une charge de travail moyenne d'intervieweur élevée (par exemple de 60 à 70 cas) est très fréquente dans les enquêtes téléphoniques (Tucker 1983 ; Groves et Magilavy 1986). Dans le cas de l'Enquête sociale européenne, Philippens et Loosveldt (2004) ont produit des boîtes à moustaches des corrélations intra-intervieweur et des charges de travail d'intervieweur pour 18 pays participants.

L'effet, ou variance, d'intervieweur est généralement défini comme étant l'accroissement de la variance totale causée uniquement par les intervieweurs. Pour un plan EPE avec charges de travail égales des intervieweurs, la variance d'intervieweur pour la moyenne d'échantillon est donnée simplement par $1 + (n_{\text{int}} - 1)\rho_{\text{int}}$, où n_{int} est la charge de travail commune des intervieweurs. Dans le cas des enquêtes complexes avec charges de travail inégales des intervieweurs, les spécialistes de la recherche sur les enquêtes utilisent fréquemment une modification simple de cette formule où la charge de travail commune des

1. Siegfried Gabler, GESIS, C.P. 12 21 55, 68072 Mannheim, Allemagne. Courriel : siegfried.gabler@gesis.org; Partha Lahiri, Université du Maryland, College Park, États-Unis. Courriel : plahiri@survey.umd.edu.

intervieweurs est remplacée par la charge de travail moyenne des intervieweurs, c'est-à-dire la formule $1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$. À la section 2, nous soutenons que cette formule standard $1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$ ne peut pas être interprétée comme un accroissement de la variance totale causé par les intervieweurs, même dans le cas d'un plan EPE avec charges de travail inégales des intervieweurs. Aux sections 2 à 4, nous observons que la définition de la variance d'intervieweur dépend de la nature du plan d'échantillonnage complexe, ainsi que de la charge de travail affectée à l'intervieweur. Dans le présent article, nous donnons des définitions appropriées de la variance d'intervieweur sous divers scénarios de sondage. Une définition fiable de la variance d'intervieweur aide à déterminer les mesures qui doivent être prises afin de réduire la variabilité d'intervieweur. Les résultats présentés dans l'article sont avant tout applicables à la planification des enquêtes plutôt qu'à l'analyse des données d'enquête. Autrement dit, nous nous concentrons ici sur les définitions et sur l'interprétation de la variabilité d'intervieweur et non sur son estimation pour une enquête particulière.

À la section 2, nous considérons un plan EPE sans grappes spatiales et nous fournissons une interprétation assistée par modèle de l'effet d'intervieweur *ieff*. Nous montrons que, pour des charges de travail d'intervieweur égales, *ieff* est simplement égal au ratio de la variance de la moyenne d'échantillon sous un modèle corrélé tenant compte de l'homogénéité des observations recueillies par un même intervieweur à celle sous un modèle non corrélé qui ne tient pas compte de cette homogénéité. Donc, en multipliant la variance de la moyenne d'échantillon pour l'échantillonnage aléatoire simple par *ieff*, nous pouvons obtenir la variance totale de la moyenne d'échantillon qui tient compte à la fois de la variabilité d'échantillonnage et de la variabilité d'intervieweur. Il s'agit d'une interprétation très intuitive de *ieff* qui complète la justification assistée par modèle donnée antérieurement par Kish (1962). Dans cette section, nous montrons aussi que, pour un plan EPE, *ieff* est plus faible que la valeur donnée par la formule de l'effet d'intervieweur assistée par modèle si la charge de travail de l'intervieweur varie et que la corrélation intra-intervieweur est positive. Donc, le concepteur d'enquêtes qui utilise *ieff* fera moins d'effort qu'il n'est réellement nécessaire pour contrôler la variabilité d'intervieweur. Dans cette situation, une formule appropriée de l'effet d'intervieweur peut être obtenue à partir de *ieff* en substituant la charge de travail moyenne pondérée d'intervieweur à la moyenne simple habituelle.

À la section 3, nous examinons la possibilité d'une pondération inégale, mais sans grappes spatiales. Nous obtenons une interprétation assistée par modèle d'*ieff* si et uniquement si les personnes interrogées par un même intervieweur ont le même poids de sondage et que la charge de travail de l'intervieweur est inversement proportionnelle au carré du poids commun pour l'intervieweur. Il est intéressant de souligner que, contrairement au plan EPE, une répartition égale des charges de travail d'intervieweur ne garantit pas nécessairement une interprétation assistée par modèle d'*ieff*. En cas de charges de travail égales des intervieweurs, s'il existe au moins un intervieweur pour lequel les enquêtés n'ont pas tous le même poids de sondage, nous montrons que *ieff* est toujours plus élevé que la valeur donnée par la formule assistée par modèle. Nous dégageons aussi les facteurs à l'origine de la différence entre ces deux formules. Ces résultats ont un intérêt pratique en ce qui concerne la réduction des coûts d'enquête. Plus précisément, le concepteur d'enquêtes qui utilise *ieff* affectera vraisemblablement un budget plus important qu'il n'est vraiment nécessaire au contrôle de la variabilité d'intervieweur. Nous avons également mentionné certaines situations où *ieff* pourrait poser un problème de sous-estimation et, par conséquent, les concepteurs d'enquêtes qui utilisent cette formule pourraient accorder trop peu d'importance au contrôle des effets d'intervieweur. Notre formule fournit une évaluation plus exacte de la variabilité d'intervieweur et permet donc d'affecter un budget plus raisonnable au contrôle de la variabilité d'intervieweur. De surcroît, la modification des formules de planification aura une incidence sur la taille de l'échantillon.

Dans de nombreuses enquêtes par sondage à grande échelle, pour diverses raisons de nature organisationnelle et financière, telles que l'absence d'un registre de population général ou la réduction du coût global d'enquête, un plan d'échantillonnage en grappes à plusieurs degrés est considéré comme une alternative moins coûteuse que l'échantillonnage aléatoire simple. Sous un plan d'échantillonnage en grappes à plusieurs degrés, des personnes vivant dans des lieux spatialement rapprochés sont sélectionnées. Les personnes qui vivent dans une même grappe spatiale ont tendance à avoir les mêmes attitudes, à cause de leur contexte socioéconomique semblable, ce qui accroît l'homogénéité interne des données d'enquête. Cette homogénéité spatiale viole l'hypothèse iid (unités indépendantes et identiquement distribuées) fréquemment émise dans les méthodes d'inférence statistique standard, au même titre que la mise en grappes des intervieweurs. Ce fait a été reconnu par de nombreux spécialistes de la recherche sur les enquêtes et l'apport de corrections à diverses procédures statistiques, ainsi que les problèmes logiciels connexes ont été décrits dans la littérature (voir, entre autres, Rao et Scott 1984 ; Skinner, Holt et Smith 1989 ; Biemer et Trewin 1997 ; Chambers et Skinner 2003). À la section 4, nous présentons une nouvelle définition de la variabilité d'intervieweur en présence de pondération inégale et de grappes

spatiales. Dans le cas de grappes spatiales, nous soutenons que la formule *ieff* a généralement tendance à surestimer la variabilité d'intervieweur. Donc, pour les enquêtes complexes avec grappes spatiales, *ieff* peut donner inutilement une fausse alarme quant à l'importance de la variabilité d'intervieweur.

À la section 5, nous discutons de l'effet résultant des effets combinés de la pondération, des grappes spatiales et de l'intervieweur. La formule de l'effet global donne une détermination exacte de la taille d'échantillon à l'étape de la planification. Nous présentons une belle ventilation de cet effet global en effets dus à la pondération, aux grappes et à l'intervieweur. Cette décomposition peut faciliter la découverte de divers moyens de réduire la variance totale. Dans sa discussion de l'article de Verma, Scott et O'Muircheartaigh (1980), Hedges a mentionné qu'une telle formule de l'effet global était nécessaire. Nous généralisons une formule proposée antérieurement par Davis et Scott (1995) à un plan non EPE et à un modèle général de corrélation valide pour les données discrètes ainsi que continues. Les preuves de tous les résultats techniques sont présentées à l'annexe.

2. Plan EPE sans grappes spatiales

Soit y_{ik} l'observation obtenue auprès de la $k^{\text{ième}}$ personne interogée par le $i^{\text{ième}}$ intervieweur ($i = 1, ..., I$; $k = 1, ... n_i$). Définissons $n = \sum_{i=1}^{I} n_i$, la taille totale d'échantillon, $\overline{y} = 1/n \sum_{i=1}^{I}\sum_{k=1}^{n_i} y_{ik}$, la moyenne d'échantillon non pondérée et $\overline{n}_{\text{int}}(\mathbf{a}) = \sum_{i=1}^{I} a_i n_i$, une moyenne pondérée des charges de travail d'intervieweur, où a_i est un poids arbitraire appliqué à la charge de travail du $i^{\text{ième}}$ intervieweur et $\mathbf{a} = (a_1, ..., a_I)$.

Nous commencerons par donner une justification assistée par modèle de la formule classique d'effet d'intervieweur, c'est-à-dire $ieff = 1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$, où $\overline{n}_{\text{int}}$ est la moyenne non pondérée des charges de travail de l'intervieweur. Notons que $\overline{n}_{\text{int}} = \overline{n}_{\text{int}}(\mathbf{a}_0)$, avec $\mathbf{a}_0 = (a_{01}, ..., a_{0I})$, $a_{0i} = 1/I$ et $ieff = ieff(\mathbf{a}_0)$. En utilisant le résultat 1 donné à l'annexe, nous obtenons

$$ieff(\mathbf{a}_1) = \frac{\text{Var}_{M_2}(\overline{y})}{\text{Var}_{M_1}(\overline{y})} = 1 + [\overline{n}_{\text{int}}(\mathbf{a}_1) - 1]\rho_{\text{int}},$$

où $\mathbf{a}_1 = (a_{11}, ..., a_{1I})$, avec $a_{1i} = n_i / n$. Dans la formule susmentionnée, $\text{Var}_{M_1}(\overline{y})$ et $\text{Var}_{M_2}(\overline{y})$ sont les variances de $\overline{y}$ sous les deux modèles suivants, respectivement,

$$M_1 : \text{Cov}(y_{ik}, y_{i'k'}) = \begin{cases} \sigma^2 & \text{si } i = i',\ k = k', \\ 0 & \text{autrement,} \end{cases}$$

$$M_2 : \text{Cov}(y_{ik}, y_{i'k'}) = \begin{cases} \sigma^2 & \text{si } i = i',\ k = k', \\ \rho_{\text{int}}\sigma^2 & \text{si } i = i',\ k \neq k', \\ 0 & \text{autrement.} \end{cases}$$

Notons que, contrairement au modèle M_1, le modèle M_2 introduit l'homogénéité des observations recueillies par un même intervieweur.

Remarque 2.1 : Il découle du corollaire du résultat 1, donné à l'annexe, que pour $\rho_{\text{int}} > 0$, $ieff(\mathbf{a}_1) = ieff$ si et uniquement si $n_i = n/I$ pour tout i, c'est-à-dire si et uniquement si la charge de travail est la même pour chaque intervieweur. Pour le cas équilibré, Kish (1962) a donné une justification assistée par modèle de *ieff* en utilisant un modèle linéaire mixte qui est un cas particulier de M_2. Pour le cas non équilibré, il est intéressant de souligner la similarité entre la formule de la variabilité d'intervieweur $ieff(\mathbf{a}_1)$ et la formule des effets de plan (A3) donnée dans l'article de Holt discutant de Verma et coll. (1980).

Remarque 2.2 : Il découle du corollaire du résultat 1 que, si $\rho_{\text{int}} > 0$ et que les n_i ne sont pas égaux, alors $ieff(\mathbf{a}_1) > ieff$.

Dans l'exemple qui suit, nous démontrons la mesure dans laquelle $ieff(\mathbf{a}_1)$ et *ieff* pourraient différer pour divers profils de charges de travail des intervieweurs.

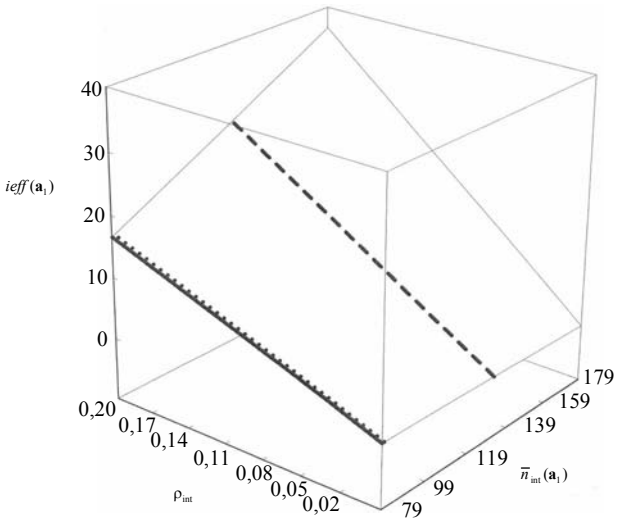
Exemple 1 : Dans le tableau 1, nous considérons trois affectations de charges de travail différentes pour dix intervieweurs, chacune avec $n = 790$. Le cas A) représente l'affectation des charges de travail la plus variable avec un écart-type = 68,3, le cas B) est presque équilibré avec un écart-type = 9,5 et le cas C) correspond à l'affectation de charges de travail égales aux intervieweurs.

Tableau 1
Trois affectations différentes des charges de travail aux intervieweurs (exemple 1)

	Profil des charges de travail des intervieweurs		
Intervieweur	A)	B)	C)
1	4	70	79
2	10	70	79
3	20	70	79
4	34	70	79
5	52	70	79
6	74	88	79
7	100	88	79
8	130	88	79
9	164	88	79
10	202	88	79
n	790	790	790
$\overline{n}_{\text{int}}(\mathbf{a}_1)$	132	80	79

Soit $ieff(\mathbf{a}_{1;A})$, $ieff(\mathbf{a}_{1;B})$ et $ieff(\mathbf{a}_{1;C}) = ieff$ qui désignent $ieff(\mathbf{a}_1)$, la formule de variance d'intervieweur assistée par modèle correspondant aux cas A, B et C, respectivement. Pour $\rho_{\text{int}} > 0$, la fonction $ieff(\mathbf{a}_1)$ est Schur-convexe, ce qui explique que $ieff(\mathbf{a}_{1;A}) \geq ieff(\mathbf{a}_{1;B}) \geq ieff(\mathbf{a}_{1;C}) = ieff$. La figure 1 donne les

valeurs de la variance d'intervieweur obtenues au moyen de la formule standard (c'est-à-dire, *ieff*) et de notre formule de variance d'intervieweur assistée par modèle pour toutes les combinaisons des deux facteurs influents, c'est-à-dire la moyenne pondérée des charges de travail d'intervieweur et la corrélation intra-intervieweur. En examinant la figure 1, il est intéressant de noter que *ieff* pourrait donner une sous-estimation d'environ 100 %.



Intervieweur A : trait interrompu
Intervieweur B : pointillé
Intervieweur C : trait plein

Figure 1 Graphiques de $ieff(\mathbf{a}_1)$ en fonction de ρ_{int} pour différents $\bar{n}_{int}(\mathbf{a}_1)$

3. Pondération inégale sans grappes spatiales

À la présente section, nous examinons la situation où les poids sont inégaux. Soit w_{ik} le poids de sondage associé à la $k^{\text{ième}}$ personne interrogée par le $i^{\text{ième}}$ intervieweur. Dans cette situation, la moyenne pondérée $\bar{y}_w = \sum_i \sum_k w_{ik} y_{ik} / \sum_i \sum_k w_{ik}$ est un estimateur fréquemment utilisé de la moyenne de population finie (voir Brewer 1963 ; Hájek 1971) et la formule de la variance d'intervieweur assistée par modèle est donnée par

$$ieff_w = \frac{\mathrm{Var}_{M_2}(\bar{y}_w)}{\mathrm{Var}_{M_1}(\bar{y}_w)} = 1 + \rho_{\text{int}} \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right).$$

Voir le résultat 1 en annexe.

Définissons $\bar{w}_i = 1/n_i \sum_{k=1}^{n_i} w_{ik}$, le poids de sondage moyen pour le $i^{\text{ième}}$ intervieweur et $\sigma_i^2 = 1/n_i \sum_k w_{ik}^2 - \bar{w}_i^2$, la variance des poids de sondage pour le $i^{\text{ième}}$ intervieweur.

On peut montrer que

$$ieff_w = 1 + \rho_{\text{int}}(\bar{n}_w - 1),$$

où

$$\bar{n}_w = \frac{\sum_i n_i^2 \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2 + \sum_i n_i \sigma_i^2}.$$

Notons qu'en général, $ieff_w$ ne peut pas s'écrire sous la forme $ieff_w = 1 + \rho_{\text{int}}(\bar{n}_{\text{int}}(\mathbf{a}) - 1)$ avec $\sum_i a_i = 1$.

Remarque 3.1 : Du résultat 2 en annexe, il découle que

$$ieff_w \le ieff(\mathbf{a}_2),$$

où

$$\mathbf{a}_2 = (a_{21}, \ldots, a_{2I}), \text{ avec } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2}.$$

Dans l'équation qui précède, pour $\rho_{\text{int}} > 0$, $ieff_w = ieff(\mathbf{a}_2)$ si et uniquement si tous les σ_i^2 sont nuls. Donc, $ieff(\mathbf{a}_2)$ peut être interprété comme une variance d'intervieweur prudente.

L'égalité est vérifiée si et uniquement si $w_{ik} = \bar{w}_i$ pour tout i et k, auquel cas

$$ieff_w = ieff(\mathbf{a}_2^*),$$

où

$$\mathbf{a}_2^* = (a_{21}^*, \ldots, a_{2I}^*), \text{ avec } a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}.$$

Donc, les formules $ieff_w$ et $ieff(\mathbf{a}_2^*)$ sont équivalentes si et uniquement si les poids de sondage sont tous les mêmes pour un intervieweur donné. Un exemple de plan de sondage de ce genre est un plan EPE pour lequel nous avons

$$a_{2i}^* = \frac{n_i}{n}$$

et

$$ieff_w = ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_1).$$

Maintenant, nous allons essayer de voir quels facteurs expliquent la différence entre $ieff_w$ et $ieff$. Pour cela, définissons

$\bar{w} = 1/n \sum_{i=1}^{I} \sum_{k=1}^{n_i} w_{ik} = \sum_{i=1}^{I} n_i/n \, \bar{w}_i$, le poids de sondage moyen pour l'ensemble des intervieweurs,

$SCE = \sum_{i=1}^{I} n_i (\overline{w}_i - \overline{w})^2$, la somme des carrés entre intervieweurs des poids de sondage,

$SCI = \sum_{i=1}^{I} \sum_{k=1}^{n_i} (w_{ik} - \overline{w}_i)^2 = \sum_{i=1}^{I} n_i \sigma_i^2$, la somme des carrés intra-intervieweur des poids de sondage,

$SCT = SCE + SCI$, la somme totale des carrés des poids de sondage,

$\tau_w = SCI/SCT$, un indicateur de la contribution relative de la variabilité intra-intervieweur des poids de sondage à la variabilité totale,

$CV_w = \sqrt{SCT/n} \; / \; \overline{w}$, le coefficient de variation des poids de sondage dans l'échantillon complet.

Nous pouvons montrer que (voir le résultat 4)

$$ieff_w - ieff = \frac{\overline{n}_{\text{int}}}{SCT + n\overline{w}^2}\left[\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right) n_i \overline{w}_i^2 - SCI\right]\rho_{\text{int}} \qquad (1)$$

$$= \frac{\overline{n}_{\text{int}}}{(1 + CV_w^{-2})\, SCT}\left[\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right) n_i \overline{w}_i^2 - SCI\right]\rho_{\text{int}} \qquad (2)$$

$$= \frac{\overline{n}_{\text{int}} \tau_w}{1 + CV_w^{-2}}\left(\frac{\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right) n_i \overline{w}_i^2}{SCI} - 1\right)\rho_{\text{int}}. \qquad (3)$$

Remarque 3.2 : Nous pouvons utiliser la formule (1) dans toute situation. Pour les plans EPE, nous avons

$$ieff_w - ieff = \rho_{\text{int}} \frac{\overline{n}_{\text{int}}}{n} \sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right) n_i.$$

Notons qu'une application de l'inégalité de Cauchy-Schwarz donne à penser que $ieff_w - ieff \geq 0$ avec égalité si et uniquement si $n_i = n/I$ pour tout i.

Remarque 3.3 : Nous pouvons utiliser (2) si $SCT \neq 0$, c'est-à-dire si le plan n'est pas EPE. Si $\rho_{\text{int}} > 0$, (2) implique que

$$ieff_w - ieff \leq 0 \text{ si et uniquement si } \sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right) n_i \overline{w}_i^2 \leq SCI.$$

Si une charge d'intervieweur élevée a tendance à être associée à de petits poids moyens de sondage et inversement et que $SCI \neq 0$, nous pouvons nous attendre à ce que $ieff$ soit une valeur prudente de la variance d'intervieweur réelle $ieff_w$. Dans l'exemple 2, c) et d), nous sommes en présence d'une telle situation.

Maintenant, nous avons $ieff_w = ieff$ si et uniquement si $w_{ik} = \overline{w}_i$ (ou, de manière équivalente, $SCI = 0$) et $n_i \overline{w}_i^2 / \sum_i n_i \overline{w}_i^2 = 1/I$ pour tout i et k, c'est-à-dire $ieff_w = ieff$, si et uniquement si $w_{ik} = \overline{w}_i$ et $\overline{w}_i \propto 1/\sqrt{n_i}$ pour tout i et k.

Donc, pour un plan non EPE, une charge de travail d'intervieweur égale ne fournit pas nécessairement une interprétation assistée par modèle de $ieff$. Par exemple, si les poids de sondage varient pour au moins un intervieweur, nous n'aurons pas d'interprétation assistée par modèle de $ieff$. Évidemment, pour un plan EPE, les deux formules sont équivalentes si et uniquement si les charges de travail des intervieweurs sont égales.

Remarque 3.4 : Si la charge de travail est la même pour tous les intervieweurs, nous avons

$$ieff_w - ieff = -\frac{\overline{n}_{\text{int}} \tau_w}{1 + CV_w^{-2}} \rho_{\text{int}}$$

(en supposant que $SCT \neq 0$). Donc, $ieff$ est une valeur prudente de l'effet réel d'intervieweur $ieff_w$. En outre, $|ieff_w - ieff|$ est une fonction croissante de la charge de travail d'intervieweur commune $\overline{n}_{\text{int}}$ et $\tau_w / (1 + CV_w^{-2})$ (pour CV_w^{-2} constant, cette dernière expression est une fonction croissante de τ_w). La même charge de travail d'intervieweur est donnée dans l'exemple 2 a).

Remarque 3.5 : Nous pouvons utiliser la formule (3) si $SCI > 0$, c'est-à-dire s'il existe au moins un intervieweur pour lequel les poids ne sont pas égaux.

Exemple 2.

Le tableau 2 donne huit combinaisons différentes de $(n_i, \overline{w}_i, \sigma_i^2)$. La première suppose que les valeurs de n_i sont égales, mais que les poids sont inégaux. La deuxième suppose que $\overline{w}_i^2 \propto \sigma_i^2$. Les six autres combinaisons montrent tous les classements possibles de $\overline{n}_{\text{int}}$, $\overline{n}_{\text{int}}(\mathbf{a}_1)$, $\overline{n}_w$, $\overline{n}_{\text{int}}(\mathbf{a}_2)$ et, par conséquent $ieff$, $ieff(\mathbf{a}_1)$, $ieff_w$, $ieff(\mathbf{a}_2)$ si l'on tient compte du fait que $ieff \leq ieff(\mathbf{a}_1)$ et $ieff_w \leq ieff(\mathbf{a}_2)$.

Dans l'exemple, $\sum_i n_i \overline{w}_i = n$. Nous expliquons maintenant les huit profils différents.

a) Puisque tous les n_i sont égaux, $ieff = ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$. En outre, $ieff_w$ est plus petit que tous les autres, parce que $\sigma_i^2 > 0$.

b) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. En outre, $\sigma_i^2 = c \cdot \overline{w}_i^2$ implique que $ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$.

c) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. En outre, puisque $\overline{w}_i^2 + \sigma_i^2$ et n_i augmentent tous deux, nous avons $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$.

d) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. Puisque $\overline{w}_i^2 + \sigma_i^2$ diminue et que n_i augmente, nous avons $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

e) Puisque les σ_i^2 sont relativement grands, $ieff_w < ieff$. En outre, $\overline{w}_i^2$ et σ_i^2 diminuent et n_i augmente, ce qui implique que $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

f) Le fait que $\overline{w}_i^2$ et n_i augmentent implique que $ieff_w > ieff$; puisque σ_i^2 et n_i augmentent tous deux, nous avons $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$.

g) Puisque $\overline{w}_i^2$ et n_i augmentent, nous avons $ieff_w > ieff$ et, puisque σ_i^2 augmente, nous avons $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$. En outre, $ieff_w < ieff(\mathbf{a}_1)$ puisque σ_i^2 est plus petit que dans f).

h) Puisque $\overline{w}_i^2$ et n_i augmentent, nous avons $ieff_w > ieff$ et puisque σ_i^2 diminue, nous avons $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

Tableau 2
Classement des formules d'effet d'intervieweur pour plusieurs combinaisons de paramètres (exemple 2) ; dans la dernière colonne $\rho_{int} = 0{,}01$

	n_i	$\overline{w}_i$	σ_i^2	$\overline{n}_{int}$	$\overline{n}_{int}(\mathbf{a}_1)$	$\overline{n}_w$	$\overline{n}_{int}(\mathbf{a}_2)$	Effets d'intervieweur	$ieff$ / $ieff_w$
a)	25	1,022	0,299	25	25	19,20	25	$ieff = ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2) > ieff_w$	1,003
	25	1,036	0,375						
	25	0,998	0,276						
	25	0,945	0,260						
b)	10	1	1	25	30	15	30	$ieff_w < ieff < ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$	1,007
	20	1	1						
	30	1	1						
	40	1	1						
c)	10	1	1	25	30	7,5	32,5	$ieff_w < ieff < ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$	1,023
	20	1	2						
	30	1	3						
	40	1	4						
d)	10	1	4	25	30	10	26,7	$ieff_w < ieff < ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$	1,015
	20	1	3						
	30	1	2						
	40	1	1						
e)	10	4	144	25	30	1,80	11,71	$ieff_w < ieff(\mathbf{a}_2) < ieff < ieff(\mathbf{a}_1)$	0,998
	20	2	9						
	30	0,333	0,555						
	40	0,250	0,125						
f)	10	0,333	0,025	25	30	31,82	35,26	$ieff < ieff(\mathbf{a}_1) < ieff_w < ieff(\mathbf{a}_2)$	1,015
	20	0,666	0,075						
	30	1	0,125						
	40	1,333	0,175						
g)	10	1	0,010	25	30	29,13	30,10	$ieff < ieff_w < ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$	0,999
	20	1	0,020						
	30	1	0,030						
	40	1	0,040						
h)	10	1	0,004	25	30	29,94	29,99	$ieff < ieff_w < ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$	0,998
	20	1	0,003						
	30	1	0,002						
	40	1	0,001						

4. Pondération inégale et grappes spatiales

À la présente section, nous obtenons une formule appropriée de la variance d'intervieweur en présence de grappes spatiales et de probabilités inégales de sélection. Considérons la situation où plus d'un intervieweur travaillent indépendamment dans la même UPE et où, dans chaque UPE, les enquêtés sont affectés aléatoirement aux intervieweurs. Nous supposons qu'aucun intervieweur ne travaille dans plus d'une UPE. Une telle situation a été considérée par Biemer et Stokes (1985). Maintenant, nous allons séparer l'effet d'intervieweur de l'effet d'UPE (c'est-à-dire l'existence de grappes spatiales) et de la pondération inégale. Soit y_{pik} et w_{pik} l'observation et le poids de sondage connexe pour la $k^{\text{ième}}$ personne dans la $p^{\text{ième}}$ UPE interrogée par le $i^{\text{ième}}$ intervieweur ($p = 1, ..., P$; $i = 1, ...I_p$; $k = 1, ..., n_{pi}$). Soit $n_p = \sum_{i=1}^{I_p} n_{pi}$ le nombre d'unités d'échantillonnage dans l'UPE p.

Dans ce cas, nous utilisons la moyenne pondérée qui suit pour estimer la moyenne de population finie :

$$\overline{y}_w = \frac{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik} y_{pik}}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}}.$$

Définissons

$$ieff_{s,\,w} = \frac{\mathrm{Var}_{M_4}(\overline{y}_w)}{\mathrm{Var}_{M_3}(\overline{y}_w)},$$

où les indices inférieurs s et w signifient la présence de grappes spatiales et de poids inégaux. Dans l'expression ci-dessus, $\mathrm{Var}_{M_3}(\overline{y}_w)$ et $\mathrm{Var}_{M_4}(\overline{y}_w)$ sont les variances de $\overline{y}_w$ sous les deux modèles suivants, respectivement

$$M_3 : \mathrm{Cov}(y_{pik},\, y_{p'i'k'}) = \begin{cases} \sigma^2 & \text{si } p=p',\ i=i',\ k=k' \\ \rho_C\sigma^2 & \text{si } p=p',\ k\neq k' \\ 0 & \text{autrement} \end{cases}$$

$$M_4 : \mathrm{Cov}(y_{pik}, y_{p'i'k'}) = \begin{cases} \sigma^2 & \text{si } p=p',\ i=i',\ k=k' \\ \rho_C\sigma^2 & \text{si } p=p',\ i\neq i' \\ \rho\sigma^2 & \text{si } p=p',\ i=i',\ k\neq k' \\ 0 & \text{si } p\neq p' \end{cases}$$

Dans ces modèles, ρ_C est la corrélation intra-UPE et ρ est la corrélation intra-classe combinée d'intervieweurs et d'UPE. Définissons $\rho_{\text{int}} = \rho - \rho_C$, la corrélation intra-intervieweur. Habituellement, $\rho_{\text{int}} > 0$.

Du résultat 5, il découle que

$$ieff_{s,\,w} = 1 + \rho_{\text{int}}\frac{\overline{n}_{\text{int}}(\mathbf{A}_w) - 1}{1 + \rho_C(\overline{n}_{\text{upe}}(\mathbf{b}_w) - 1)},$$

où

$$\mathbf{A}_w = ((a_{wpi}))_{\substack{i=1,\,...,\,I_p \\ p=1,\,...,\,P}} \quad \text{et} \quad a_{wpi} = \frac{n_{pi}\overline{w}_{pi}^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}$$

avec

$$\overline{w}_{pi} = \frac{1}{n_{pi}}\sum_{k}^{n_{pi}} w_{pik},$$

$$\overline{n}_{\text{int}}(\mathbf{A}_w) = \sum_{p=1}^{P}\sum_{i}^{I_p} a_{wpi} n_{pi} = \frac{\sum_{p=1}^{P}\sum_{i}^{I_p}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}$$

et

$$\mathbf{b}_w = (b_{wp})_{p=1,\ldots,P} \quad \text{et} \quad b_{wp} = \frac{n_p\overline{w}_p^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2},$$

avec

$$\overline{w}_p = \frac{1}{n_p}\sum_{i=1}^{I_p}\sum_{k}^{n_{pi}} w_{pik} = \frac{1}{n_p}\sum_{i=1}^{I_p} n_{pi}\overline{w}_{pi}$$

et

$$\overline{n}_{\text{upe}}(\mathbf{b}_w) = \sum_{p=1}^{P} b_{wp} n_p = \frac{\sum_{p=1}^{P}\left(\sum_{i}^{I_p}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}.$$

Notons que $\overline{n}_{\text{int}}(\mathbf{A}_w) \le \overline{n}_{\text{upe}}(\mathbf{b}_w)$ avec égalité si et uniquement si $I_P = 1$. Notons aussi que $\overline{n}_{\text{int}}(\mathbf{A}_w)$ ne varie pas en fonction de l'affectation des intervieweurs aux UPE, tandis que $\overline{n}_{\text{upe}}(\mathbf{b}_w)$ varie.

Remarque 4.1 : Si $\rho_C = 0$, nous obtenons

$$ieff_{s,w} = 1 + \rho_{\text{int}}(\overline{n}_{\text{int}}(\mathbf{A}_w) - 1).$$

Cette formule est semblable à $ieff_w$ donné à la section 2. Donc, tous les commentaires formulés à la remarque 2.1 s'appliquent ici. Notons que $\overline{n}_{\text{int}}(\mathbf{A}_w)$, tout comme $\overline{n}_w$, ne peut généralement pas s'écrire sous la forme $\overline{n}_{\text{int}}(\mathbf{A}_w) = \sum_{p=1}^{P}\sum_{i=1}^{I_p} a_{wpi} n_{pi}$ avec $\sum_{p=1}^{P}\sum_{i=1}^{I_p} a_{wpi} = 1$; le même commentaire s'applique à $\overline{n}_{\text{upe}}(\mathbf{b}_w)$.

Remarque 4.2 : Définissons

$$\overline{n}_{\text{int}}(\mathbf{A}) = \sum_{p=1}^{P}\sum_{i}^{I_p} a_{pi} n_{pi}, \text{ où } \mathbf{A} = ((a_{pi})), \text{ avec } a_{pi} = \frac{n_{pi}}{n},$$

et

$$\overline{n}_{\text{upe}}(\mathbf{b}) = \sum_{p=1}^{P} b_p n_p, \text{ où } \mathbf{b} = (b_1, ..., b_P) \text{ avec } b_p = \frac{n_p}{n}.$$

Si $\rho_C \neq 0$, mais que nous utilisons un plan EPE, nous laissons tomber l'indice inférieur w dans $ieff_{s,w}$. Notons que

$$ieff_s = 1 + \rho_{\text{int}} \frac{\overline{n}_{\text{int}}(\mathbf{A}) - 1}{1 + \rho_C [\overline{n}_{\text{upe}}(\mathbf{b}) - 1]}$$

$$= 1 + \frac{\rho_{\text{int}}}{\rho_C} \cdot \frac{\overline{n}_{\text{int}}(\mathbf{A}) - 1}{\overline{n}_{\text{upe}}(\mathbf{b}) - 1} \cdot \frac{\rho_C (\overline{n}_{\text{upe}}(\mathbf{b}) - 1)}{1 + \rho_C (\overline{n}_{\text{upe}}(\mathbf{b}) - 1)}$$

de sorte que

$$ieff_s < 1 + \frac{\rho_{\text{int}}}{\rho_C} \cdot \frac{\overline{n}_{\text{int}}(\mathbf{A}) - 1}{\overline{n}_{\text{upe}}(\mathbf{b}) - 1} < 1 + \frac{\rho_{\text{int}}}{\rho_C} \cdot \frac{\overline{n}_{\text{int}}(\mathbf{A})}{\overline{n}_{\text{upe}}(\mathbf{b})}.$$

Il est facile de voir que le deuxième membre de l'inégalité augmente avec les ratios $\rho_{\text{int}} / \rho_C$ et

$$\frac{\overline{n}_{\text{int}}(\mathbf{A}) - 1}{\overline{n}_{\text{upe}}(\mathbf{b}) - 1}.$$

Nous avons

$$ieff_s - ieff = \rho_{\text{int}} \frac{\frac{\overline{n}_{\text{int}}(\mathbf{A}) - 1}{\overline{n}_{\text{int}} - 1} - [1 + \rho_C (\overline{n}_{\text{upe}}(\mathbf{b}) - 1)]}{1 + \rho_C (\overline{n}_{\text{upe}}(\mathbf{b}) - 1)} (\overline{n}_{\text{int}} - 1).$$

Donc, pour $\rho_{\text{int}} > 0$,

$ieff_s < ieff$ si et uniquement si

$$Deff_s := 1 + \rho_C (\overline{n}_{\text{upe}}(\mathbf{b}) - 1) > \frac{\overline{n}_{\text{int}}(\mathbf{A}) - 1}{\overline{n}_{\text{int}} - 1},$$

c'est-à-dire si et uniquement si l'effet de plan dû aux grappes spatiales est plus grand que le ratio de la moyenne pondérée des charges de travail d'intervieweur -1 à la moyenne des charges de travail d'intervieweur -1. Si la charge de travail est la même pour tous les intervieweurs, le deuxième membre de l'inégalité est égal à 1 et, par conséquent, l'inégalité est toujours valide. Il est intéressant de souligner que $ieff \approx 4 \cdot ieff_s$ si $\rho_{\text{int}} = 0{,}1$, $\rho_C = 0{,}05$, $\overline{n}_{\text{upe}}(b) = 140$, et $\overline{n}_{\text{int}} = 70$.

Remarque 4.3 : Dans le cas général, nous avons

$$ieff_{s,w} - ieff = \rho_{\text{int}} \left(\frac{\overline{n}_{\text{int}}(\mathbf{A}_w) - 1}{1 + \rho_C (\overline{n}_{\text{upe}}(\mathbf{b}_w) - 1)} - (\overline{n}_{\text{int}} - 1) \right).$$

Donc, pour $\rho_{\text{int}} > 0$,

$ieff_{s,w} < ieff$ si et uniquement si

$$Deff_{s,w} := 1 + \rho_C (\overline{n}_{\text{upe}}(\mathbf{b}_w) - 1) > \frac{\overline{n}_{\text{int}}(\mathbf{A}_w) - 1}{\overline{n}_{\text{int}} - 1},$$

c'est-à-dire si et uniquement si

$$\rho_C > \frac{\overline{n}_{\text{int}}(\mathbf{A}_w) - \overline{n}_{\text{int}}}{(\overline{n}_{\text{int}} - 1)(\overline{n}_{\text{upe}}(\mathbf{b}_w) - 1)} =: \rho_C^*, \text{ disons.}$$

Dans l'exemple 2 (voir le tableau 3), $ieff$ est une valeur prudente de $ieff_{s,w}$ pour a) à e) si $\rho_C > 0$. Il en est de même pour f) à h) si $\rho_C > 0{,}004$.

Si un ménage et une personne dans ce ménage sont sélectionnés au hasard, les poids sont souvent indépendants de l'UPE et de l'intervieweur, et dépendent uniquement de la taille du ménage. Le cas échéant, les tailles de ménage forment les classes de pondération. Pour ces dernières, nous définissons

m_{pij} : nombre d'unités d'échantillonnage dans l'UPE p affectées à l'intervieweur i appartenant à la classe de pondération j,

$m_{pj} = \sum_{i=1}^{I_p} m_{pij}$: nombre d'unités d'échantillonnage dans l'UPE p appartenant à la classe de pondération j,

$m_j = \sum_{p=1}^{P}\sum_{i=1}^{I_p} m_{pij}$: nombre d'unités d'échantillonnage appartenant à la classe de pondération j.

Donc,

$n_{pi} = \sum_{j=1}^{J} m_{pij}$: nombre d'unités d'échantillonnage dans l'UPE p affectées à l'intervieweur i,

$n_p = \sum_{i=1}^{I_p}\sum_{j=1}^{J} m_{pij}$: nombre d'unités d'échantillonnage dans l'UPE p,

$n = \sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{j=1}^{J} m_{pij}$: taille de l'échantillon.

En outre,

$$\overline{n}_{\text{int}}(\mathbf{A}_w) = \frac{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\left(\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2} = \frac{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\left(\sum_{j=1}^{J} w_j m_{pij}\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}$$

et

$$\overline{n}_{\text{upe}}(\mathbf{b}_w) = \frac{\sum_{p=1}^{P}\left(\sum_{i}^{I_p}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2} = \frac{\sum_{p=1}^{P}\left(\sum_{j} w_j m_{pj}\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}$$

sont les ratios des formes quadratiques dans $\mathbf{w} = (w_1, ..., w_J)$.

Tableau 3
Charges de travail d'intervieweur moyennes pour plusieurs combinaisons de paramètres (exemple 2) ; $ieff$ / $ieff_{s,w}$ pour $\rho_{\text{int}} = 0{,}01$ et $\rho_C = 0{,}02$

					$IA = (1,3)$				$IA = (2,2)$				$IA = (3,1)$			
	n_i	$\bar{w}_i$	σ_i^2	$\bar{n}_{\text{int}}$	$\bar{n}_{\text{int}}(A_w)$	$\bar{n}_{\text{upe}}(\mathbf{b}_w)$	ρ_C^*	$\frac{ieff}{ieff_{s,w}}$	$\bar{n}_{\text{int}}(A_w)$	$\bar{n}_{\text{upe}}(\mathbf{b}_w)$	ρ_C^*	$\frac{ieff}{ieff_{s,w}}$	$\bar{n}_{\text{int}}(A_w)$	$\bar{n}_{\text{upe}}(\mathbf{b}_w)$	ρ_C^*	$\frac{ieff}{ieff_{s,w}}$
a)	25	1,022	0,299	25	19,202	47,528	-0,005	1,133	19,202	38,389	-0,006	1,123	19,202	49,039	-0,005	1,135
	25	1,036	0,375													
	25	0,998	0,276													
	25	0,945	0,260													
b)	10	1	1	25	15	41	-0,010	1,151	15	29	-0,015	1,138	15	26	-0,017	1,134
	20	1	1													
	30	1	1													
	40	1	1													
c)	10	1	1	25	7,5	20,5	-0,037	1,185	7,5	14,5	-0,054	1,180	7,5	13	-0,061	1,178
	20	1	2													
	30	1	3													
	40	1	4													
d)	10	1	4	25	10	27,333	-0,024	1,171	10	19,333	-0,034	1,163	10	17,333	-0,038	1,161
	20	1	3													
	30	1	2													
	40	1	1													
e)	10	4	144	25	1,801	2,755	-0,551	1,230	1,801	3,603	-0,371	1,231	1,801	4,344	-0,289	1,231
	20	2	9													
	30	0,333	0,555													
	40	0,250	0,125													
f)	10	0,333	0,025	25	31,820	75,685	0,004	1,104	31,820	58,427	0,005	1,084	31,820	40,629	0,007	1,058
	20	0,666	0,075													
	30	1	0,125													
	40	1,333	0,175													
g)	10	1	0,010	25	29,126	79,612	0,002	1,118	29,126	56,311	0,003	1,094	29,126	50,485	0,003	1,086
	20	1	0,020													
	30	1	0,030													
	40	1	0,040													
h)	10	1	0,004	25	29,940	81,836	0,003	1,117	29,940	57,884	0,004	1,092	29,940	51,896	0,004	1,084
	20	1	0,003													
	30	1	0,002													
	40	1	0,001													

5. Effet global

L'effet global tient compte de la pondération inégale, des grappes spatiales et des effets d'intervieweur, et peut être considéré comme une généralisation de l'effet de plan classique. La multiplication de la variance sous EAS pour la moyenne d'échantillon non pondérée par l'effet global fournira l'estimateur de variance totale

$$eff = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_1^*}(\bar{y})} = eff_w \times eff_s \times eff_{\text{int}},$$

où

$$eff_w = \frac{\text{Var}_{M_1^*}(\bar{y}_w)}{\text{Var}_{M_1^*}(\bar{y})},$$

$$eff_s = \frac{\text{Var}_{M_3}(\bar{y}_w)}{\text{Var}_{M_1^*}(\bar{y}_w)},$$

$$eff_{\text{int}} = ieff_{s,w} = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_3}(\bar{y}_w)}.$$

Ci-dessus, $\text{Var}_{M_1^*}$ est calculé par rapport au modèle suivant :

$$M_1^*: \ \text{Cov}(y_{pik}, y_{p'i'k'}) = \begin{cases} \sigma^2 \text{ si } p = p', i = i', k = k', \\ 0 \quad \text{autrement.} \end{cases}$$

Nous pouvons montrer que

$$eff = \frac{n\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}$$

$$\times\left[1+\rho_C\left(\frac{\sum_{p=1}^{P}\left(\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}-1\right)+\rho_{\text{int}}\left(\frac{\sum_{p=1}^{P}\sum_{i}^{I_p}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}-1\right)\right].$$

Les contributions relatives de la pondération, des grappes spatiales et des effets d'intervieweur à l'effet global sont données par

$$eff_w \text{ rel} = \frac{n\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2} \Big/ eff,$$

$$eff_s \text{ rel} = \frac{1+\rho_C(\overline{n}_{\text{upe}}(\mathbf{b}_w)-1)}{eff},$$

$$eff_I \text{ rel} = \frac{1+\rho_{\text{int}}\dfrac{\overline{n}_{\text{upe}}(A_w)-1}{1+\rho_C(\overline{n}_{\text{upe}}(\mathbf{b}_w)-1)}}{eff}.$$

À la figure 2, nous présentons des graphiques tridimensionnels des contributions relatives de la pondération, des grappes spatiales et des effets d'intervieweur à l'effet global pour diverses combinaisons de corrélations intra-grappe et intra-intervieweur pour différents profils de pondération donnés par les cas a), f) et h) du tableau 3 avec $IA = (1, 3)$, où $IA = (a, b)$ indique que les a premiers des quatre intervieweurs sont dans l'UPE 1 et que les b derniers intervieweurs sont dans l'UPE 2.

Remarque 5.1 : Du résultat 6, il découle que

$$eff \geq 1 + \rho_C\left(\frac{n}{P}-1\right) + \rho_{\text{int}}\left(\frac{n}{I}-1\right).$$

Le deuxième membre est l'effet global si le même nombre d'intervieweurs ayant une même charge de travail sont affectés à chaque UPE. Il convient de souligner la similarité entre le deuxième membre de l'inégalité qui précède et la formule de l'effet de plan (3.1) donnée dans Hansen, Hurwitz et Madow (1953, vol. I, page 370). Pour pouvoir affirmer la similarité, nous devons traiter les unités secondaires d'échantillonnage comme des unités appartenant à un intervieweur. Dans cette connexion, nous notons aussi la formule (3.7) donnée dans Hansen et coll. (1953, vol. II, page 292) pour le cas $I = P$.

Remarque 5.2 : Si les classes de pondération sont les mêmes pour les diverses interactions UPE × intervieweur, nous avons

$$eff = \frac{n\sum_{j=1}^{J} w_j^2 m_j}{\left(\sum_{j=1}^{J} w_j m_j\right)^2} \times \left[1+\rho_C\left(\frac{\sum_{p=1}^{P}\left(\sum_{j=1}^{J} w_j m_{pj}\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}-1\right)+\rho_{\text{int}}\left(\frac{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\left(\sum_{j=1}^{J} w_j m_{pij}\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}-1\right)\right].$$

Remarque 5.3 : Considérons le cas particulier

$$m_{pij} = \frac{n_{pi} m_j}{n}$$

dans lequel nous permettons aux poids de varier dans les classes UPE × intervieweur, mais nous les contraignons d'avoir la même distribution de fréquences relatives dans chaque classe, c'est-à-dire que les moyennes et les variances des poids à l'intérieur des classes ne dépendent pas de la classe (Lynn et Gabler 2004). Il est facile de voir que, dans ces conditions,

$$eff = \frac{n\sum_{j=1}^{J} w_j^2 m_j}{\left(\sum_{j=1}^{J} w_j m_j\right)^2} \times \left[1+\rho_C\left(\frac{\left(\sum_{j=1}^{J} w_j m_j\right)^2 \sum_{p=1}^{P} n_p^2}{\sum_{j=1}^{J} w_j^2 m_j \, n^2}-1\right) + \rho_{\text{int}}\left(\frac{\left(\sum_{j=1}^{J} w_j m_j\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}\sum_{p=1}^{P}\sum_{i=1}^{I_p}\frac{n_{pi}^2}{n^2}-1\right)\right].$$

En utilisant le même argument que celui donné dans la preuve du résultat 6, nous obtenons

$$eff \geq 1 + \rho_C\left(\frac{\sum_{p=1}^{P} n_p^2}{n^2}-1\right) + \rho_{\text{int}}\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\frac{n_{pi}^2}{n}-1\right)$$
$$= 1 + \rho_C(\overline{n}_{\text{upe}}(\mathbf{b})-1) + \rho_{\text{int}}(\overline{n}_{\text{int}}(\mathbf{A})-1).$$

Cela signifie que l'effet global est plus grand que celui observé pour un plan EPE (voir la remarque 5.4).

Remarque 5.4 : Pour un plan EPE, nous avons

$$eff = 1 + \rho_C(\overline{n}_{\text{upe}}(\mathbf{b})-1) + \rho_{\text{int}}(\overline{n}_{\text{int}}(\mathbf{A})-1),$$

où

$$\overline{n}_{\text{upe}}(\mathbf{b}) = \frac{\sum_{p=1}^{P} n_p^2}{n} \text{ et } \overline{n}_{\text{int}}(\mathbf{A}) = \frac{\sum_{p=1}^{P}\sum_{i}^{I_p} n_{pi}^2}{n}.$$

Notons que Davis et Scott (1995) ont obtenu cette formule pour le cas particulier du modèle linéaire mixte suivant :

$$y_{pik} = \mu + \alpha_i + \beta_p + \varepsilon_{pik},$$

où μ est l'effet global, α_i, β_p sont les effets aléatoires dus à l'intervieweur i et à l'UPE p, respectivement et ε_{pik} est l'erreur pure. Ils ont supposé que les effets aléatoires étaient indépendants avec $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_p \sim N(0, \sigma_\beta^2)$ et $\varepsilon_{pik} \sim N(0, \sigma_\varepsilon^2)$.

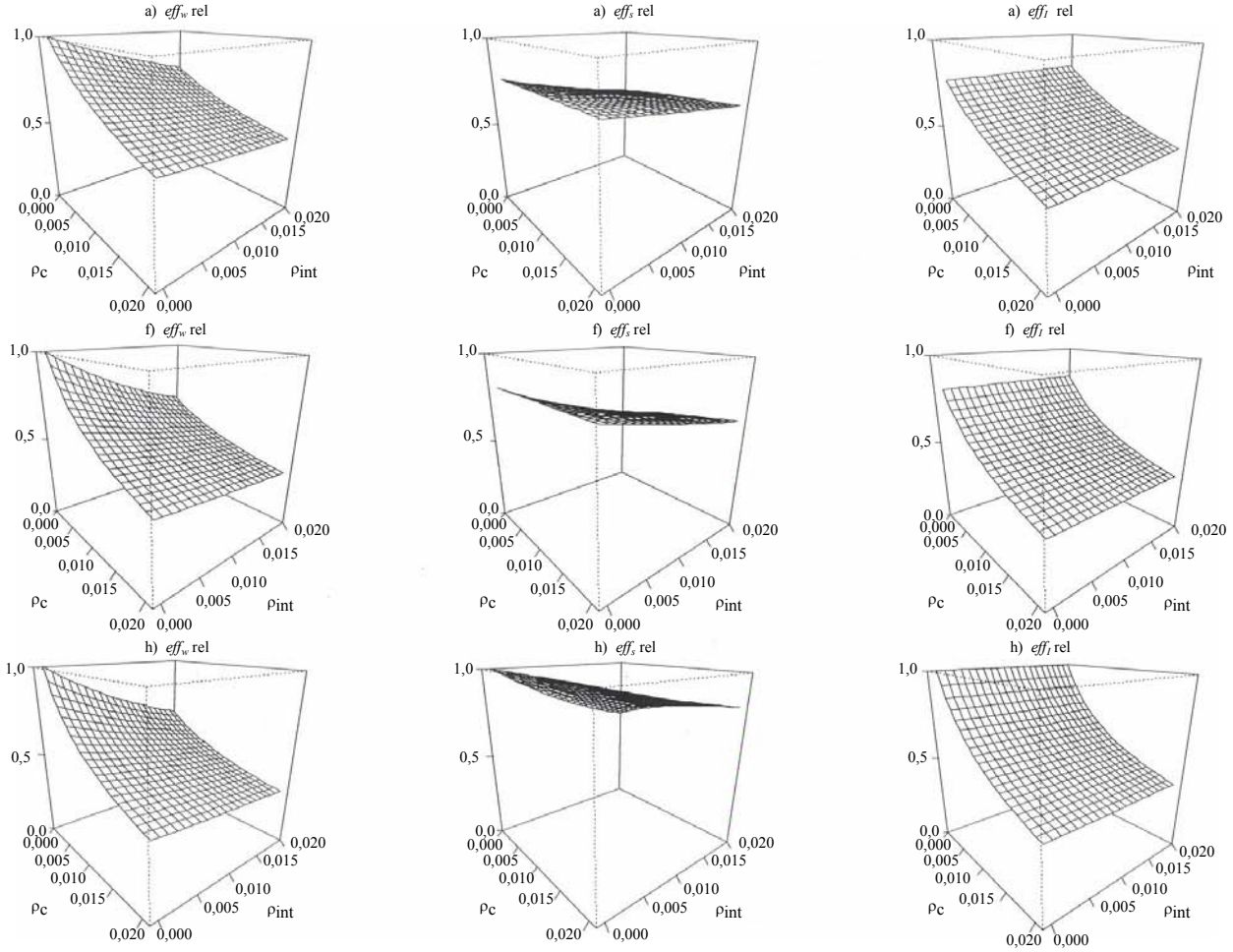


Figure 2 Contributions relatives des effets de pondération, de plan et d'intervieweur à l'effet global pour les cas a), f) et h) de l'exemple 2 pour le cas *IA* = (1, 3)

Pour le modèle linéaire mixte susmentionné, il est facile de vérifier que

$$\rho_{\text{int}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2} \text{ et } \rho_c = \frac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2} .$$

Cependant, il est intéressant de souligner que la définition *eff* ne nécessite pas que ρ_{int} et ρ_c soient strictement positifs et elle s'étend au-delà du modèle linéaire mixte. Par exemple, elle s'applique à l'exemple suivant :

Exemple 3 : Modèle simple pour des données binaires.

En supposant que $0 < \min(\alpha, \beta) < \theta < 1$, nous définissons le modèle suivant :

Pour chacune des n_{pi} personnes interrogées par l'intervieweur i dans l'UPE p,

$P(Y_{pik} = x_1, Y_{pik'} = x_2)$

x_1 \ x_2	1	0	Total
1	α	$\theta - \alpha$	θ
0	$\theta - \alpha$	$1 - 2\theta + \alpha$	$1 - \theta$
Total	θ	$1 - \theta$	1

Pour chacune des n_{pi} personnes interrogées par l'intervieweur i dans l'UPE p et chacune des $n_{pi'}$ personnes interrogées par l'intervieweur i' dans l'UPE p,

$P(Y_{pik} = x_1, Y_{pi'k'} = x_2)$

x_1 \ x_2	1	0	Total
1	β	$\theta - \alpha$	θ
0	$\theta - \alpha$	$1 - 2\theta + \beta$	$1 - \theta$
Total	θ	$1 - \theta$	1

Pour chacune des n_p personnes interrogées de l'UPE p et chacune des $n_{p'}$ personnes interrogées de l'UPE p',

$P(Y_{pik} = x_1, Y_{p'i'k'} = x_2)$

x_1 \ x_2	1	0	Total
1	θ^2	$\theta(1-\theta)$	θ
0	$\theta(1-\theta)$	$(1-\theta)^2$	$1-\theta$
Total	θ	$1-\theta$	1

Par conséquent, nous avons

$$E(Y_{pik}) = \theta \text{ pour tout } p, i, k,$$

$$\mathrm{Var}(Y_{pik}) = \theta(1-\theta) \text{ pour tout } p, i, k,$$

$$\rho = \frac{\mathrm{Cov}(Y_{pik}, Y_{pik'})}{\sqrt{\mathrm{Var}(Y_{pik})\mathrm{Var}(Y_{pik'})}} = \frac{\alpha - \theta^2}{\theta(1-\theta)} \text{ pour tout } p, i \text{ et } k \neq k',$$

$$\rho_C = \frac{\mathrm{Cov}(Y_{pik}, Y_{pi'k'})}{\sqrt{\mathrm{Var}(Y_{pik})\mathrm{Var}(Y_{pi'k'})}} = \frac{\beta - \theta^2}{\theta(1-\theta)} \text{ pour tout } p \text{ et } i \neq i',$$

ce qui est un cas particulier du modèle M_4 avec $\sigma^2 = \mathrm{Var}(Y_{pik}) = \theta(1-\theta)$. Notons que ρ_C ainsi que ρ peuvent être négatifs et que $\rho_{\text{int}} = \rho - \rho_C$ est positif si et uniquement si $\alpha > \beta$.

Remarque 5.5 : Pour un plan EPE avec taille d'UPE commune $b = n/P$, nous avons

$$eff = 1 + \rho_C(b-1) + \rho_{\text{int}}(\bar{n}_{\text{int}}(A) - 1).$$

Remarque 5.6 : Dans sa discussion de Verma et coll. (1980), Holt a considéré le cas où il n'existe aucune variabilité d'intervieweur et où l'UPE est la classe de pondération, c'est-à-dire le cas où $\rho_{\text{int}} = 0$ et $w_{pik} = w_p$ pour tout p, i, k. Dans ces conditions, eff se réduit à

$$eff = \frac{n\sum_{p=1}^{P} n_p w_p^2}{\left(\sum_{p=1}^{P} n_p w_p\right)^2} \times \left[1 + \rho_C\left(\frac{\sum_{p=1}^{P} n_p^2 w_p^2}{\sum_{p=1}^{P} n_p w_p^2} - 1\right)\right].$$

Notons que la formule ci-dessus peut être obtenue à partir de l'équation (A4) donnée par Holt dans sa discussion de Verma et coll. (1980), après correction d'une coquille manifeste (c'est-à-dire en supprimant n dans le dénominateur), en adoptant les poids de sondage qu'il a choisis et en effectuant certaines opérations algébriques. Les formules de l'effet de plan en l'absence d'effets d'intervieweur ont été examinées par de nombreux auteurs. Voir Kish (1965), Verma et coll. (1980), Skinner (1986), Valliant (1987), Skinner et coll. (1989), Gabler, Häder et Lahiri (1999), Lynn et Gabler (2004), Kalton, Brick et Lê (2005), et d'autres.

6. Conclusion

Nous avons constaté que la formule standard de l'effet d'intervieweur pourrait poser un problème de surestimation ou de sous-estimation selon la situation. Par exemple, elle pourrait sous-estimer gravement les effets d'intervieweur dans le cas d'un plan d'échantillonnage avec probabilité égale (EPE) quand les charges de travail des intervieweurs sont différentes. Curieusement, la corrélation spatiale peut transformer cette sous-estimation en une surestimation. Dans le premier cas, le concepteur d'enquêtes qui utilise la formule standard d'effet d'intervieweur pourrait ne pas accorder suffisamment d'attention au contrôle de cet effet. Dans le second cas, une valeur élevée de l'effet d'intervieweur pourrait susciter inutilement des préoccupations au sujet de la qualité des données associées à l'intervieweur. Cela pourrait déclencher l'affectation d'une part plus grande qu'il n'est nécessaire du budget à la réduction de l'effet d'intervieweur, lequel pourrait déjà être nettement plus faible que la valeur obtenue par une application de la formule standard. L'objectif de l'article est de définir et d'interpréter les effets d'intervieweur qui sont appropriés dans diverses situations d'enquêtes complexes.

Nous avons considéré le cas où un intervieweur est affecté à une seule UPE. Le cas où un intervieweur travaille dans différentes UPE est également important et sera examiné dans un futur article. Les poids utilisés dans les formules proposées tiennent compte uniquement des poids d'échantillonnage, car ils sont planifiés à l'étape de la conception de l'enquête, mais ne reflètent pas nécessairement les poids réels associés à chaque cas, une fois que les données sont recueillies. Autrement dit, nos formules d'effet d'intervieweur ne tiennent pas compte des effets dus à la non-réponse et aux corrections par poststratification. Les formules présentées dans l'article sont surtout utiles à l'étape de la planification et de la conception, quand nous avons une certaine idée des corrélations intra-intervieweur et spatiale.

L'estimation fiable de ρ_{int} et ρ_c est importante. Bien que certains articles publiés traitent de cette estimation, il est sans aucun doute nécessaire de faire progresser la recherche dans ce domaine important. En comparant les deux sources d'homogénéité, Hansen et coll. (1961) ont constaté que la variabilité d'intervieweur était souvent plus

grande que la variabilité d'échantillonnage. Dans de nombreuses enquêtes, ce genre d'évaluation, qui nécessite l'estimation des corrélations intra-intervieweur et intra-grappe, est difficile, voire même impossible, parce que les effets d'intervieweur sont souvent confondus avec les effets de grappes spatiales. L'utilisation d'un plan à échantillons superposés, proposé pour la première fois par Mahalanobis (1946), dans lequel les personnes interrogées sont affectées aléatoirement aux intervieweurs, est un moyen de contourner le problème. En pratique, la mise en œuvre d'un plan de ce genre dans une enquête par sondage à grande échelle est difficile, mais il est possible d'appliquer certains plans à échantillons superposés approximatifs (Hansen et coll. 1961, Bailar, Bailey et Stevens 1977, Bailey, Moore et Bailar 1978, Collins et Butcher 1982, O'Muircheartaigh et Campanelli 1998). Les modèles multiniveaux ont été utilisés comme remède partiel au problème (Hox et De Leeuw 1994, Davis et Scott 1995, O'Muircheartaigh et Campanelli 1998, Scott et Davis 2001). Nous n'avons pas envisagé la question de l'estimation des corrélations intra-intervieweur et intra-grappe. Il s'agit d'un problème important dont nous traiterons dans un futur article.

En pratique, les effets d'intervieweur ou de plan sont calculés pour de nombreuses questions en utilisant la même formule, et une mesure sommaire, telle que l'effet médian d'intervieweur ou de plan, est utilisée pour la planification et la conception de l'enquête. En ce qui concerne les problèmes associés au traitement de questions multiples, le concepteur d'enquêtes peut poursuivre son propre protocole ; la seule modification que nous pourrions proposer consiste à utiliser nos nouvelles définitions de l'effet d'intervieweur ou de l'effet global dans la mesure du possible. L'utilisation de notre formule peut suggérer des effets globaux qui sont nettement plus faibles que ceux indiqués par la formule standard. Cela, à son tour, pourrait suggérer une plus petite taille d'échantillon et, par conséquent, permettre de réaliser une économie.

Annexe

Résultat 1. $$ieff_w = \frac{\mathrm{Var}_{M_2}(\bar{y}_w)}{\mathrm{Var}_{M_1}(\bar{y}_w)} = 1 + \rho_{\mathrm{int}} \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right).$$

Preuve : Le résultat s'obtient en notant que

$$\mathrm{Var}_{M_1}(\bar{y}_w) = \mathrm{Var}_{M_1} \left[\frac{\sum_i \sum_k w_{ik} y_{ik}}{\sum_i \sum_k w_{ik}} \right] = \frac{\sigma^2 \sum_i \sum_k w_{ik}^2}{\left(\sum_i \sum_k w_{ik} \right)^2},$$

et que

$$\mathrm{Var}_{M_2}(\bar{y}_w) = \frac{\sigma^2 \left[\sum_i \sum_k w_{ik}^2 + \rho_{\mathrm{int}} \sum_i \sum_{k \neq k'} w_{ik} w_{ik'} \right]}{\left(\sum_i \sum_k w_{ik} \right)^2},$$

et en effectuant certaines opérations algébriques.

Corollaire : Supposons que $\rho_{\mathrm{int}} > 0$ et $w_{ik} = 1/n$. En utilisant le résultat 1 et l'inégalité de Cauchy-Schwarz, nous obtenons

$$ieff(\mathbf{a}_1) = 1 + \rho_{\mathrm{int}} \left(\frac{\sum_i n_i^2}{n} - 1 \right) \geq 1 + \rho_{\mathrm{int}} \left(\frac{n}{I} - 1 \right) = ieff.$$

Résultat 2. $ieff_w \leq ieff(\mathbf{a}_2)$, où

$$\mathbf{a}_2 = (a_{21}, \ldots, a_{2I}) \text{ avec } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2}.$$

Preuve : En utilisant l'inégalité de Cauchy-Schwarz, nous avons

$$\sum_i \left(\sum_k w_{ik} \right)^2 \leq \sum_i n_i \sum_k w_{ik}^2$$

avec égalité si et uniquement si $w_{ik} = \bar{w}_i$ pour tout i et k, où

$$\bar{w}_i = \frac{\sum_{k=1}^{n_i} w_{ik}}{n_i}$$

est le poids de sondage moyen pour le $i^{\text{ième}}$ intervieweur. Donc, nous avons $ieff_w \leq 1 + [\bar{n}_{\mathrm{int}}(\mathbf{a}_2) - 1]\rho_{\mathrm{int}} = ieff(\mathbf{a}_2)$.

L'égalité est vérifiée si et uniquement si $w_{ik} = \bar{w}_i$ pour tout i et k, auquel cas $ieff_w = ieff(\mathbf{a}_2^*)$, où

$$\mathbf{a}_2^* = (a_{21}^*, \ldots, a_{2I}^*), \text{ avec } a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}.$$

Si tous les poids sont non négatifs, alors

$$\sigma_i^2 = \frac{1}{n_i} \sum_k (w_{ik} - \bar{w}_i)^2 \leq (n_i - 1)\bar{w}_i^2,$$

puisque σ_i^2 est Schur-convexe. Définir

$$x_i = \frac{1 + \frac{\sigma_i^2}{\bar{w}_i^2}}{n_i} \text{ implique que } \frac{1}{n_i} \leq x_i \leq 1$$

et

$$\bar{n}_{\text{int}}(\mathbf{a}_2) = \frac{\sum_i n_i \sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2} = \frac{\sum_i n_i^2 \bar{w}_i^2 + \sum_i n_i^2 \sigma_i^2}{\sum_i n_i \bar{w}_i^2 + \sum_i n_i \sigma_i^2}$$

$$= \frac{\sum_i n_i^3 \bar{w}_i^2 - \sum_i n_i^2 ((n_i - 1)\bar{w}_i^2 - \sigma_i^2)}{\sum_i n_i^2 \bar{w}_i^2 - \sum_i n_i ((n_i - 1)\bar{w}_i^2 - \sigma_i^2)}$$

$$= \frac{\sum_i n_i^3 \bar{w}_i^2 x_i}{\sum_i n_i^2 \bar{w}_i^2 x_i} \leq \frac{\sum_i n_i^3 \bar{w}_i^2}{\sum_i n_i^2 \bar{w}_i^2} = \sum_i n_i \frac{n_i^2 \bar{w}_i^2}{\sum_i n_i^2 \bar{w}_i^2}$$

avec égalité si et uniquement si $\sigma_i^2 = (n_i - 1)\bar{w}_i^2$ pour tout i ou si tous les n_i sont égaux.

L'inégalité découle de la concavité logarithmique de $\bar{n}_{\text{int}}(\mathbf{a}_2)$ en tant que fonction de $(x_1, ..., x_I)$.

Résultat 3. Pour $\mathbf{a}_2^* = (a_{21}^*, ..., a_{2I}^*)$ avec $a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}$

et

$$\mathbf{a}_2 = (a_{21}, ..., a_{2I}) \text{ avec } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2},$$

nous avons

$$ieff(\mathbf{a}_2^*) \underset{\geq}{\leq} ieff(\mathbf{a}_2) \text{ si et uniquement si}$$

$$\sum_i n_i \sigma_i^2 \sum_i n_i^2 \bar{w}_i^2 \underset{\geq}{\leq} \sum_i n_i^2 \sigma_i^2 \sum_i n_i \bar{w}_i^2.$$

Preuve. Nous avons

$$ieff(\mathbf{a}_2^*) - ieff(\mathbf{a}_2) = \frac{\sum_i n_i \sigma_i^2 \sum_i n_i^2 \bar{w}_i^2 - \sum_i n_i^2 \sigma_i^2 \sum_i n_i \bar{w}_i^2}{\left(\sum_i n_i \sigma_i^2 + \sum_i n_i \bar{w}_i^2\right) \sum_i n_i \bar{w}_i^2}.$$

Pour $n_i = n/I$ pour tout i, nous obtenons

$$ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_2).$$

Pour $w_{ik} = \bar{w}_i$ pour tout i, c'est-à-dire $\sigma_i^2 = 0$, nous obtenons

$$ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_2).$$

Pour $\bar{w}_i = \bar{w}$ pour tout i et $\sigma_i^2 = \sigma^2 > 0$ pour tout i, nous obtenons

$$ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_2).$$

Pour $\bar{w}_i = \bar{w}$ pour tout i, nous obtenons

$$ieff(\mathbf{a}_2^*) \underset{\geq}{\leq} ieff(\mathbf{a}_2) \text{ ssi } \sum_i n_i \sigma_i^2 \sum_i n_i^2 \underset{\geq}{\leq} n \sum_i n_i^2 \sigma_i^2.$$

Pour $\sigma_i^2 = \sigma^2 > 0$ pour tout i, nous obtenons

$$ieff(\mathbf{a}_2^*) \underset{\geq}{\leq} ieff(\mathbf{a}_2) \text{ ssi } n \sum_i n_i^2 \bar{w}_i^2 \underset{\geq}{\leq} \sum_i n_i \bar{w}_i^2 \sum_i n_i^2.$$

Résultat 4. Nous avons

$$ieff_w - ieff = \frac{\bar{n}_{\text{int}}}{SCT + n\bar{w}^2} \left[\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SCI \right] \rho_{\text{int}}$$

$$= \frac{\bar{n}_{\text{int}}}{(1 + CV_w^{-2}) SCT} \left[\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SCI \right] \rho_{\text{int}}$$

$$= \frac{\bar{n}_{\text{int}} \tau_w}{1 + CV_w^{-2}} \left(\frac{\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2}{SCI} - 1 \right) \rho_{\text{int}}.$$

Preuve.

$$ieff_w - ieff$$

$$= 1 + \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right) \rho_{\text{int}} - 1 - (\bar{n}_{\text{int}} - 1) \rho_{\text{int}}$$

$$= \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - \bar{n}_{\text{int}} \right) \rho_{\text{int}}$$

$$= \left(\frac{\sum_i n_i^2 \bar{w}_i^2}{SCT + n\bar{w}^2} - \bar{n}_{\text{int}} \right) \rho_{\text{int}}$$

$$= \frac{\bar{n}_{\text{int}}}{SCT + n\bar{w}^2} \left(\frac{\sum_i n_i^2 \bar{w}_i^2}{\bar{n}_{\text{int}}} - (SCT + n\bar{w}^2) \right) \rho_{\text{int}}$$

$$= \frac{\bar{n}_{\text{int}}}{SCT + n\bar{w}^2} \left(\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 + \sum_{i=1}^{I} n_i \bar{w}_i^2 - (SCT + n\bar{w}^2) \right) \rho_{\text{int}}.$$

Maintenant, le résultat s'ensuit en effectuant certaines opérations algébriques.

Résultat 5.

$$ieff_{s,w} = \frac{\mathrm{Var}_{M_4}\left(\overline{y}_w\right)}{\mathrm{Var}_{M_3}\left(\overline{y}_w\right)}$$

$$= 1 + \rho_{\mathrm{int}} \frac{\dfrac{\sum_{p=1}^{P}\sum_{i}^{I_P}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2} - 1}{1 + \rho_C\left(\dfrac{\sum_{p=1}^{P}\left(\sum_{i}^{I_P}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2} - 1\right)}.$$

Preuve. Le résultat s'obtient en notant que

$$\frac{\mathrm{Var}_{M_4}\left(\overline{y}_w\right)}{\mathrm{Var}_{M_3}\left(\overline{y}_w\right)} =$$

$$\frac{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2 + \rho_C \sum_{p=1}^{P}\sum_{i\neq i'}\sum_{k}^{n_{pi}}\sum_{k'}^{n_{pi'}} w_{pik} w_{pi'k'} + \rho \sum_{p=1}^{P}\sum_{i}^{I_P}\sum_{k\neq k'}^{n_{pi}} w_{pik} w_{pik'}}{\sum_{p=1}^{P}\left(\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2 + \rho_C \sum_{i,i'}^{I_p}\sum_{k\neq k'} w_{pik} w_{pi'k'}\right)}$$

et en effectuant certaines opérations algébriques.

Résultat 6. Pour $0 < \rho_C < 1$ et $0 < \rho_{\mathrm{int}} < 1$,

$$eff \geq 1 + \rho_C\left(\frac{n}{P} - 1\right) + \rho_{\mathrm{int}}\left(\frac{n}{I} - 1\right),$$

avec égalité si et uniquement si les poids sont tous égaux et que les intervieweurs ont tous la même charge de travail.

Si, dans chaque UPE, nous n'avons qu'un seul intervieweur, alors

$$eff \geq 1 + (\rho_C + \rho_{\mathrm{int}})\left(\frac{n}{P} - 1\right).$$

Preuve : En utilisant certaines notions algébriques et l'inégalité générale,

$$\sum_j^J p_j x_j^2 \geq \left(\sum_j^J p_j x_j\right)^2$$

avec

$$p_j \geq 0 \text{ et } \sum_{j=1}^{J} p_j = 1,$$

nous avons

$$eff = \frac{n\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}(1 - \rho_C - \rho_{\mathrm{int}})$$

$$+ n\rho_C \frac{\sum_{p=1}^{P}\left(\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2} + n\rho_{\mathrm{int}} \frac{\sum_{p=1}^{P}\sum_{i}^{I_P}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}$$

$$\geq 1 - \rho_C - \rho_{\mathrm{int}} + \rho_C \frac{n}{P} + \rho_{\mathrm{int}} n \frac{I\sum_{p=1}^{P}\frac{I_p}{I}\left(\frac{1}{I_p}\sum_{i}^{I_P}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}$$

$$\geq 1 - \rho_C - \rho_{\mathrm{int}} + \rho_C \frac{n}{P} + \rho_{\mathrm{int}} \frac{n}{I}$$

$$= 1 + \rho_C\left(\frac{n}{P} - 1\right) + \rho_{\mathrm{int}}\left(\frac{n}{I} - 1\right).$$

Remerciements

Nous remercions le rédacteur en chef et les examinateurs de leurs suggestions et commentaires constructifs, qui ont amélioré la version originale de cet article.

Bibliographie

Bailar, B.A., Bailey, L. et Stevens, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.

Bailey, L., Moore, T.F. et Bailar, B.A. (1978). An interviewer variance study for eight impact cities of the National Crime Survey Cities Sample. *Journal of the American Statistical Association*, 73, 16-23.

Biemer, P.P., et Stokes, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 369, 158-166.

Biemer, P., et Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. Dans *Survey Measurement and Process Quality*, (Éds., L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York : John Wiley & Sons, Inc., 603-632.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, 5, 93-105.

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

Collins, M., et Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

Davis, P.D., et Scott, A.J. (1995). La variance de l'intervieweur et ses effets sur les comparaisons de domaines. *Techniques d'enquête*, 21, 111-118.

Feather, J. (1973). A study of interviewer variance. WHO International Collaborative Study of Medical Care Utilization, Saskatchewan Study Area Reports, Séries II, Monographie No. 3.

Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

Gabler, S., Häder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Gray, P.G. (1956). Examples of interviewer variability taken from two sample surveys. *Applied Statistics*, V, 73-85.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.

Groves, R.M., et Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol I, II. New York : John Wiley & Sons, Inc.

Hansen, M.H., Hurwitz, W.N. et Bershad, M.A. (1961). Measurement errors in census and surveys. *Bulletin of the ISI 38*, 2, 351-374.

Hanson, R.H., et Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

Hájek, J. (1971). Comments, In Foundations of Statistical Inference, (Éds. V.P. Godambe et D.A. Sprott). Toronto : Holt, Rinchart, and Winston.

Heeb, J.-L., et Gmel, G. (2001). Interviewers and respondents effects on self-reported alcohol consumption in Swiss Health Survey. *Journal of Studies on Alcohol*, 62, 434-442.

Hox, J.J., et De Leeuw, E.D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modeling to meta-analysis. *Quality & Quantity*, 329-344.

Kalton, G., Brick, J.M. et Lê, T.h. (2005). Estimating components of design effects for use in sample design. Dans : *Household Sample Surveys in Developing and Transition Countries*, Chapitre VI. Disponible au http://unstats.un.org/unsd/hhsurveys/pdf/Chapter_6.pdf.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

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

Lynn, P., et Gabler, S. (2004). Approximations to *b** in the estimation of design effects due to clustering. *Working Papers of the Institute for Social and Economic Research*, papier 2004-07. Colchester : University of Essex. Disponible au http://www. iser.essex.ac.uk/ pubs/workpaps/pdf/2004-07.pdf.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, Série A, 109, 325-378, ré-édité dans *Sankhyā* (1958), 1-68.

O'Muircheartaigh, C., et Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society*, Séries A, 161, 63-77.

Philippens, M., et Loosveldt, G. (2004). Interviewer-related variance in the European Social Survey. Article présenté à la sixième conférence internationale sur la méthodologie des sciences sociales, 17-20 août, Amsterdam.

Rice, S.A. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology*, 35, 420-423.

Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multi-way contingency tables with proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

Schnell, R., et Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.

Scott, A.J., et Davis, P.D. (2001). Estimation des effets de l'intervieweur sur des réponses binaires. *Recueil* : *Symposium 2001*, *La Qualité des Données d'un Organisme Statistique*.

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

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

Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47, 84-95.

Valliant, R.M. (1987). Generalized variance functions in stratified two-stage sampling, 82, 499-508.

Verma, V., Scott, C. et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Série A, 143, 431-473.

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 107-121
Statistique Canada, N° 12-001-X au catalogue

Indicateurs de la représentativité de la réponse aux enquêtes

Barry Schouten, Fannie Cobben et Jelke Bethlehem [1]

Résumé

De nombreux organismes statistiques considèrent le taux de réponse comme étant l'indicateur de la qualité à utiliser en ce qui concerne l'effet du biais de non-réponse. Ils prennent donc diverses mesures en vue de réduire la non-réponse ou de maintenir la réponse à un niveau jugé acceptable. Cependant, à lui seul, le taux de réponse n'est pas un bon indicateur du biais de non-réponse. En général, un taux de réponse élevé n'implique pas que le biais dû à la non-réponse est faible. On trouve à cet égard de nombreux exemples dans la littérature (par exemple, Groves et Peytcheva 2006 ; Keeter, Miller, Kohut, Groves et Presser 2000 ; Schouten 2004).

Nous introduisons un certain nombre de concepts et un nouvel indicateur en vue d'évaluer la similarité entre la réponse à une enquête et l'échantillon de cette enquête. Cet indicateur de la qualité, que nous appelons indicateur R, peut servir de complément aux taux de réponse et est destiné principalement à évaluer le biais de non-réponse. Il peut faciliter l'analyse de la réponse aux enquêtes en fonction du temps, ou pour diverses stratégies d'enquête sur le terrain ou divers modes de collecte des données. Nous appliquons l'indicateur R à deux exemples pratiques.

Mots clés : Qualité ; non-réponse ; réduction de la non-réponse ; correction de la non-réponse.

1. Introduction

Il est bien décrit dans la littérature sur la méthodologie d'enquête que les taux de réponse sont, à eux seuls, de médiocres indicateurs du biais de non-réponse ; voir, par exemple, Curtin, Presser et Singer (2000), Groves, Presser et Dipko (2004), Groves (2006), Groves et Peytcheva (2006), Keeter et coll. (2000), Merkle et Edelman (2002), Heerwegh, Abts et Loosveldt (2007), ainsi que Schouten (2004). Cependant, les spécialistes du domaine n'ont pas encore établi d'autres indicateurs de la non-réponse qui seraient moins ambigus en tant qu'indicateurs de la qualité d'une enquête.

Nous proposons un indicateur, que nous appelons indicateur R (« R » pour représentativité), de la similarité entre la réponse à une enquête, d'une part, et l'échantillon ou la population à l'étude, d'autre part. Cette similarité peut être appelée « réponse représentative ». La littérature spécialisée offre de nombreuses interprétations du concept de « représentativité ». Consulter Kruskal et Mosteller (1979a, b et c) pour une étude approfondie de la littérature statistique et non statistique. Rubin (1976) a introduit le concept de non-réponse ignorable, c'est-à-dire les conditions minimales permettant d'obtenir une estimation sans biais d'une statistique. Certains auteurs définissent explicitement la représentativité. Hájek (1981) relie le terme « représentatif » à l'estimation de paramètres de population ; la paire formée par un estimateur et un mécanisme de création de données manquantes est représentative si, avec une probabilité de un, l'estimateur est égal au paramètre de population. Suivant la définition de Hajèk, les estimateurs par calage (par exemple, Särndal, Swensson et Wretman 2003) sont représentatifs pour les variables auxiliaires qui sont calées. Bertino (2006) définit un indice de représentativité dit univarié pour des variables aléatoires continues. Cet indice est une mesure exempte de distribution basée sur la statistique de Cramér-Von Mises. Kohler (2007) définit ce qu'il appelle un critère interne de représentativité. Ce critère univarié ressemble à la statistique Z pour les moyennes de population.

Nous isolons le concept de représentativité de l'estimation d'un paramètre de population particulier, mais nous le relions à l'effet sur la composition globale de la réponse. Dissocier les indicateurs d'un paramètre particulier permet de les utiliser comme outils pour comparer diverses enquêtes entre elles ou les résultats d'une enquête au cours du temps, ainsi que pour comparer des stratégies et modes différents de collecte des données. En outre, la mesure donne une perspective multidimensionnelle de la dissemblance entre l'échantillon et la réponse.

L'indicateur R que nous proposons s'appuie sur des probabilités de réponse estimées. L'estimation de ces probabilités implique que l'indicateur R proprement dit est une variable aléatoire et, par conséquent, qu'il possède une précision et éventuellement un biais. La taille de l'échantillon d'une enquête joue donc un rôle important dans l'évaluation de l'indicateur R, comme nous le montrerons. Cependant, cette dépendance existe pour toute mesure ; les petites enquêtes ne permettent tout simplement pas de tirer des conclusions catégoriques au sujet du mécanisme de création des données manquantes.

Nous montrons que l'indicateur R proposé est apparenté à la mesure V de Cramèr pour l'association entre la réponse

1. Barry Schouten, Fannie Cobben et Jelke Bethlehem, Statistics Netherlands, Department of Methodology and Quality, PO Box 4000, 2370 JM Voorburg, The Netherlands. Courriel : bstn@cbs.nl.

et les variables auxiliaires. En fait, nous considérons l'indicateur R comme une mesure du manque d'association. La situation est d'autant meilleure que l'association est faible, car cela signifie qu'il n'existe aucune preuve que la non-réponse a affecté la composition des données observées.

Afin de pouvoir utiliser les indicateurs R comme outils de surveillance et de comparaison de la qualité des enquêtes dans l'avenir, ceux-ci doivent avoir les caractéristiques d'une mesure. Autrement dit, nous voulons un indicateur R qui peut être interprété, mesuré et normalisé tout en ayant les propriétés mathématiques d'une mesure. Et ce, surtout parce que l'interprétation et la normalisation ne sont pas des caractéristiques directes.

Nous appliquons l'indicateur R à deux études menées à Statistiques Pays-Bas en 2005 et en 2006. Elles avaient pour objectifs de comparer différentes stratégies de collecte des données. Elles comportaient divers modes de collecte des données et diverses stratégies de suivi des cas de non-réponse. Chacune de ces études a été suivie d'une analyse détaillée des données et de la production de rapports. Par conséquent, elles conviennent bien pour une validation empirique de l'indicateur R. Nous comparons les valeurs de l'indicateur R aux conclusions des analyses. Nous renvoyons le lecteur à Schouten et Cobben (2007), ainsi qu'à Cobben et Schouten (2007) pour d'autres exemples et études empiriques.

À la section 2, nous commençons par discuter du concept de réponse représentative. Puis, à la section 3, nous définissons la notation mathématique pour notre indicateur. La section 4 est consacrée aux caractéristiques de l'indicateur R. La section 5 décrit l'application de cet indicateur R aux études sur le terrain. Enfin, à la section 6, nous présentons une discussion.

2. Le concept de réponse représentative

En premier lieu, nous expliquons ce que nous entendons quand nous disons qu'un ensemble de répondants à une enquête est représentatif de l'échantillon. Ensuite, nous rendons le concept de représentativité mathématiquement rigoureux en lui donnant une définition.

2.1 Que signifie représentatif ?

La littérature nous avertit de ne pas concentrer tous nos efforts sur les taux de réponse pour évaluer la qualité des enquêtes. Nous pouvons illustrer facilement ce point au moyen d'un exemple tiré de l'enquête néerlandaise POLS (acronyme de *Permanent Onderzoek Leefsituatie* ou Enquête intégrée sur les conditions de vie des ménages en français).

Le tableau 1 contient les estimations, fondées sur les réponses à l'enquête POLS après un mois et après deux mois, de la proportion de la population néerlandaise qui recevait une forme d'allocations sociales et de la proportion dont au moins un parent était né ailleurs qu'aux Pays-Bas. Pour les deux variables, les données sont extraites de registres et sont traitées artificiellement comme des réponses à une enquête en supprimant leurs valeurs pour les non-répondants. Les proportions dans l'échantillon sont également données au tableau 1. Après un mois, le taux de réponse était de 47,2 %, tandis qu'après la période complète d'interview de deux mois, il était de 59,7 %. Dans le cas de l'enquête POLS de 1998, le premier mois, la collecte a été effectuée par IPAO (interview sur place assistée par ordinateur). Après le premier mois, les non-répondants ont été affectés à un mode de collecte par ITAO (interview téléphonique assistée par ordinateur) s'ils étaient abonnés à un service téléphonique par fil dont le numéro était publié. Sinon, ils ont de nouveau été affectés à un mode de collecte par IPAO. Donc, le deuxième mois d'interview a donné 12,5 % supplémentaires de réponse. Cependant, l'examen du tableau 1 montre qu'après le deuxième mois, les estimations fondées sur les données de l'enquête présentent un biais plus important qu'après le premier mois.

Tableau 1
Moyennes des réponses à l'enquête POLS pour le premier mois de collecte et pour la période complète de collecte de deux mois

Variable	Après 1 mois	Après 2 mois	Échantillon
Recevant des allocations sociales	10,5 %	10,4 %	12,1 %
Parent non natif	12,9 %	12,5 %	15,0 %
Taux de réponse	47,2 %	59,7 %	100 %

L'exemple semble montrer clairement que l'effort accru a abouti à une réponse moins représentative en ce qui a trait aux deux variables auxiliaires. Mais d'une manière générale, qu'entendons-nous par représentatif ?

Le terme « représentatif » est souvent utilisé avec hésitation dans la littérature statistique. Kruskal et Mosteller (1979a, b et c) procèdent à un inventaire approfondi de l'utilisation du mot « représentatif » dans la littérature et relève neuf interprétations distinctes. Un certain nombre de ces interprétations sont omniprésentes dans les textes statistiques. Les interprétations statistiques que Kruskal et Mosteller ont nommées « absence de forces sélectives », « miniature de la population » et cas « typiques ou idéaux » se rapportent à l'échantillonnage probabiliste, l'échantillonnage par quota et l'échantillonnage raisonné. À la section suivante, nous proposons une définition de la représentativité qui correspond à l'« absence de forces sélectives ». Mais avant cela, nous expliquons pourquoi nous faisons ce choix.

Le concept de réponse représentative est aussi relié étroitement aux mécanismes de création des données manquantes désignés MCAR (pour *Missing-Completely-at-Random* ou Manquent entièrement au hasard), MAR (pour *Missing-at-Random* ou Manquent au hasard) et NMAR (pour *Not-Missing-at-Random* ou Ne manquent pas au hasard) qui sont souvent mentionnés dans la littérature ; voir Little et Rubin (2002). Un mécanisme de création de données manquantes est de type MCAR quand la probabilité de réponse ne dépend pas du sujet d'enquête d'intérêt. Le mécanisme est de type MAR si la probabilité de réponse dépend des données observées seulement, ce qui est donc une hypothèse plus faible que l'hypothèse MCAR. Si la probabilité dépend également des données manquantes, le mécanisme est de type NMAR. En fait, ces mécanismes ont pour origine la théorie statistique fondée sur la modélisation. Interprété plus ou moins librement en ce qui a trait à un sujet d'enquête, MCAR signifie que les répondants sont, en moyenne, les mêmes que les non-répondants, MAR signifie que, au sein de sous-populations connues, les répondants sont en moyenne les mêmes que les non-répondants, et NMAR implique que, même au sein de sous-populations, les répondants sont différents des non-répondants. L'ajout du sujet d'enquête est essentiel. Dans un même questionnaire, certains items peuvent être de type MCAR, tandis que d'autres sont de type MAR ou NMAR. En outre, l'hypothèse d'un mécanisme MAR pour un item tient pour une stratification particulière de la population. Un autre item pourrait nécessiter une stratification différente.

Étant donné que nous désirons surveiller et comparer la réponse à des enquêtes qui diffèrent par le sujet étudié ou par la période où elles sont exécutées, il n'est pas intéressant de définir une réponse représentative comme étant dépendante du sujet d'enquête proprement dit ni comme étant dépendante de l'estimateur utilisé. Nous préférons nous concentrer sur la qualité de la collecte des données plutôt que sur l'estimation. Ces conditions nous amènent à comparer la composition de la réponse (groupe de répondants) à celle de l'échantillon. Clairement, les sujets étudiés influencent la probabilité que les ménages participent à l'enquête, mais cette influence ne peut pas être mesurée ni testée et, par conséquent, de notre point de vue, ils ne peuvent pas représenter les données d'entrée utilisées pour évaluer la qualité de la réponse. Nous proposons de juger la composition de la réponse au moyen d'ensembles prédéfinis de variables qui sont observés en dehors du cadre de l'enquête et qui peuvent être employés pour chaque enquête examinée. Nous voulons que la sélection des répondants soit aussi proche que possible d'un « échantillon aléatoire simple de l'échantillon utilisé pour l'enquête », c'est-à-dire présentant une relation aussi faible que possible entre la réponse et les caractéristiques qui distinguent les unités les unes des autres. Le dernier énoncé peut être interprété comme signifiant l'absence de forces sélectives dans la sélection des répondants, ou comme un mécanisme de type MCAR en ce qui concerne toutes les variables étudiées possibles.

2.2 Définition d'un sous-ensemble de réponses représentatif

Soit $i = 1, 2, 3, \ldots, N$ les étiquettes d'unité pour la population. Nous désignons par s_i l'indicateur d'échantillon 0-1, c'est-à-dire que l'unité i prend la valeur 1 si elle est échantillonnée, et 0 autrement. Nous désignons par r_i l'indicateur de réponse 0-1 pour l'unité i. Si l'unité i est échantillonnée et qu'elle répond, alors $r_i = 1$. Sinon, sa valeur est 0. La taille d'échantillon est n. Enfin, π_i désigne la probabilité d'inclusion de premier ordre de l'unité i.

La clé de nos définitions réside dans les propensions à répondre individuelles. Soit ρ_i la probabilité que l'unité i réponde quand elle est échantillonnée.

L'interprétation d'une propension à répondre n'est pas directe. Nous suivons une approche assistée par modèle, ce qui signifie que seuls l'échantillon et les indicateurs de réponse ont un caractère aléatoire. Une probabilité de réponse est une caractéristique d'une unité étiquetée et identifiable, si l'on peut dire une pièce de monnaie biaisée que l'unité garde en poche, et est, par conséquent, inséparable de cette unité. Toutefois, moyennant un peu d'effort, tous les concepts peuvent être traduits dans un contexte fondé sur un modèle.

Pour commencer, nous donnons une définition forte.

Définition (*forte*) : *Un sous-ensemble de réponses est représentatif de l'échantillon si les propensions à répondre* ρ_i *sont les mêmes pour toutes les unités de la population*

$$\rho_i = P[r_i = 1 \mid s_i = 1] = \rho, \quad \forall i, \qquad (1)$$

et si la réponse d'une unité est indépendante de la réponse de toutes les autres unités.

Si un mécanisme de données manquantes satisfait la définition forte, il sera du type Manquent entièrement au hasard (MCAR) pour toutes les questions d'enquête possibles. Bien que la définition soit séduisante, il est impossible de tester sa validité en pratique. Nous ne possédons aucune répétition de la réponse d'une unité unique. Par conséquent, nous construisons aussi une définition faible qui peut être testée en pratique.

Définition (*faible*) : *Un sous-ensemble de réponses est représentatif d'une variable catégorique* X *possédant* H *catégories si la propension à répondre moyenne sur les catégories est constante, soit*

$$\bar{\rho}_h = \frac{1}{N_h}\sum_{k=1}^{N_h} \rho_{hk} = \rho, \quad \text{pour} \quad h = 1, 2, \ldots, H, \qquad (2)$$

où N_h est la taille de la population de la catégorie h, ρ_{hk} est la propension à répondre de l'unité k dans la classe h et la somme est faite sur toutes les unités dans cette catégorie.

La définition faible correspond à un mécanisme de création de données manquantes de type MCAR par rapport à X, car selon l'hypothèse MCAR, nous ne pouvons pas faire la distinction entre les répondants et les non-répondants en nous fondant sur la connaissance de X.

3. Indicateurs R

À la section précédente, nous avons défini une réponse fortement représentative et faiblement représentative. Les deux définitions s'appuient sur les probabilités de réponse individuelles qui sont inconnues en pratique. Pour commencer, nous prenons un indicateur R de population. Puis, nous basons le même indicateur R sur un échantillon et sur les propensions à répondre estimées.

3.1 Indicateurs R de population

Nous considérons d'abord la situation hypothétique où les propensions à répondre individuelles sont connues. Manifestement, dans ce cas, nous pouvons même tester la définition forte et nous cherchons simplement à mesurer la quantité de variation dans les propensions à répondre ; la représentativité au sens fort est d'autant moindre que la variation est forte. Soit $\rho = (\rho_1, \rho_2, \ldots, \rho_N)'$ un vecteur de propensions à répondre, soit $\mathbf{1} = (1, 1, \ldots, 1)'$ le vecteur de dimension N de valeurs 1, et soit $\rho_0 = \mathbf{1} \times \bar{\rho}$ le vecteur constitué de la propension à répondre moyenne de la population.

Toute fonction de distance d dans $[0, 1]^N$ suffirait à mesurer l'écart par rapport à une réponse fortement représentative en calculant $d(\rho, \rho_0)$. Notons que la hauteur de la réponse globale ne joue aucun rôle. La distance euclidienne est une fonction de distance simple. Appliquée à une distance entre ρ et ρ_0, cette mesure est proportionnelle à l'écart-type des probabilités de réponse

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (\rho_i - \bar{\rho})^2}. \qquad (3)$$

Il n'est pas difficile de montrer que

$$S(\rho) \le \sqrt{\bar{\rho}(1-\bar{\rho})} \le \frac{1}{2}. \qquad (4)$$

Nous voulons que l'indicateur R prenne les valeurs comprises dans l'intervalle [0, 1], la valeur 1 représentant une représentativité forte et la valeur 0 étant l'écart maximal par rapport à la représentativité forte. Nous proposons l'indicateur R, que nous désignons par R, défini par

$$R(\rho) = 1 - 2S(\rho). \qquad (5)$$

Notons que la valeur minimale de (5) dépend du taux de réponse (voir la figure 1). Pour $\bar{\rho} = 1/2$, il possède une valeur minimale de 0. Pour $\bar{\rho} = 0$ et $\bar{\rho} = 1$, aucune variation n'est manifestement possible et la valeur minimale est 1. Paradoxalement, la borne inférieure augmente quand le taux de réponse diminue pour passer de $1/2$ à 0. Dans le cas d'un taux de réponse faible, la possibilité que la variation des propensions à répondre individuelles soit forte est moindre.

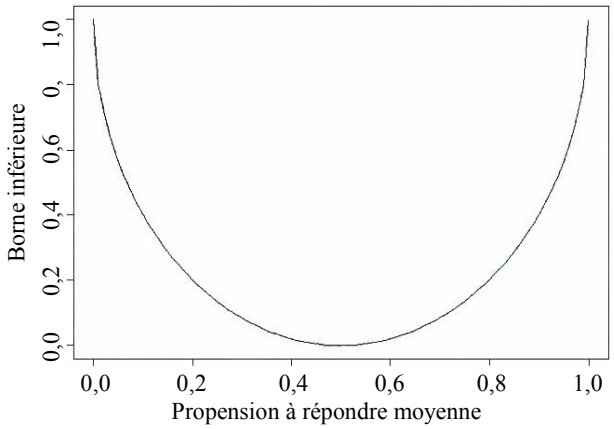


Figure 1 Valeur minimale de l'indicateur R (5) en fonction de la propension à répondre moyenne

Nous pouvons considérer R comme étant une mesure du manque d'association. Quand $R(\rho) = 1$, il n'existe aucune relation entre tout item d'une enquête et le mécanisme de création de données manquantes. Nous montrons que R est en fait étroitement lié à la statistique χ^2 bien connue qui est souvent utilisée pour tester l'indépendance et l'adéquation.

Supposons que les propensions à répondre diffèrent uniquement pour les classes h définies par une variable catégorique X. Soit $\bar{\rho}_h$ et f_h, respectivement, la propension à répondre et la fonction de population de la classe h, c'est-à-dire

$$f_h = \frac{N_h}{N}, \quad \text{pour} \quad h = 1, 2, \ldots, H. \qquad (6)$$

Donc, pour tout i avec $X_i = h$, la propension à répondre est $\rho_i = \bar{\rho}_h$.

Puisque la variance des propensions à répondre est la somme des variances « interclasses » et « intraclasse » sur

l'ensemble des classes h, et que nous supposons que les variances intraclasse sont nulles, il est vérifié que

$$S^2(\tilde{\rho}) = \frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2$$

$$= \frac{N}{N-1}\sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2 \approx \sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2. \qquad (7)$$

La statistique χ^2 mesure la distance entre les proportions observées et prévues. Cependant, il ne s'agit d'une fonction de distance vraie qu'au sens mathématique pour les distributions marginales fixes f_h et $\bar{\rho}$. Nous pouvons appliquer la statistique χ^2 à X afin de « mesurer » la distance entre le comportement de réponse réel et le comportement de réponse qui est attendu quand la réponse est indépendante de X. Autrement dit, nous mesurons l'écart par rapport à la représentativité faible en ce qui concerne X.

Nous pouvons réécrire la statistique χ^2 pour obtenir

$$\chi^2 = \sum_{h=1}^{H} \frac{(N_h\bar{\rho}_h - N_h\bar{\rho})^2}{N_h\bar{\rho}}$$

$$+ \sum_{h=1}^{H} \frac{(N_h(1-\bar{\rho}_h) - N_h(1-\bar{\rho}))^2}{N_h(1-\bar{\rho})}$$

$$= \sum_{h=1}^{H} \frac{Nf_h(\bar{\rho}_h - \bar{\rho})^2}{\bar{\rho}} + \sum_{h=1}^{H} \frac{Nf_h(\bar{\rho}_h - \bar{\rho})^2}{(1-\bar{\rho})}$$

$$= \frac{N}{\bar{\rho}(1-\bar{\rho})}\sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2$$

$$= \frac{N-1}{\bar{\rho}(1-\bar{\rho})} S^2(\tilde{\rho}). \qquad (8)$$

Une mesure d'association qui converti la statistique χ^2 à l'intervalle [0, 1] (voir, par exemple, Agresti 2002) est le V de Cramèr

$$V = \sqrt{\frac{\chi^2}{N(\min\{C, R\} - 1)}}, \qquad (9)$$

où C et R sont, respectivement, le nombre de colonnes et de lignes dans la table de contingence sous-jacente. Le V de Cramèr atteint une valeur 0 si les proportions observées concordent exactement avec les proportions espérées et sa valeur maximale est 1. Dans notre cas, le dénominateur est égal à N puisque l'indicateur de réponse ne possède que deux catégories : réponse et non-réponse. Par conséquent, (9) se transforme en

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N-1}{N\bar{\rho}(1-\bar{\rho})}}\, S(\tilde{\rho}). \qquad (10)$$

Partant de (10), nous voyons que, si N est grand, le V de Cramèr est approximativement égal à l'écart-type des propensions à répondre normalisées par l'écart-type maximal $\sqrt{\bar{\rho}(1-\bar{\rho})}$ pour une propension à répondre moyenne fixe $\bar{\rho}$.

3.2 Indicateurs R fondés sur la réponse

À la section 3.1, nous avons supposé que les propensions à répondre individuelles étaient connues. Naturellement, en pratique, elles ne le sont pas. En outre, dans un sondage, nous ne possédons des renseignements que sur le comportement de réponse des unités échantillonnées. Par conséquent, nous devons trouver d'autres solutions pour les indicateurs R. Un moyen évident consiste à utiliser des estimateurs fondés sur la réponse pour les propensions à répondre individuelles et la propension à répondre moyenne.

Soit $\hat{\rho}_i$ un estimateur de ρ_i qui utilise toutes les variables auxiliaires disponibles ou un sous-ensemble de celles-ci. Les méthodes qui permettent de calculer ce genre d'estimation sont, par exemple, les modèles de régression logit ou probit (Agresti 2002) et les arbres de classification CHAID (Kass 1980). Soit $\hat{\bar{\rho}}$ la moyenne pondérée d'échantillon des propensions à répondre estimées, c'est-à-dire

$$\hat{\bar{\rho}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\rho}_i \frac{s_i}{\pi_i}, \qquad (11)$$

où nous utilisons les poids d'inclusion.

Nous remplaçons R par l'estimateur $\hat{R}$

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\bar{\rho}})^2}. \qquad (12)$$

Notons que, dans (12), il existe en fait deux étapes d'estimation fondées sur des mécanismes probabilistes différents. Les propensions à répondre proprement dites sont estimées, ainsi que la variation dans les propensions. Nous revenons sur les conséquences de l'estimation en deux étapes à la section 4.

3.3 Exemple

Nous appliquons les indicateurs R proposés à des données provenant de l'enquête POLS de 1998 décrite à la section 2.1. Rappelons que l'enquête consistait en une combinaison d'interviews sur place et d'interviews téléphoniques dans laquelle les interviews du premier mois étaient de type IPAO uniquement. La taille de l'échantillon était de près de 40 000 et le taux de réponse était de 60 % environ. Nous avons relié des données administratives concernant le travail sur le terrain à l'échantillon et déduit si chaque tentative de prise de contact avait abouti à une réponse, ce

qui nous a permis de suivre la courbe de l'indicateur R durant la période de collecte des données.

Pour l'estimation des taux de réponse, nous nous sommes servis d'un modèle de régression logistique comprenant la région, les antécédents ethniques et l'âge comme variables indépendantes. La région était une classification à 16 catégories, c'est-à-dire les 12 provinces et les quatre villes les plus grandes (Amsterdam, Rotterdam, La Haye et Utrecht) comme catégories distinctes. Les antécédents ethniques étaient répartis en sept catégories : natif, Marocain, Turc, Surinamais, Néerlando-antillais, autre non-natif non occidental et autre non-natif occidental. La classification est fondée sur le pays de naissance des parents de la personne sélectionnée. La variable d'âge comprend trois catégories : de 0 à 34 ans, de 35 à 54 ans et de 55 ans et plus.

À la figure 2, $\hat{R}$ est représenté graphiquement en fonction du taux de réponse pour les six premières tentatives de prise de contact durant l'enquête POLS. La valeur la plus à gauche correspond à l'ensemble de répondants après une tentative. Pour chaque tentative supplémentaire, le taux de réponse augmente, mais l'indicateur révèle une diminution de la représentativité. Ce résultat confirme les constatations faites par Schouten (2004).

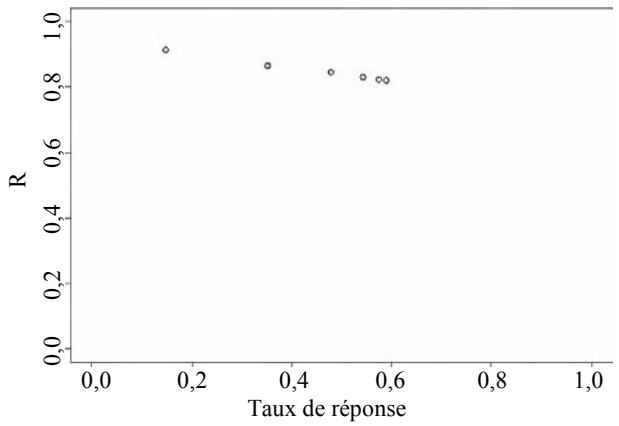


Figure 2 Indicateur R pour les six premières tentatives de prise de contact dans l'enquête POLS de 1998

4. Caractéristiques des indicateurs R

À la section 3, nous proposons un indicateur de la représentativité. Cependant, il est possible d'en construire d'autres. Les mesures d'association ou les indices d'adéquation sont nombreux, par exemple, Goodman et Kruskal (1979), Bentler (1990) et Marsh, Balla et McDonald (1988). La relation entre les mesures d'association et les indicateurs R est forte. Essentiellement, ces indicateurs ont pour objectif de mesurer, dans des conditions multivariées, le manque d'association. À la présente section, nous discutons des caractéristiques souhaitées des indicateurs R. Nous montrons que l'indicateur R proposé permet d'obtenir une borne supérieure simple pour le biais de non-réponse.

4.1 Caractéristiques générales

Nous voulons que les indicateurs R soient fondés sur une fonction ou une mesure de distance au sens mathématique. La propriété d'inégalité triangulaire d'une fonction de distance permet un classement partiel de la variation dans les propensions à répondre qui facilite l'interprétation. Une fonction de distance peut être dérivée facilement de toute norme mathématique. À la section 3, nous choisissons la norme euclidienne qui est celle utilisée couramment. Elle nous mène à un indicateur R qui s'appuie sur l'écart-type des propensions à répondre. D'autres normes, comme la norme supremum, ou norme sup, nous mèneraient à d'autres fonctions de distance. Toutefois, à la section 4.3, nous montrons que les indicateurs R fondés sur la norme euclidienne ont des caractéristiques de normalisation intéressantes.

Nous devons faire une distinction subtile entre les indicateurs R et les fonctions de distance. Ces dernières sont symétriques, tandis qu'un indicateur R mesure un écart par rapport à un point particulier, à savoir celui où toutes les propensions à répondre sont égales. Si nous modifions le vecteur des propensions individuelles, dans la plupart des cas, ce point se déplace. Toutefois, si nous fixons la valeur de la propension à répondre moyenne, la fonction de distance facilite l'interprétation.

À part une relation avec une fonction de distance, nous voulons pouvoir mesurer, interpréter et normaliser les indicateurs R. À la section 3.2, nous avons déjà dérivé des estimateurs fondés sur la réponse pour les indicateurs R « de population » qui ne sont pas mesurables quand les propensions à répondre sont inconnues et que l'on ne dispose que des réponses à un sondage. Donc, nous avons rendu les indicateurs R mesurables en passant à des estimateurs. Les deux autres caractéristiques sont examinées séparément dans les deux sections qui suivent.

4.2 Interprétation

La deuxième caractéristique des indicateurs R est la facilité avec laquelle nous pouvons interpréter leur valeur et le concept qu'ils mesurent. Nous sommes passés à un estimateur pour un indicateur R qui est fondé sur les échantillons des sondages et sur des estimateurs de probabilité de réponse individuelle. Ces deux éléments ont des conséquences très importantes en ce qui concerne l'interprétation et la comparaison des indicateurs R.

Puisque l'indicateur R est lui-même un estimateur, il est aussi une variable aléatoire. Autrement dit, il dépend de

l'échantillon, c'est-à-dire qu'il peut présenter un biais et qu'il possède une certaine exactitude. Mais qu'estime-t-il ?

Supposons d'abord que la taille d'échantillon est arbitrairement grande de sorte que la précision ne joue aucun rôle et que la sélection d'un modèle pour les propensions à répondre ne pose pas de problème. Autrement dit, nous sommes capables d'ajuster tout modèle à tout ensemble fixe de variables auxiliaires.

Il existe une forte relation entre l'indicateur R, d'une part, et la disponibilité et l'utilisation de variables auxiliaires, d'autre part. À la section 2, nous avons défini la représentativité forte et faible. Même dans le cas où nous arrivons à ajuster n'importe quel modèle, nous ne sommes pas capables d'estimer les propensions à répondre au-delà du « pouvoir de résolution » des variables auxiliaires disponibles. Donc, nous ne pouvons tirer des conclusions qu'au sujet de la représentativité faible par rapport à l'ensemble de variables auxiliaires. Cela veut dire que, chaque fois que nous utilisons un indicateur R, nous devons utiliser comme complément de sa valeur l'ensemble de covariables qui ont servi de grille pour estimer les propensions à répondre individuelles. Si l'indicateur R est utilisé pour des comparaisons, les ensembles de variables doivent être les mêmes. Il convient de souligner qu'il n'est pas nécessaire d'utiliser toutes les variables auxiliaires pour l'estimation des propensions à répondre, puisqu'elles peuvent ne pas accroître la puissance explicative du modèle. Cependant, des ensembles identiques devraient être disponibles. L'indicateur R mesure alors un écart par rapport à la représentativité faible.

L'indicateur R ne reflète pas les différences entre les probabilités de réponse à l'intérieur d'autres sous-groupes de la population que ceux définis par les classes de X. Si nous désignons de nouveau par $h = 1, 2, \ldots, H$ les strates définies par X, par N_h la taille de la strate h et par $\bar{\rho}_h$ la moyenne de population des probabilités de réponse dans la strate h, il n'est pas difficile de montrer que $\hat{R}$ est une estimateur convergent de

$$R_X(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2}, \qquad (13)$$

quand des modèles standard, comme la régression logistique ou la régression linéaire, sont utilisés pour estimer les probabilités de réponse. Naturellement, les indicateurs (13) et (5) peuvent être différents.

En pratique, la taille d'échantillon n'est pas arbitrairement grande. Elle a une incidence sur les deux étapes d'estimation, c'est-à-dire l'estimation des propensions à répondre et celle de l'indicateur R en utilisant un échantillon.

Si nous connaissions les propensions à répondre individuelles, l'estimation fondée sur l'échantillon de l'indicateur R ne donnerait lieu qu'à une variance et ne produirait pas de biais. Nous serions capables d'estimer l'indicateur R de population sans biais. Donc, pour les petites tailles d'échantillon, les estimateurs auraient une faible précision dont il pourrait être tenu compte en utilisant des intervalles de confiance au lieu de simples estimateurs ponctuels.

Les incidences sur l'estimation des probabilités de réponse sont toutefois différentes, à cause de la sélection et de l'ajustement du modèle. Il existe deux options. Nous pouvons imposer un modèle pour estimer les propensions à répondre en fixant les covariables au préalable ou nous pouvons laisser le modèle dépendre des covariables dont la contribution est significative par rapport à un niveau prédéfini. Dans le premier cas, nous n'introduirons de nouveau aucun biais, mais l'erreur-type pourrait être affectée par un surajustement. Dans le deuxième cas, le modèle pour l'estimation des propensions à répondre dépend de la taille de l'échantillon ; plus l'échantillon est grand, plus le nombre d'interactions acceptées comme étant significatives est élevé. Même si l'ajustement de modèles en se fondant sur un niveau de signification est une pratique statistique courante, la sélection du modèle peut introduire un biais et une variance dans l'estimation de tout indicateur R. Ce point est facile à comprendre si l'on considère la situation extrême d'un échantillon de taille 10. Pour un si petit échantillon, aucune interaction entre le comportement de réponse et les caractéristiques auxiliaires ne sera acceptée, ce qui laisse un modèle vide et un indicateur R estimé de 1. Les petits échantillons ne permettent tout simplement pas d'estimer les propensions à répondre. En général, un échantillon de petite taille conduira donc à une vue plus optimiste de la représentativité.

Nous devons faire ici une autre distinction subtile. Il se pourrait que, dans un sondage, un grand nombre d'interactions réunies contribuent à la prédiction des propensions à répondre, mais que les contributions individuelles soient très faibles, tandis que dans un autre sondage, une seule interaction intervienne, mais que sa contribution soit grande. Il se peut qu'individuellement, aucune petite contribution ne soit significative, mais qu'ensemble, elles soient aussi fortes que la grande contribution unique qui est significative. Donc, nous serions plus optimistes dans le premier exemple, même pour des tailles d'échantillon comparables.

Ces observations montrent qu'il faut toujours utiliser les indicateurs R avec prudence. Ils ne peuvent pas être considérés indépendamment des variables auxiliaires qui ont été utilisées pour les calculer. En outre, la taille de l'échantillon a une incidence sur le biais ainsi que sur la précision.

4.3 Normalisation

La troisième caractéristique importante d'un indicateur R est la normalisation. Nous voulons pouvoir donner des bornes à un indicateur R afin que son échelle, et donc, les variations de sa valeur, aient une signification. De toute

évidence, les problèmes d'interprétation soulevés à la section précédente ont également une incidence sur la normalisation de l'indicateur R. Par conséquent, ici, nous supposons que nous nous trouvons dans la situation idéale où nous pouvons estimer les propensions à répondre sans biais. Cette hypothèse tient pour les grandes enquêtes. Nous discutons la normalisation de l'indicateur R correspondant à $\hat{R}$.

4.3.1 Biais absolu maximal et racine carrée de l'erreur quadratique moyenne

Nous montrons que, pour tout item d'une enquête Y, l'indicateur R peut être utilisé pour imposer des bornes supérieures au biais de non-réponse et à la racine carrée de l'erreur quadratique moyenne (REQM) des moyennes ajustées des réponses. Nous appliquons ces bornes à l'indicateur R pour montrer l'effet sous les pires scénarios.

Soit Y une variable qui est mesurée dans une enquête et soit $\hat{\bar{y}}_{\mathrm{HT}}$ l'estimateur Horvitz-Thompson de la moyenne de population basée sur les réponses à l'enquête. On peut montrer (par exemple, Bethlehem 1988, Särndal et Lundström 2005) que son biais $B(\hat{\bar{y}}_{\mathrm{HT}})$ est approximativement égal à

$$B(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{C(y,\ \rho)}{\bar{\rho}}, \tag{14}$$

avec $C(y,\ \rho) = 1/N \sum_{i=1}^{N}(y_i - \bar{y})\,(\rho_i - \bar{\rho})$ la covariance de population entre les items étudiés et les probabilités de réponse. Pour une approximation étroite de la variance $s^2(\hat{\bar{y}}_{\mathrm{HT}})$ de $\hat{\bar{y}}_{\mathrm{HT}}$, nous nous référons à Bethlehem (1988).

L'inégalité de Cauchy-Schwarz permet de trouver une normalisation de R. Cette inégalité énonce que la covariance entre deux variables quelconques est bornée au sens absolu par le produit des écarts-types des deux variables. Nous pouvons traduire cela en bornes pour le biais (14) de $\hat{\bar{y}}_{\mathrm{HT}}$

$$|B(\hat{\bar{y}}_{\mathrm{HT}})| \le \frac{S(\rho)S(y)}{\bar{\rho}} = \frac{(1 - R(\rho))S(y)}{2\bar{\rho}} = B_m(\rho,\ y). \tag{15}$$

Manifestement, nous ne connaissons pas la borne supérieure $B_m(\rho,\ y)$ dans (15), mais nous pouvons l'estimer en utilisant l'échantillon et les probabilités de réponse estimées. Nous désignons l'estimateur par $\hat{B}_m(\hat{\rho},\ y)$.

De la même manière, nous pouvons établir une borne pour la racine carrée de l'erreur quadratique moyenne (REQM) de $\hat{\bar{y}}_{\mathrm{HT}}$. Il est vérifié approximativement que

$$\mathrm{REQM}(\hat{\bar{y}}_{\mathrm{HT}}) = \sqrt{B^2(\hat{\bar{y}}_{\mathrm{HT}}) + s^2(\hat{\bar{y}}_{\mathrm{HT}})} \le \sqrt{B_m^2(\rho,\ y) + s^2(\hat{\bar{y}}_{\mathrm{HT}})} = E_m(\rho,\ y). \tag{16}$$

De nouveau, nous ne connaissons pas $E_m(\rho,\ y)$. À sa place, nous utilisons l'estimateur fondé sur l'échantillon qui emploie les probabilités de réponse estimées, désignées par $\hat{E}_m(\hat{\rho},\ y)$.

Les bornes $\hat{B}_m(\hat{\rho},\ y)$ et $\hat{E}_m(\hat{\rho},\ y)$ sont différentes pour chaque item y de l'enquête. Pour les comparaisons, il est par conséquent commode de définir un item d'enquête hypothétique. Nous supposons que $\hat{S}(y) = 0{,}5$. Nous désignons les bornes correspondantes par $\hat{B}_m(\hat{\rho})$ et $\hat{E}_m(\hat{\rho})$. Elles sont égales à

$$\hat{B}_m(\hat{\rho}) = \frac{(1 - \hat{R}(\hat{\rho}))}{4\hat{\bar{\rho}}} \tag{17}$$

$$\hat{E}_m(\hat{\rho}) = \sqrt{\hat{B}_m^2(\hat{\rho}) + \hat{s}^2(\hat{\bar{y}}_{\mathrm{HT}})}. \tag{18}$$

Nous calculons (17) et (18) dans le cas de toutes les études décrites à la section 5. Nous devons souligner que (17) et (18) sont de nouveau des variables aléatoires qui possèdent une certaine précision et qui sont susceptibles de présenter un biais.

4.3.2 Fonctions de représentativité de la réponse

À la section précédente, nous avons utilisé l'indicateur R pour déterminer les bornes supérieures du biais de non-réponse et de la racine carrée de l'erreur quadratique moyenne de la moyenne (ajustée) des réponses. Inversement, nous pouvons établir une borne inférieure de l'indicateur R en demandant que le biais absolu de non-réponse ou la racine carrée de l'erreur quadratique moyenne soit inférieur à une valeur prescrite. Cette borne inférieure peut être choisie comme étant l'un des ingrédients des contraintes de qualité imposées aux données d'enquête par l'utilisateur de ces données. Si ce dernier ne veut pas que le biais de non-réponse ou la racine carrée de l'erreur quadratique moyenne excède une certaine valeur, l'indicateur R doit être supérieur à la borne correspondante.

Il est évident que les bornes inférieures de l'indicateur R dépendent de l'item d'enquête. Par conséquent, nous nous limitons de nouveau à un item hypothétique pour lequel $\hat{S}(y) = 0{,}5$.

Il est facile de montrer, en partant de (17), que si nous demandons que

$$\hat{B}_m(\hat{\rho}) \le \gamma, \tag{19}$$

il doit être vérifié que

$$\hat{R} \ge 1 - 4\hat{\bar{\rho}}\gamma = r_1(\gamma,\ \hat{\rho}). \tag{20}$$

De manière analogue, en utilisant (18) et en demandant que

$$\hat{E}_m(\hat{\rho}) \le \gamma, \tag{21}$$

nous arrivons à

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\sqrt{\gamma^2 - \hat{s}^2(\hat{\bar{y}}_{\mathrm{HT}})} = r_2(\gamma, \hat{\rho}). \qquad (22)$$

Dans (20) et (22), nous représentons par $r_1(\gamma, \hat{\rho})$ et $r_2(\gamma, \hat{\rho})$ les bornes inférieures de l'indicateur R. Dans la section qui suit, nous donnons à $r_1(\gamma, \hat{\rho})$ et $r_2(\gamma, \hat{\rho})$ le nom de fonctions de représentativité de la réponse. Nous les calculons pour les études décrites à la section 5.

4.3.3 Exemple

Nous illustrons de nouveau la normalisation au moyen de l'exemple utilisé aux sections 2.1 et 3.3. La figure 3 contient la fonction de représentativité de la réponse $r_1(\gamma, \hat{\rho})$ et les indicateurs R observés $\hat{R}$ pour les six tentatives de prise de contact durant l'enquête POLS de 1998. Nous avons choisi trois valeurs de γ, soit $\gamma = 0{,}1$; $\gamma = 0{,}075$ et $\gamma = 0{,}05$.

La figure 3 indique qu'après la deuxième tentative de prise de contact, les valeurs de l'indicateur R sont supérieures à la borne inférieure correspondant au niveau de 10 %. Après quatre tentatives, l'indicateur R est proche du niveau de 7,5 %. Cependant, les valeurs ne sont jamais supérieures à l'autre borne inférieure qui est fondée sur le niveau de 5 %.

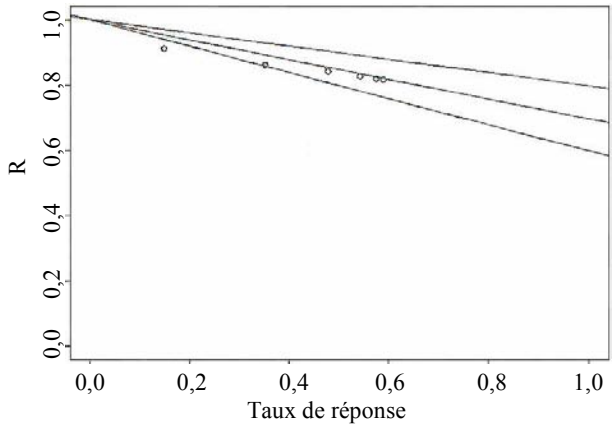


Figure 3 Bornes inférieures de l'indicateur R $\hat{R}$ pour les six premières tentatives de prise de contact durant l'enquête POLS de 1998. Les bornes inférieures sont basées sur $\gamma = 0{,}1$, $\gamma = 0{,}075$ et $\gamma = 0{,}05$

À la figure 4, le biais absolu maximal $\hat{B}_m(\hat{\rho})$ est représenté en fonction du taux de réponse pour les six tentatives de prise de contact. Après la troisième tentative, l'indicateur R a convergé vers une valeur autour de 8 %.

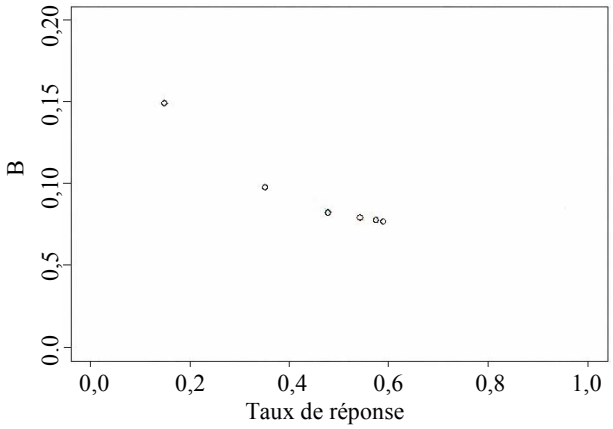


Figure 4 Biais absolu maximal pour les six premières tentatives de prise de contact dans l'enquête POLS de 1998

5. Application de l'indicateur R

À la présente section, nous appliquons l'indicateur R à des études portant sur diverses stratégies de suivi des cas de non-réponse et diverses combinaisons de modes de collecte des données. La première a trait à l'Enquête sur la population active (EPA) des Pays-Bas. Elle comprend un examen de l'approche du rappel (Hansen et Hurwitz 1946) et de celle des questions de bases (Kersten et Bethlehem 1984). La deuxième concerne les plans de collecte de données à mode mixte appliqués dans le cadre de l'enquête nationale de surveillance de la sécurité (*National Safety Monitor*) des Pays-Bas.

Aux sections 5.2 et 5.3, nous examinons ces études de plus près en ce qui a trait à la représentativité de leurs différentes stratégies de travail sur le terrain. Avant cela, à la section 5.1, nous décrivons comment nous obtenons une approximation des erreurs-types.

5.1 Erreur-type et intervalle de confiance

Si nous voulons comparer les valeurs de l'indicateur R pour diverses enquêtes ou stratégies de collecte des données, nous devons estimer leurs erreurs-types.

L'indicateur R $\hat{R}$ fait intervenir l'écart-type d'échantillon des probabilités de réponse estimées. Autrement dit, deux processus aléatoires entrent en jeu. Le premier est l'échantillonnage de la population, et le second, le mécanisme de réponse des unités échantillonnées. Si nous connaissions les probabilités de réponse réelles, le tirage d'un échantillon introduirait quand même une incertitude au sujet de l'indicateur R de population, et donc, entraînerait une certaine perte de précision. Comme nous ne connaissons pas les probabilités de réponse réelles, nous les

estimons en utilisant l'échantillon, ce qui cause une perte de précision supplémentaire.

Un calcul analytique de l'erreur-type de $\hat{R}$ n'est pas simple à cause de l'estimation des probabilités de réponse. Ici, nous nous résignons à utiliser des approximations numériques naïves de l'erreur-type. Nous estimons l'erreur-type de l'indicateur R par une méthode du bootstrap non paramétrique (Efron et Tibshirani 1993). Selon cette méthode, l'erreur-type de l'indicateur R est estimée en sélectionnant un nombre $b = 1, 2, \ldots, B$ d'échantillons bootstrap. Il s'agit d'échantillons tirés indépendamment et avec remise à partir de l'ensemble de données originales et de même taille n que cet ensemble. L'indicateur R est calculé pour chaque échantillon bootstrap b. Nous obtenons donc B répliques de l'indicateur R ; $\hat{R}_b^{\mathrm{BT}}$, $b = 1, 2, \ldots, B$. L'erreur-type pour la loi empirique de ces B répliques est une estimation de l'erreur-type de l'indicateur R, c'est-à-dire

$$s_R^{\mathrm{BT}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{R}_b^{\mathrm{BT}} - \hat{\bar{R}}^{\mathrm{BT}})^2} \qquad (23)$$

où $\hat{\bar{R}}^{\mathrm{BT}} = 1/B \sum_{b=1}^{B} \hat{R}_b^{\mathrm{BT}}$ est l'indicateur R moyen estimé.

Dans les approximations, nous prenons $B = 200$ pour toutes les études. Nous avons expérimenté de plus grandes valeurs de B allant jusqu'à $B = 500$, mais avons constaté dans tous les cas que l'estimation de l'erreur-type avait convergé dès la valeur $B = 200$.

Nous déterminons les intervalles de confiance $100(1 - \alpha)\,\%$ en supposant que la loi de $\hat{R}$ est approximativement normale et en employant les erreurs-types estimées données par (23), soit

$$\mathrm{CI}_\alpha^{\mathrm{BT}} = (\hat{R} \pm \xi_{1-\alpha} \times s_R^{\mathrm{BT}}) \qquad (24)$$

où $\xi_{1-\alpha}$ est le quantile $1 - \alpha$ de la loi normale standard.

5.2 Enquête sur la population active – Étude de suivi de 2005

De juillet à décembre 2005, Statistique Pays-Bas a réalisé un suivi à grande échelle des cas de non-réponse à l'Enquête sur la population active (EPA) des Pays-Bas. L'étude a consisté en une nouvelle tentative d'interview auprès de deux échantillons de non-répondants à l'EPA, selon l'approche du rappel (Hansen et Hurwitz 1946) dans un cas et selon l'approche des questions de base (Kersten et Bethlehem 1984) dans l'autre. Les échantillons étaient formés de ménages sélectionnés pour participer à l'EPA qui avaient refusé de participer, dont les réponses n'avaient pas été traitées ou avec lesquels il n'y avait pas eu prise de contact pour l'EPA de juillet à octobre. Pour concevoir l'étude de suivi, nous avons tenu compte des recommandations émanant des études de Stoop (2005) et de Voogt (2004).

Les principales caractéristiques des approches du rappel et des questions de base appliquées à l'EPA sont présentées au tableau 2. Pour plus de précisions, nous renvoyons le lecteur à Schouten (2007) et à Cobben et Schouten (2007). Dans l'approche du rappel, nous avons employé le questionnaire original visant les ménages administré par IPAO, tandis que dans l'approche des questions de base, nous avons utilisé un questionnaire abrégé dans des conditions de mode de collecte mixte. Le plan à mode mixte comportait une application en ligne, une application papier et une application ITAO. Cette dernière a été utilisée pour tous les ménages possédant un numéro de téléphone publié. Ceux n'ayant pas de numéro de téléphone publié ont reçu une lettre de présentation envoyée à l'avance, un questionnaire imprimé et la procédure d'entrée en communication avec un site Internet sécurisé contenant le questionnaire en ligne. Les répondants étaient libres de remplir le questionnaire imprimé ou le questionnaire en ligne.

Tableau 2
Caractéristiques des deux approches de l'étude de suivi

Approche du rappel	Approche des questions de base
• Questionnaire de l'EPA auquel devaient répondre tous les membres du ménage par IPAO. • 28 intervieweurs sélectionnés géographiquement parmi les intervieweurs qui, selon les antécédents, obtenaient les meilleurs résultats. • Intervieweur différent de celui qui avait reçu la non-réponse. • Formation supplémentaire concernant l'interaction au pas de la porte offerte aux intervieweurs. • Période de travail sur le terrain étendue de deux mois. • Intervieweur autorisé à offrir des mesures d'incitation. • Intervieweur susceptible de recevoir une prime. • Résumé imprimé des caractéristiques du ménage non répondant envoyé à l'intervieweur. • Affectation de l'adresse une semaine après la non-réponse.	• Questionnaire fortement condensé contenant les questions clés de l'EPA et auquel on peut répondre en une à trois minutes. • Plan de collecte des données à mode mixte (Internet, questionnaire papier et ITAO). • Questionnaire rempli par une personne par ménage, selon la méthode du prochain anniversaire. • Ménage contacté une semaine après qu'il ait été traité comme un cas de non-réponse.

La taille de l'échantillon de l'enquête pilote de l'EPA était de $n = 18\,074$ ménages, parmi lesquels 11 275 ont répondu. Les ménages non répondants ont été stratifiés en fonction de la cause de la non-réponse. Étaient admissibles au suivi les ménages qui n'avaient pas été traités, ceux avec

lesquels il n'y avait pas eu prise de contact et ceux qui avaient refusé de participer. Il a été jugé contraire à l'éthique de faire un suivi auprès des ménages qui n'avaient pas répondu pour d'autres raisons, telles qu'une maladie. En tout, 6 171 ménages étaient admissibles. Parmi ceux-ci, nous avons tiré deux échantillons aléatoires simples de taille 775. Dans les analyses, les ménages admissibles non échantillonnés ont été laissés de côté et ceux qui avaient été échantillonnés ont été pondérés en conséquence. Les 11 275 ménages ayant répondu à l'EPA et les 628 ménages inadmissibles ont tous reçu un poids de 1. Cela signifie que les probabilités d'inclusion sont inégales dans notre exemple.

Schouten (2007) a comparé les répondants à l'EPA aux non-répondants convertis et aux non-répondants persistants dans le cas de l'approche du rappel en utilisant un grand ensemble de caractéristiques démographiques et socioéconomiques. Pour prédire le type de réponse, il s'est servi de modèles de régression logistique. Il a conclu que les non-répondants convertis dans le cas de l'approche de rappel diffèrent des répondants à l'EPA en ce qui concerne les variables auxiliaires choisies. En outre, il n'a découvert aucune preuve que les non-répondants convertis différaient des non-répondants persistants en ce qui a trait aux mêmes caractéristiques. Ces constatations l'ont porté à conclure que la réponse combinée de l'EPA et de l'approche du rappel est plus représentative pour ce qui est des variables auxiliaires choisies.

Dans le cas de l'approche des questions de base, la réponse supplémentaire a été analysée par Cobben et Schouten (2007) en se servant du même ensemble de variables auxiliaires et des mêmes modèles de régression logistique. Dans le cas de cette forme de suivi, les constatations n'ont pas été les mêmes pour les ménages possédant et ne possédant pas de numéro de téléphone publié. Pour l'échantillon limité aux ménages ayant un numéro publié, ils ont obtenu les mêmes résultats que pour l'approche du rappel ; la réponse devient plus représentative après l'ajout des répondants ayant un numéro de téléphone publié qui ont répondu aux questions de base. Par contre, pour l'ensemble de la population, c'est-à-dire en incluant les ménages n'ayant pas de numéro de téléphone publié, l'inverse a été observé pour l'approche des questions de base. Celle-ci fournit « un plus grand nombre de résultats pareils » et, donc, accentue le contraste entre les répondants et les non-répondants. La combinaison des réponses à l'EPA et des réponses aux questions de base produit une composition moins représentative. Dans les modèles de régression logistique utilisés par Cobben et Schouten (2007), les indicateurs 0-1 pour la possession d'un numéro de téléphone publié et celle d'un travail rémunéré ont donné une contribution importante.

Cobben et Schouten (2007), ainsi que Schouten (2007) ont utilisé l'ensemble de variables auxiliaires énumérées au tableau 3. Ces variables ont été appariées à l'échantillon provenant de divers registres et données administratives. Elles ont été incluses dans les modèles de régression logistique pour les probabilités de réponse quand elles donnaient une contribution significative au seuil de signification de 5 %. Sinon, elles ont été exclues.

Tableau 3
Variables auxiliaires utilisées dans les études de Schouten (2007) et de Cobben et Schouten (2007). Le noyau du ménage est constitué du chef du ménage et de son (sa) partenaire si présent(e)

Variable
Le ménage possède un numéro de téléphone publié
Région du pays en 4 classes
Province et les 4 villes les plus grandes
Âge moyen en 6 classes
Groupe ethnique en 4 classes
Degré d'urbanisation en 5 classes
Type de ménage en 6 classes
Sexe
Valeur moyenne du logement au niveau du code postal en 11 classes
Au moins un membre du noyau du ménage est un travailleur autonome
Au moins un membre du noyau du ménage est inscrit au centre « Emploi et Revenu » (CWI)
Au moins un membre du noyau du ménage reçoit des allocations sociales
Au moins un membre du noyau du ménage possède un emploi rémunéré
Au moins un membre du noyau du ménage reçoit des allocations d'invalidité

Le tableau 4 donne la taille pondérée de l'échantillon, le taux de réponse, $\hat{R}$, $\mathrm{CI}_{0,05}^{\mathrm{BT}}$, $\hat{B}_m$ et $\hat{E}_m$ pour la réponse à l'EPA, la réponse à l'EPA combinée à la réponse au rappel, et la réponse à l'EPA combinée à la réponse aux questions de base. Les erreurs-types sont relativement grandes en ce qui concerne les études décrites aux sections suivantes, à cause de la pondération. La valeur de $\hat{R}$ augmente quand les répondants au rappel sont ajoutés aux répondants à l'EPA. Comme le taux de réponse et l'indicateur R augmentent tous deux, le biais absolu maximal $\hat{B}_m$ diminue. Les intervalles de confiance $\mathrm{CI}_{0,05}^{\mathrm{BT}}$ pour la réponse à l'EPA et pour les réponses combinées à l'EPA et au rappel se chevauchent. Cependant, l'application d'un test unilatéral mène au rejet de l'hypothèse nulle H_0: $R_{\mathrm{LFS}} - R_{\mathrm{LFS+CB}} \geq 0$ au niveau de signification de 5 %.

Tableau 4
Taille pondérée de l'échantillon, taux de réponse, indicateur R, intervalle de confiance, biais maximal et REQM maximale pour l'EPA, l'EPA plus le rappel, et l'EPA plus les questions de base pour l'ensemble étendu de variables auxiliaires

Réponse	n	Taux	$\hat{R}$	$\mathrm{CI}_{0,05}^{\mathrm{BT}}$	$\hat{B}_m$	$\hat{E}_m$
EPA	18 074	62,2 %	80,1 %	(77,5-82,7)	8,0 %	8,0 %
EPA + rappel	18 074	76,9 %	85,1 %	(82,4-87,8)	4,8 %	4,9 %
EPA + questions de base	18 074	75,6 %	78,0 %	(75,6-80,4)	7,3 %	7,3 %

Dans le tableau 4, nous notons une diminution de $\hat{R}$ quand nous comparons la réponse à l'EPA à la combinaison de cette réponse avec l'approche des questions de base. Cette diminution n'est pas significative. $\hat{B}_m$ diminue légèrement. Au tableau 5, cette comparaison est limitée aux ménages ayant un numéro de téléphone publié. En général, l'indicateur R est beaucoup plus élevé que pour l'ensemble des ménages. Comme la taille de l'échantillon est maintenant plus petite, les erreurs-types estimées sont plus grandes, comme en témoigne la largeur de l'intervalle de confiance. La valeur de $\hat{B}_m$ est plus faible. Pour la réponse combinée à l'EPA et à l'approche des questions de base, nous constatons un accroissement de $\hat{R}$, mais de nouveau, il n'est pas significatif. $\hat{B}_m$ diminue.

Tableau 5
Taille de l'échantillon, taux de réponse, indicateur R, intervalle de confiance, biais maximal et REQM maximale pour l'EPA et pour l'EPA plus les questions de base pour les ménages ayant un numéro de téléphone publié seulement et pour l'ensemble étendu de variables auxiliaires

Réponse	n	Taux	$\hat{R}$	$CI^{BT}_{0,05}$	$\hat{B}_m$	$\hat{E}_m$
EPA	10 135	68,5 %	86,3 %	(83,1-89,5)	5,0 %	5,1 %
EPA + questions de base	10 135	83,0 %	87,5 %	(84,3-90,7)	3,8 %	3,8 %

Dans le cas de l'exemple du suivi de l'EPA, nous constatons que les indicateurs R confirment les conclusions pour l'approche du rappel et celle des questions de base. En outre, l'accroissement de l'indicateur R consécutif à l'ajout des réponses au rappel est significatif au seuil de signification de 5 %.

5.3 Surveillance de la sécurité : essai pilote à mode mixte de 2006

En 2006, Statistique Pays-Bas a réalisé un essai pilote relatif à l'enquête de surveillance de la sécurité (*Safety Monitor*) en vue d'étudier les stratégies de collecte des données à mode mixte. Pour des renseignements détaillés, consulter Cobben, Janssen, Berkel et Brakel (2007). L'enquête de surveillance de la sécurité ordinaire, réalisée auprès de la population des Pays-Bas de 15 ans et plus, porte sur des questions relatives à la sécurité et à la façon dont la police remplit ses fonctions. Il s'agit d'une enquête à mode de collecte mixte. Les personnes dont le numéro de téléphone est publié sont approchées par ITAO. Celles qui ne peuvent pas être rejointes par téléphone sont approchées par IPAO. L'essai pilote de 2006 avait pour but d'évaluer la possibilité d'utiliser Internet comme l'un des modes dans une stratégie de collecte à mode mixte. Les personnes sélectionnées pour y participer ont d'abord été approchées au moyen d'une enquête en ligne. Les non-répondants à l'enquête en ligne ont été approchés de nouveau par ITAO s'ils possédaient un numéro de téléphone publié et par IPAO autrement. Le tableau 6 donne les taux de réponse pour l'enquête ordinaire, la réponse à l'essai pilote pour l'enquête en ligne uniquement et la réponse à l'essai pilote dans son ensemble. La réponse à l'enquête en ligne uniquement est faible. Seulement 30 % de personnes ont rempli le questionnaire en ligne. Cela signifie que près de 70 % des unités échantillonnées ont été réaffectées à l'IPAO ou à l'ITAO. Cela a produit une réponse supplémentaire d'environ 35 %. Le taux global de réponse est légèrement plus faible que celui observé pour l'enquête ordinaire.

Fouwels, Janssen et Wetzels (2006) ont procédé à une analyse univariée des compositions des réponses. Ils soutiennent que le taux de réponse est plus faible pour l'essai pilote, mais que cette diminution est relativement stable chez les divers sous-groupes démographiques. Ils constatent une baisse univariée du taux de réponse pour les variables auxiliaires d'âge, de groupe ethnique, de degré d'urbanisation et de type de ménage. Toutefois, ils relèvent des indices que la réponse devient moins représentative quand la comparaison est limitée aux personnes ayant répondu en ligne seulement. Cette constatation tient surtout pour l'âge des personnes échantillonnées, ce qui n'est pas étonnant.

Le tableau 6 donne la taille de l'échantillon, le taux de réponse, $\hat{R}$, $CI^{BT}_{0,05}$, $\hat{B}_m$ et $\hat{E}_m$ pour trois groupes : l'enquête ordinaire, l'enquête pilote limitée au mode de collecte en ligne et l'enquête pilote dans son ensemble. Les variables auxiliaires d'âge, de groupe ethnique, de degré d'urbanisation et de type de ménage ont été appariées d'après des registres et ont été sélectionnées dans le modèle logistique pour les probabilités de réponse. Le tableau 6 montre que l'indicateur R est plus faible pour la réponse en ligne que pour l'enquête ordinaire. La valeur p correspondante est proche de 5 %. Étant donné à la fois le faible taux de réponse et le faible indicateur R, le biais absolu maximal $\hat{B}_m$ est plus de deux fois plus grand que pour l'enquête ordinaire. Par contre, pour l'essai pilote dans son ensemble, l'indicateur R ainsi que $\hat{B}_m$ sont proches des valeurs observées pour l'enquête ordinaire. À cause de la plus petite taille de l'échantillon de l'essai pilote, les erreurs-types estimées sont plus grandes que pour l'enquête ordinaire.

Tableau 6
Taille de l'échantillon, taux de réponse, indicateur R, intervalle de confiance, biais maximal et REQM maximale pour la réponse à l'enquête de surveillance de la sécurité ordinaire, l'enquête pilote avec composantes en ligne uniquement et l'enquête pilote avec composantes en ligne et suivi par IPAO/ITAO

Réponse	n	Taux	$\hat{R}$	$CI^{BT}_{0,05}$	$\hat{B}_m$	$\hat{E}_m$
Ordinaire	30 139	68,9 %	81,4 %	(80,3-82,4)	6,8 %	6,8 %
Pilote – en ligne	3 615	30,2 %	77,8 %	(75,1-80,5)	18,3 %	18,4 %
Pilote – en ligne plus	3 615	64,7 %	81,2 %	(78,3-84,0)	7,3 %	7,4 %

Les résultats présentés au tableau 6 ne contredisent pas ceux de Fouwels et coll. (2006). Nous constatons aussi que la composition de la réponse en ligne à l'essai pilote est moins équilibrée, tandis que celle de la réponse complète à l'essai pilote n'est pas sensiblement moins bonne que celle de la réponse à l'enquête de surveillance de la sécurité proprement dite.

6. Discussion

Nous poursuivons trois grands objectifs dans le présent article : donner une définition mathématiquement rigoureuse et une perception de la représentativité de la réponse, construire un indicateur possible de la représentativité et illustrer empiriquement l'utilisation de l'indicateur. Comme nous l'avons vu, l'indicateur proposé est un exemple de ce que nous appelons les indicateurs R, où « R » signifie représentativité. Au moyen de l'exemple empirique, nous cherchons à appuyer l'idée que ces indicateurs R sont des outils valables pour comparer différentes enquêtes ou stratégies de collecte des données. Les indicateurs R sont utiles s'ils permettent de confirmer les résultats d'analyses complexes effectuées dans le cadre d'études ayant trait à des enquêtes réalisées sur de multiples périodes ou à de multiples enquêtes portant sur un sujet particulier.

L'indicateur R décrit dans le présent article est prometteur, parce qu'il est facile à calculer et peut être interprété et normalisé quand les propensions à répondre peuvent être estimées sans erreur. L'application à des données d'enquête réelles montre que l'indicateur R confirme les analyses antérieures de la composition de la non-réponse. Naturellement, d'autres indicateurs R peuvent être construits en choisissant simplement différentes fonctions de distance entre les vecteurs des propensions à répondre. L'indicateur R et les graphiques présentés dans l'article peuvent être calculés en utilisant la plupart des progiciels statistiques standard.

Le calcul des indicateurs R est fondé sur un échantillon et s'appuie sur des modèles pour les propensions à répondre individuelles. Donc, ces indicateurs sont eux-mêmes des variables aléatoires et leur estimation comporte deux étapes qui influencent leur biais et leur variance. Cependant, c'est surtout la modélisation des propensions à répondre qui a des conséquences importantes. Le fait de se limiter à l'échantillon pour estimer les indicateurs R implique que ceux-ci sont moins précis, mais asymptotiquement, cette restriction n'introduit pas de biais. Le choix et l'ajustement du modèle sont habituellement effectués en imposant un niveau de signification et en ajoutant uniquement dans le modèle les interactions dont la contribution est significative. Ce dernier point signifie que la taille de l'échantillon et la disponibilité de données sur les variables auxiliaires jouent un rôle important dans l'estimation des propensions à répondre. La stratégie de sélection du modèle peut introduire un biais. Diverses approches évidentes peuvent être utilisées pour traiter la dépendance à l'égard de la taille de l'échantillon. On pourrait ne pas sélectionner un modèle, mais fixer une stratification d'avance. De cette façon, on évite d'introduire un biais, mais les erreurs-types ne sont pas contrôlées et peuvent être considérables. On pourrait aussi prendre la validation empirique comme point de départ pour élaborer les « pratiques exemplaires » pour les indicateurs R.

Nous avons appliqué l'indicateur R proposé à deux études réalisées à Statistique Pays-Bas ces dernières années et dont les résultats avaient été examinés en détail par d'autres auteurs. L'accroissement ou la diminution de la valeur de l'indicateur R concorde avec les résultats des analyses plus détaillées effectuées par ces auteurs. Par conséquent, nous concluons que les indicateurs R peuvent être des outils valables. Cependant, un plus grand nombre de données empiriques sont manifestement nécessaires.

L'application de l'indicateur R a révélé qu'il n'existe aucune relation claire entre le taux de réponse et la représentativité de la réponse. Les taux de réponse plus élevés ne mènent pas nécessairement à une réponse plus équilibrée. Naturellement, nous constatons que les taux de réponse plus élevés réduisent le risque d'un biais de non-réponse. Plus le taux de réponse est élevé, plus le biais absolu maximal observé pour les items d'intérêt est faible.

L'application aux études choisies montre que les erreurs-types diminuent quand la taille de l'échantillon augmente, comme il fallait s'y attendre, mais elles demeurent relativement grandes pour des tailles d'échantillon modestes. Par exemple, pour une taille d'échantillon de 3 600, nous observons une erreur-type d'environ 1,3 %. Donc, si nous émettons l'hypothèse d'une loi normale, l'intervalle de confiance à 95 % a une largeur approximative de 5,4 %. La taille de l'échantillon de l'EPA est d'environ 30 000 unités. L'erreur-type est de l'ordre de 0,5 % et l'intervalle de confiance à 95 % correspondant à une largeur d'environ 2 %. Les erreurs-types sont plus grandes que nous ne l'avions pensé.

Le présent article contient une première étude empirique d'un indicateur R et de son erreur-type. Des travaux de recherche théoriques et empiriques beaucoup plus approfondis seront nécessaires pour comprendre pleinement les indicateurs R et leurs propriétés. En premier lieu, nous n'avons pas du tout tenu compte des items de l'enquête. Manifestement, il est absolument nécessaire que nous le fassions dans l'avenir. Cependant, comme nous l'avons déjà soutenu, les indicateurs R dépendent de l'ensemble de variables auxiliaires. Nous pouvons, par conséquent, conjecturer que, comme dans le cas des méthodes de correction de la non-réponse, la mesure dans laquelle les indicateurs R

prédisent le biais de non-réponse des variables étudiées dépend du mécanisme de création de données manquantes. Si celui-ci est fortement non ignorable, les indicateurs R ne donneront pas de bons résultats. Cependant, sans aucun renseignement sur le mécanisme de données manquantes, aucun autre indicateur ne le ferait non plus. C'est pourquoi nous avons établi la notion de biais absolu maximal, car elle donne une limite au biais de non-réponse sous le pire scénario. Un deuxième sujet de futurs travaux de recherche est le calcul théorique de l'erreur-type de l'indicateur R utilisé dans le présent article. Les erreurs obtenues par le bootstrap non paramétrique ne donnent que des approximations naïves. Or, si nous voulons que les indicateurs R jouent un rôle plus actif dans la comparaison de diverses stratégies, nous devons obtenir des formes (approximativement) explicites. Troisièmement, nous devrions étudier la relation entre le choix et le nombre de variables auxiliaires, d'une part, et les erreurs-types de l'indicateur R, d'autre part.

Remerciements

Les auteurs remercient Bob Groves, Björn Janssen, Geert Loosveldt, ainsi que le rédacteur associé et deux examinateurs de leurs suggestions et commentaires constructifs.

Bibliographie

Agresti, A. (2002). Categorical data analysis. *Wiley Series in Probability and Statistics*. New York : John Wiley & Sons, Inc., NY, USA.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 2, 238-246.

Bertino, S. (2006). A measure of representativeness of a sample for inferential purposes. *Revue Internationale de Statistique*, 74, 149-159.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 3, 251-260.

Cobben, F., Janssen, B., Berkel, K. van et Brakel, J. van den (2007). Statistical inference in a mixed-mode data collection setting. Document présenté au ISI 2007, 23 au 29 août 2007, Lisbon, Portugal.

Cobben, F., et Schouten, B. (2007). An empirical validation of R-indicators. Papier de discussion, CBS, Voorburg.

Curtin, R., Presser, S. et Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-428.

Efron, B., et Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Fouwels, S., Janssen, B. et Wetzels, W. (2006). Experiment mixed-mode waarneming bij de VMR. Rapport technique SOO-2007-H53, CBS, Heerlen.

Goodman, L.A., et Kruskal, W.H. (1979). Measures of association for cross-classifications. Springer-Verlag, Berlijn, Duitsland.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 5, 646-675.

Groves, R.M., et Peytcheva, E. (2006). The impact of nonresponse rates on nonresponse bias: A meta-analysis. Document présenté au 17th International Workshop on Household Survey Nonresponse, 28 au 30 août, Omaha, NE, USA.

Groves, R.M., Presser, S. et Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68, 2-31.

Hájek, J. (1981). Sampling from finite populations. New York : Marcel Dekker, USA.

Hansen, M.H., et Hurwitz, W.H. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Heerwegh, D., Abts, K. et Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 1, 3-10.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 2, 119-127.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. et Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.

Kersten, H.M.P., et Bethlehem, J.G. (1984). Exploring an reducing the nonresponse bias by asking the basic question. *Statistical Journal of the United Nations*, ECE 2, 369-380.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.

Kruskal, W., et Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *Revue Internationale de Statistique*, 47, 13-24.

Kruskal, W., et Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *Revue Internationale de Statistique*, 47, 111-123.

Kruskal, W., et Mosteller, F. (1979c). Representative sampling III: Current statistical literature. *Revue Internationale de Statistique*, 47, 245-265.

Little, R.J.A., et Rubin, D.B. (2002). Statistical analysis with missing data. Wiley Series in Probability and Statistics. New York : John Wiley & Sons, Inc., NY, USA.

Marsh, H.W., Balla, J.R. et McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 3, 391-410.

Merkle, D.M., et Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. Dans *Survey Nonresponse* (Éds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A Little). New York : John Wiley & Sons, Inc., 243-258.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Särndal, C., et Lundström, S. (2005). Estimation in Surveys with Nonresponse. Wiley Series in Survey Methodology, John Wiley & Sons, Chichester, England.

Särndal, C., Swensson, B. et Wretman, J. (2003). Model-assisted survey sampling. Springer Series in Statistics, Springer, New York.

Schouten, B. (2004). Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998. Papier de discussion 04001, CBS, Voorburg, disponible au site Web http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2004/default.htm.

Schouten, B., et Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes, Article de discussion 07002, CBS, Voorburg. Disponible au http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2007/default.htm.

Stoop, I. (2005). Surveying nonrespondents. *Field Methods*, 16, 23-54.

Voogt, R. (2004). I am not interested: Nonresponse bias, response bias and stimulus effects in election research. Thèse de doctorat, University of Amsterdam, Amsterdam.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 123-127
Statistique Canada, N° 12-001-X au catalogue

Échantillonnage équilibré stratifié

Guillaume Chauvet [1]

Résumé

Lors de la sélection d'un échantillon, une pratique courante consiste à définir un plan de sondage stratifié sur des sous-populations. La variance de l'estimateur de Horvitz-Thompson est alors réduite par rapport à un tirage direct si les strates sont bien homogènes au regard de la variable d'intérêt. Si des variables auxiliaires sont disponibles pour chaque individu, l'échantillonnage peut être amélioré par tirage équilibré au sein de chaque strate et l'estimateur de Horvitz-Thompson sera plus précis si les variables auxiliaires sont bien corrélées à la variable d'intérêt. Cependant, si la répartition d'échantillon est faible dans certaines strates, l'équilibrage ne sera respecté que de façon très approximative. Nous proposons ici une méthode de tirage permettant de sélectionner un échantillon équilibré sur l'ensemble de la population, en respectant une allocation fixée au sein de chaque strate. Nous montrons que dans le cas particulier important d'un tirage de taille 2 dans chaque strate, la précision de l'estimateur de Horvitz-Thompson est améliorée si la variable d'intérêt est bien expliquée par les variables d'équilibrage sur l'ensemble de la population. Une application au cas d'un échantillonnage rotatif est également proposée.

Mots clés : Échantillonnage rotatif ; entropie maximale ; méthode du Cube ; stratification ; tirage à probabilités inégales.

1. Introduction

Dans le cas d'un tirage stratifié, la population U est partionnée en H sous-populations U_h, $h = 1, ..., H$ appelées strates, dans lesquelles des échantillons S_h, $h = 1, ..., H$ sont respectivement sélectionnés selon des plans de sondage indépendants p_h, $h = 1, ..., H$. La probabilité d'inclusion de l'unité k est la probabilité π_k que l'unité k soit sélectionnée dans l'échantillon, et la probabilité d'inclusion jointe est la probabilité π_{kl} que deux unités distinctes k et l soient sélectionnées conjointement dans l'échantillon. On notera $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ et $\boldsymbol{\pi}^{\mathbf{h}} = (\pi_k)_{k\in U_h}$. On suppose qu'au sein de chaque strate U_h, le plan $p_h(.)$ est de taille fixe. On a alors en particulier $\sum_{k\in U_h}\pi_k = n_h$, $h = 1, ..., H$ où n_h désigne la répartition dans la strate U_h. Nous supposerons dans la suite de l'article que toutes les tailles d'échantillon par strate n_h sont entières.

L'estimateur de Horvitz-Thompson $\hat{t}_{\mathbf{z}\pi} = \sum_{k\in S}\mathbf{z}_k/\pi_k = \sum_{h=1}^{H}\hat{t}_{\mathbf{z}\pi}^h$, où $\hat{t}_{\mathbf{z}\pi}^h = \sum_{k\in S_h}\mathbf{z}_k/\pi_k$, estime sans biais $t_{\mathbf{z}} = \sum_{k=1}^{H}t_{\mathbf{z}}^h$, où $t_{\mathbf{z}}^h = \sum_{k\in U_h}\mathbf{z}_k$ désigne le total sur U_h de la variable (vectorielle) $\mathbf{z}$. Dans le cas particulier où $\mathbf{z}_k = y_k$ est scalaire, la variance de l'estimateur de Horvitz-Thompson s'obtient à l'aide de la formule de variance de Sen-Yates-Grundy :

$$\mathrm{Var}(\hat{t}_{y\pi}) = \frac{1}{2}\sum_{h=1}^{H}\sum_{k\neq l\in U_h}(\pi_k\pi_l - \pi_{kl})\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \qquad (1)$$

Cette variance sera faible si les strates sont homogènes au regard de la variable d'intérêt, plus exactement si y_k/π_k est approximativement constant au sein de chaque strate.

Si un vecteur $\mathbf{x} = (x_1, ..., x_q)$ de q variables auxiliaires est disponible avant tirage pour chaque individu de la population, l'échantillonnage au sein de chaque strate peut être amélioré à l'aide de l'algorithme du Cube (Deville et Tillé 2004) qui permet de sélectionner des échantillons équilibrés. Le plan de sondage $p_h(.)$ est dit équilibré sur les variables $\mathbf{x}$ si les équations

$$\hat{t}_{\mathbf{x}\pi}^h = t_{\mathbf{x}}^h \qquad (2)$$

sont exactement respectées. La variance de l'estimateur de Horvitz-Thompson est donc nulle pour l'estimation du total des variables d'équilibrage. Dans le cas particulier où $\mathbf{x} = \boldsymbol{\pi}$, c'est-à-dire si la probabilité d'inclusion est la seule variable d'équilibrage, les équations (2) se réduisent à

$$\sum_{k\in S_h}1 = \sum_{k\in U_h}\pi_k = n_h. \qquad (3)$$

Le tirage stratifié de taille fixe dans chaque strate est donc un cas particulier d'échantillonnage équilibré. Dans le cas d'un nombre quelconque de contraintes, un échantillon exactement équilibré ne peut généralement être trouvé. Supposons par exemple que la population U_h contienne 100 individus sur lesquels est définie une variable x à deux modalités, 0 et 1, telle que 53 individus de la population présentent la modalité 0. Sélectionner un échantillon de taille 10, à probabilités égales, équilibré sur la variable x, supposerait de sélectionner un échantillon contenant 5.3 individus présentant la modalité $x = 0$ et 4.7 individus présentant la modalité $x = 1$, ce qui est impossible. L'objectif est donc généralement de sélectionner un échantillon approximativement équilibré, c'est-à-dire tel que

1. Guillaume Chauvet, Laboratoire de Statistique d'Enquête, CREST/ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35 170 Bruz, France. Courriel : chauvet@ensai.fr.

$$\hat{t}^h_{\mathbf{x}\pi} \simeq t^h_{\mathbf{x}}. \quad (4)$$

La méthode du Cube (Deville et Tillé 2004) permet de sélectionner des échantillons approximativement équilibrés sur un nombre quelconque de variables, en respectant exactement un jeu de probabilités d'inclusion $\boldsymbol{\pi}$ préalablement choisi. Elle se décompose en deux phases appelées phase de vol et phase d'atterrissage. A chaque étape de la phase de vol, on décide aléatoirement de sélectionner ou d'écarter définitivement l'une des unités de la population. A l'issue de la phase de vol, on obtient dans chaque strate U_h un vecteur $\boldsymbol{\pi}^{\mathbf{h}*} = (\pi^*_k)_{k\in U_h} \in [0, 1]^N$ vérifiant les conditions suivantes :

$$E(\boldsymbol{\pi}^{\mathbf{h}*}) = \boldsymbol{\pi}^{\mathbf{h}}, \quad (5)$$

$$\sum_{k\in U_h} \frac{\mathbf{x}_k}{\pi_k}\pi^*_k = \sum_{k\in U_h} \mathbf{x}_k, \quad (6)$$

$$\text{Card}\{k \in U_h ; 0 < \pi^*_k < 1\} \leq q, \quad (7)$$

où E désigne l'espérance sous le mécanisme d'échantillonnage associé à la phase de vol. Le vecteur $\boldsymbol{\pi}^{\mathbf{h}*}$ donne le résultat de la phase de vol : π^*_k vaut 1 si l'unité k est sélectionnée, 0 si elle est rejetée et est comprise entre 0 et 1 strictement si la décision n'est pas encore prise pour l'unité k après la phase de vol. Les équations (5) et (6) assurent que les probabilités d'inclusion et les contraintes d'équilibrage sont parfaitement respectées à l'issue de la phase de vol. L'équation (7) assure qu'il reste à trancher pour au plus q individus dans chaque strate U_h, où q désigne le nombre de variables d'équilibrage. La phase de vol s'arrête quand les contraintes d'équilibrage ne peuvent plus être exactement respectées. La phase d'atterrissage consiste alors à définir, conditionnellement au résultat de la phase de vol, un plan de sondage optimal défini sur la population V des unités restantes. Ce plan est optimal au sens où il permet de terminer l'échantillonnage en minimisant la variance conditionnelle à la phase de vol de l'estimateur de Horvitz-Thompson des variables d'équilibrage. Les unités restantes sont échantillonnées, conditionnellement au résultat de la phase de vol, avec les probabilités d'inclusion $(\pi^*_k)_{k\in V}$, de sorte que les probabilités d'inclusion non conditionnelles $(\pi_k)_{k\in V}$ de ces unités soient exactement respectées.

La mesure d'entropie associée à un plan de sondage $p(.)$ défini sur la population U est donnée par

$$I(p) = -\sum_{s\subset U} p(s)\log(p(s)),$$

avec la convention $0\log(0) = 0$. Deville et Tillé (2005) montrent que le plan équilibré à entropie maximale parmi les plans de sondage équilibrés sur les mêmes variables et respectant les mêmes probabilités d'inclusion peut être vu comme le conditionnel d'un plan de Poisson. Sous une hypothèse de normalité asymptotique de l'estimateur de Horvitz-Thompson multivarié dans le cas d'un plan poissonien, ils en déduisent une formule approchée de variance pour l'estimateur de Horvitz-Thompson pour un plan de sondage équilibré. Dans le cas de l'échantillonnage équilibré stratifié, on obtient

$$\text{Var}(\hat{t}_{y\pi}) \simeq \sum_{h=1}^{H}\sum_{k\in U_h} \frac{b_k}{\pi^2_k}(y_k - \beta_{\mathbf{h}}\mathbf{x}_k)^2 \quad (8)$$

où $\beta_{\mathbf{h}} = (\sum_{l\in U_h} b_l\ \mathbf{x}_l/\pi_l\ \mathbf{x}'_l/\pi_l)^{-1}\sum_{l\in U_h} b_l\ \mathbf{x}_l/\pi_l\ y_l/\pi_l$. Deville et Tillé (2005) proposent plusieurs approximations pour les coefficients b_k. La plus simple consiste à utiliser $b_k = \pi_k(1 - \pi_k)$. La variance de l'estimateur de Horvitz-Thompson sera faible si dans chaque strate la variable d'intérêt y est bien expliquée par les variables d'équilibrage $\mathbf{x}$.

L'équilibrage sera bien respecté au sein de chaque strate si le nombre de variables d'équilibrage reste faible devant la taille d'échantillon. Mais dans certains cas, la répartition par strate est trop faible pour permettre l'équilibrage : si la population est stratifiée de façon très fine, une pratique courante consiste à sélectionner un échantillon de taille 2 dans chaque strate. Il n'est alors pas possible d'imposer une condition autre que la contrainte de taille fixe d'échantillon dans chaque strate.

Nous proposons dans la section suivante un algorithme d'échantillonnage adapté de la méthode du Cube, assurant un équilibrage sur l'ensemble de la population pour des variables choisies et permettant le respect strict de la répartition souhaitée au sein de chaque strate. Les échantillons ne sont alors plus sélectionnés indépendamment dans chaque strate. La précision est améliorée par rapport à un sondage stratifié avec tirage de taille fixe dans chaque strate si sur l'ensemble de la population les variables d'équilibrage sont bien corrélées à la variable d'intérêt. Cet algorithme présente également l'avantage de permettre un équilibrage approximatif dans chaque strate, qui sera d'autant mieux respecté que la taille d'échantillon qui y est allouée est importante.

2. Échantillonnage équilibré stratifié avec mise en commun des phases d'atterrissage

Si l'échantillon S est sélectionné dans U selon la procédure d'échantillonnage équilibré stratifié présentée en section 1, l'équilibrage sera bien respecté au sein de chaque strate si la phase d'atterrissage porte sur un faible nombre d'individus par rapport à la taille d'échantillon. Plus précisément, l'équation (7) montre que le nombre de variables d'équilibrage doit être faible par rapport à la répartition

d'échantillon dans chaque strate. Dans certains cas, cette contrainte ne peut être respectée. Il est en effet fréquent d'utiliser un découpage très fin de la population afin d'améliorer sa pertinence, ce qui revient à réduire la taille d'échantillon sélectionnée dans chaque strate en se fixant généralement la limite d'un échantillon de taille 2 pour chacune afin de pouvoir obtenir un estimateur sans biais de variance.

Nous nous plaçons à nouveau dans le cas d'une population U découpée en H strates U_1, ..., U_H, et dans laquelle un vecteur $\mathbf{x}_k = (\pi_k, \mathbf{z}_k')'$ de variables auxiliaires est connu. On suppose que la variable π_k fait partie des contraintes d'équilibrage, afin d'assurer un échantillonnage de taille fixe. Dans le cas où la répartition par strate est trop faible pour que l'équilibrage permette d'imposer d'autres contraintes que celle de taille fixe dans chaque strate, l'algorithme 1 fournit une méthode alternative d'échantillonnage. Une phase de vol est réalisée indépendamment sur chacune des H strates : on note $\boldsymbol{\pi}^{\mathbf{h}*} = (\pi_k^*)_{k\in U_h}$, $h = 1, \ldots H$ les vecteurs de probabilités obtenus à l'issue de ces phases de vol, $\boldsymbol{\pi}^* = (\pi_k^*)_{k\in V}$, où V désigne les unités qui n'ont pas encore été échantillonnées ou rejetées, et $\mathbf{x}_k^* = (\pi_k^* 1_{k\in U_1}, \ldots, \pi_k^* 1_{k\in U_H}, \mathbf{z}_k' \ \pi_k^*/\pi_k)'$. Le vecteur des probabilités obtenues après une dernière phase de vol sur l'ensemble de ces unités restantes est noté $\boldsymbol{\pi}^{**} = (\pi_k^{**})_{k\in V}$. L'ensemble des unités de la strate U_h qui n'ont pas encore été échantillonnées ou rejetées à l'issue de cette nouvelle phase de vol est noté W_h.

Algorithme 1 : Échantillonnage équilibré stratifié avec mise en commun des phases d'atterrissage

Étape 1. Réaliser une phase de vol, avec les variables d'équilibrage $\mathbf{x}_k$ et les probabilités d'inclusion π_k, indépendamment dans chaque strate U_h.

Étape 2. Réaliser une phase de vol, avec les variables d'équilibrage $\mathbf{x}_k^*$ et les probabilités d'inclusion π_k^*, sur l'ensemble V des unités restantes à l'issue de l'étape 1.

Étape 3. Sélectionner un échantillon de taille fixe dans chaque sous-population W_h, avec des probabilités d'inclusion π_k^{**}.

L'algorithme s'inspire d'une méthode utilisée à l'Institut National de la Statistique et des Études Économiques (Insee) pour la sélection des unités primaires de l'Échantillon-maître 1999. L'Échantillon-Maître est un échantillon de logements, sélectionné dans le Recensement de 1999, et servant de base de sondage pour les enquêtes sur les ménages. On trouvera une description détaillée du plan de sondage de l'Échantillon-Maître dans Bourdalle, Christine et Wilms (2000). Les logements sont à l'origine regroupés au sein d'unités urbaines ou d'unités rurales. Dans la sous-population des unités de moins de 100 000 habitants, un échantillon d'environ 6 % est sélectionné. On dispose de quatre variables auxiliaires (revenu net imposable, et effectif selon trois tranches d'âge). Le nombre attendu d'unités échantillonnées est trop faible pour permettre un tirage stratifié selon la région, avec équilibrage selon ces quatre variables dans chaque région. Les régions ont donc été regroupées en 8 super-régions, et les tirages sont coordonnés de manière à assurer, d'une part un équilibrage global pour les quatre variables auxiliaires sur chaque super-région, et d'autre part une taille fixe d'échantillon dans chaque région.

Une méthode similaire est également proposée par Rousseau et Tardieu (2004) pour la sélection d'échantillons équilibrés dans de grandes bases de sondage en utilisant la macro CUBE disponible sur le site Internet de l'Insee. Le temps d'exécution de cette macro est en effet approximativement proportionnel au carré de la taille de la population. Notons que Chauvet et Tillé (2006) proposent une méthode rapide d'échantillonnage équilibré dont le temps de calcul ne dépend plus que de la taille de la population, et permet de sélectionner directement des échantillons équilibrés sur de très grandes populations. L'algorithme a été programmé sous forme d'une macro SAS, voir Chauvet et Tillé (2005), et il est également disponible dans le R Sampling Package de Matei et Tillé (2006). Dans chacun des deux programmes, la seconde phase de vol est réalisée en ajoutant aux variables d'équilibrage $\mathbf{x}_k^*$ une contrainte associée à chaque strate, et permettant de maintenir la condition de taille fixe dans chacune d'entre elles.

Utiliser le vecteur de probabilités d'inclusion $\boldsymbol{\pi}^*$ conditionnellement au résultat obtenu à l'issue de l'étape 1, assure que le vecteur $\boldsymbol{\pi}$ des probabilités d'inclusion est respecté en déconditionnant par rapport au résultat de l'étape 1. A l'issue de cette 1ère étape, l'équation (6) implique que

$$\forall h = 1 \ldots H \quad \sum_{k\in U_h/0<\pi_k^*<1} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U_h} \mathbf{x}_k - \sum_{k\in U_h/\pi_k^*=1} \frac{\mathbf{x}_k}{\pi_k}$$

et en sommant ces expressions

$$\sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U} \mathbf{x}_k - \sum_{k\in U/\pi_k^*=1} \frac{\mathbf{x}_k}{\pi_k}.$$

A l'issue de l'étape 2, on obtient à l'aide de l'équation (6)

$$\sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} = \sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^*$$

et en comparant ces deux dernières expressions

$$\sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} + \sum_{k\in U/\pi_k^*=1} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U} \mathbf{x}_k \qquad (9)$$

ce qui assure que l'équilibrage sur les variables $\mathbf{x}_k$ est exactement respecté à l'issue de l'étape 2. L'étape 3 permet

de terminer l'échantillonnage en respectant la contrainte de taille fixe à l'intérieur de chaque strate U_h, et peut être réalisé au moyen d'un programme linéaire afin de limiter le défaut d'équilibrage (*cf.* Deville et Tillé 2004).

La variance peut être approchée à l'aide de la formule de variance proposée par Deville et Tillé (2005), si chaque phase de vol de l'algorithme 1 est réalisée avec une entropie forte. L'entropie peut être fortement augmentée en triant aléatoirement la population concernée préalablement au tirage. Les variables d'équilibrage sont ici d'une part les variables $\mathbf{z}_k$, et d'autre part les variables données par le produit des probabilités d'inclusion et des indicatrices d'appartenance aux strates et assurant une taille fixe d'échantillon dans chaque strate. On a

$$\mathrm{Var}(\hat{t}_{y\pi}) \simeq \sum_{k\in U} \frac{b_k}{\pi_k^2} (y_k - \gamma' \mathbf{a}_k)^2 \qquad (10)$$

avec $\mathbf{a}_k = (\pi_k \, 1_{k\in U_1}, \ldots, \pi_k \, 1_{k\in U_H}, \mathbf{z}_k')'$ et

$$\gamma = \left(\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

On peut utiliser l'estimateur de variance

$$v(\hat{t}_{y\pi}) = \sum_{k\in S} \frac{b_k}{\pi_k^3} (y_k - \hat{\gamma}' \mathbf{a}_k)^2 \qquad (11)$$

proposé par Deville et Tillé (2005), page 578, avec

$$\hat{\gamma} = \left(\sum_{l\in S} \frac{b_l}{\pi_l} \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l\in S} \frac{b_l}{\pi_l} \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

Ainsi qu'il apparaît dans la formule (10) de variance approchée, il est important de noter que l'indépendance des tirages dans les différentes strates est perdue avec la méthode d'échantillonnage équilibrée stratifiée proposée. Les tirages dans les strates $U_1, \ldots, U_H$ sont en effet coordonnés de façon à assurer un équilibrage global sur l'ensemble de la population, ce qui supprime l'indépendance. L'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}^h$ du total t_{yh} reste sans biais. Sa variance approchée se déduit de la formule (10) en remplaçant y_k par $y_k \, 1_{k\in U_h}$, et est donnée par

$$\mathrm{Var}(\hat{t}_{y\pi}^h) \approx \sum_{k\in U} \frac{b_k}{\pi_k^2} (y_k \, 1_{k\in U_h} - (\gamma^h)' \mathbf{a}_k)^2 \qquad (12)$$

avec $\mathbf{a}_k = (\pi_k \, 1_{k\in U_1}, \ldots, \pi_k \, 1_{k\in U_H}, \mathbf{z}_k')'$ et

$$\gamma^h = \left(\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l\in U_h} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

Dans le cas particulier où l'inférence ne porte pas sur la population entière mais sur un domaine D comportant un nombre limité de strates, le bénéfice de l'équilibrage global sur les variables $\mathbf{z}$ sera faible. La variance de l'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}^D$ du total t_y^D de la variable y sur ce domaine sera proche de la variance obtenue dans le cas d'un échantillonnage stratifié et donnée par la formule (1).

3. Résultats numériques

Nous réalisons une courte étude par simulations pour tester les performances de notre algorithme d'échantillonnage. Nous générons tout d'abord une population finie de taille 1 000, partitionnée en 25 strates de même taille, et contenant 4 variables : 2 variables d'intérêt y_1 et y_2, et 2 variables auxiliaires x_1 et x_2. Tout d'abord, les variables x_1 et x_2 sont générées selon une distribution Gamma de paramètres 4 et 25. La variable y_1 est générée au sein de la strate U_h selon le modèle

$$y_1 = \alpha_{1h} + \varepsilon_h. \qquad (13)$$

Les ε_h sont générés selon une distribution normale de moyenne 0 et de variance σ_h^2. Le modèle utilisé pour générer les valeurs de y_1 est donné par (13), avec $\alpha_{1h} = 20\,h$ et une variance σ_h^2 choisie pour donner un coefficient de détermination R^2 approximativement égal à 0,60 au sein de chaque strate. La variable y_2 est générée selon le modèle

$$y_2 = \alpha_2 + \beta_2 \, x_1 + \gamma_2 \, x_2 + \eta. \qquad (14)$$

Les η sont générés selon une distribution normale de moyenne 0 et de variance ρ^2. Le modèle utilisé pour générer les valeurs de y_2 est donné par (14), avec $\alpha_2 = 500$, $\beta_2 = \gamma_2 = 5$, et une variance ρ^2 choisie pour donner un coefficient de détermination R^2 approximativement égal à 0,60.

On s'intéresse à l'estimation du total des variables y_1 et y_2. On sélectionne un échantillon de $n = 25$ (respectivement $n = 50$) unités à probabilités égales selon trois plans de sondage :

Plan 1 : tirage stratifié avec sondage aléatoire simple dans chaque strate,
Plan 2 : tirage équilibré sur les variables π, x_1 et x_2,
Plan 3 : tirage équilibré sur les variables π, x_1 et x_2, stratifié avec mise en commun des phases d'atterrissage.

Dans le cas du tirage stratifié, on a donc une allocation de taille 1 (respectivement 2) dans chaque strate. Dans les tirages équilibrés, chaque phase de vol est précédée d'un tri aléatoire de la population concernée. La variance associée au plan 1 est calculée directement. La variance associée aux plans 2 et 3 est approchée sur la base de 10 000 simulations. Le tableau 1 compare les résultats obtenus.

Tableau 1
Variance associée à l'estimation du total de 2 variables pour les plans de sondage stratifié, équilibré, et stratifié équilibré avec mise en commun des phases d'atterrissage

	$n = 25$		$n = 50$	
	Var. total	Var. total	Var. total	Var. total
Méthode	y_1 ($\times 10^8$)	y_2 ($\times 10^9$)	y_1 ($\times 10^8$)	y_2 ($\times 10^9$)
Plan 1	6,05	7,13	2,95	3,48
Plan 2	14,31	3,05	7,02	1,40
Plan 3	6,00	3,63	2,98	1,54

Dans chaque cas, le plan de sondage proposé est comparable avec la meilleure des deux stratégies. Si la variable d'intérêt est approximativement constante par strate, l'algorithme proposé donne les mêmes résultats que le plan de sondage stratifié. Si les variables d'équilibrage sont bien explicatives, les résultats obtenus avec notre algorithme et avec un tirage équilibré direct sont équivalents. La légère perte de précision provient de l'étape d'atterrissage : dans le cas du tirage équilibré direct, on cherche à terminer l'échantillonnage en limitant le défaut d'équilibrage. Avec l'algorithme proposé, la solution retenue est sous-optimale car on ajoute la contrainte supplémentaire d'une taille fixe dans chaque strate.

Dans le cas du tirage stratifié équilibré avec mise en commun des phases d'atterrissage, le tableau 2 compare la variance donnée par 10 000 simulations avec la formule de variance approchée donnée en (10).

Tableau 2
Comparaison entre la variance donnée par 10 000 simulations et la formule de variance approchée dans le cas de l'estimation de deux totaux pour un plan de sondage stratifié équilibré avec mise en commun des phases d'atterrissage

	$n = 25$		$n = 50$	
	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)
Var. simulations	6,0	3,6	3,0	1,5
Var. approchée	5,9	2,7	2,9	1,3

La formule approchée proposée par Deville et Tillé (2005) est proche de la précision exacte si la variance associée à la phase d'atterrissage est faible devant la variance associée à la phase de vol. Dans le cas de la variable y_2, les variables d'équilibrage sont fortement explicatives. La variance liée à la phase d'atterrissage est alors plus importante, relativement à la phase de vol, et la formule approchée sous-estime la vraie variance. La prise en compte de la variance associée à la phase d'atterrissage fera l'objet de travaux ultérieurs.

Remerciements

L'auteur remercie les examinateurs et un éditeur associé pour leurs commentaires et suggestions constructifs.

Bibliographie

Bourdalle, G., Christine, M. et Wilms, L. (2000). Échantillons maître et emploi. Série INSEE Méthodes, Paris, France, 21, 139-173.

Chauvet, G., et Tillé, Y. (2005). New SAS macros for balanced sampling. INSEE, Journées de Méthodologie Statistique, Paris.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Matei, A., et Tillé, Y. (2006). The R 'sampling' package. *European Conference on Quality in Survey Statistics*, Cardiff.

Rousseau, S., et Tardieu, F. (2004). *La macro SAS CUBE d'échantillonnage équilibré - Documentation de l'utilisateur*. Rapport technique, INSEE, France.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

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 24, No. 4, 2008

Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities
Joseph Sedransk 495

Discussion
Nathaniel Schenker, Trivellore E. Raghunathan 507

Discussion
David A. Binder 513

Model Averaging Methods for Weight Trimming
Michael R. Elliott 517

A Note on the Asymptotic Equivalence of Jackknife and Linearization Variance Estimation for the Gini Coefficient
Yves G. Berger 541

On Some Common Practices of Systematic Sampling
Li-Chun Zhang 557

Does a Final Coverage Check Identify and Reduce Census Coverage Errors?
Elizabeth Martin, Don A. Dillman 571

Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse
Nancy Bates, James Dahlhamer, Eleanor Singer 591

Seasonality in Investment Plans and Their Revisions
Alex Teterukovsky 612

Editorial Collaborators 623

Contents
Volume 25, No. 1, 2009

Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models
Michael R. Elliott 1

The Use of Sample Weights in Hot Deck Imputation
Rebecca R. Andridge, Roderick J. Little 21

Small Area Estimation in the Presence of Correlated Random Area Effects
Monica Pratesi, Nicola Salvati 37

Nonparametric Variance Estimation for Nearest Neighbor Imputation
Jun Shao 55

On the Inter-Regional Mover Problem in Panel Household Surveys
Takis Merkouris 63

Compensating for Noncoverage of Nontelephone Households in Random-Digit-Dialing Surveys:
A Comparison of Adjustments Based on Propensity Scores and Interruptions in Telephone Service
K.P. Srinath, Martin R. Frankel, David C. Hoaglin, Michael P. Battaglia 77

Population Shifts and Demographic Methods
C. Matthew Snipp, Juanita Tamayo Lott 99

Privacy-Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products
Alan F. Karr, Xiaodong Lin, Ashish P. Sanil, Jerome P. Reiter 125

Using a Weighted Average of Base Period Price Indexes to Approximate a Superlative Index
Janice Lent, Alan H. Dorfman 139

Book and Software Reviews
Jana Asher, Trent D. Buskirk, John Hall, Gordon Willis 151

In Other Journals 165

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 36, No. 3, September/septembre 2008

Christopher A. FIELD
Modeling biological data: several vignettes 341

Marco A.R. FERREIRA & Marc A. SUCHARD
Bayesian analysis of elapsed times in continuous-time Markov chains 355

Amélie FILS-VILLETARD, Armelle GUILLOU & Johan SEGERS
Projection estimators of Pickands dependence functions 369

Simon GUILLOTTE & François PERRON
A Bayesian estimator for the dependence function of a bivariate extreme-value distribution....383

Jesúus LÓPEZ-FIDALGO, Raul MARTÍN-MARTÍN & Douglas P.WIENS
Marginally restricted sequential D-optimal designs 397

Kerrie P. NELSON & Don EDWARDS
On population-based measures of agreement for binary classifications 411

Lei NIE, Haitao CHU & Valeriy R. KOROSTYSHEVSKIY
Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits 427

Jean-François PLANTE
Nonparametric adaptive likelihood weights 443

Yichuan ZHAO & Hongkun WANG
Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data....463

Volume 36, No. 4, December/décembre 2008

Mylène BÉDARD & Jeffrey S. ROSENTHAL
Optimal scaling of Metropolis algorithms: heading toward general target distributions 483

Jörn DANNEMANN & Hajo HOLZMANN
Testing for two states in a hidden Markov model 505

Christopher A. FIELD, Zhen PANG & Alan H. WELSH
Bootstrapping data with multiple levels of variation 521

Feifang HU, Li-Xin ZHANG, Siu H. CHEUNG & Wai S. CHAN
Doubly adaptive biased coin designs with delayed responses 541

Sai Man Simon KWOK & Wai Keung LI
On diagnostic checking of the autoregressive conditional intensity model 561

Omer OZTURK
Inference in the presence of ranking error in ranked set sampling 577

Matías SALIBIÁN-BARRERA & Ying WEI
Weighted quantile regression with nonelliptically structured covariates 595

Ana-Maria STAICU & Nancy M. REID
On probability matching priors 613

Robin WILLINK
Shrinkage confidence intervals for the normal mean: using a guess for greater efficiency 623