Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Un plan de sondage assisté par un modèle est le minimax pour la prédiction fondée sur un modèle

par Robert Graham Clark

Date de diffusion : le 30 juin 2020

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Techniques d'enquête, juin 2020
Vol. 46, nº1, pp. 81-97
Statistique Canada, nº 12-001-X au catalogue

Un plan de sondage assisté par un modèle est le minimax pour la prédiction fondée sur un modèle

Robert Graham Clark[1]

Résumé

Les plans de sondage probabilistes sont parfois utilisés en conjonction avec des prédicteurs fondés sur un modèle de quantités de population finie. Ces plans devraient réduire au minimum la variance anticipée (VA), qui est la variance du prédicteur d'intérêt sur la superpopulation et les processus d'échantillonnage. Le plan optimal pour la VA est bien connu pour les estimateurs assistés par un modèle qui atteignent la borne inférieure de Godambe et Joshi pour la VA des estimateurs sans biais sous le plan de sondage. Cependant, aucun plan de sondage probabiliste optimal n'a été trouvé pour la prédiction fondée sur un modèle, sauf dans des conditions telles que les estimateurs fondés sur un modèle et assistés par un modèle coïncident. Ces cas peuvent être limitatifs. Le présent article montre que la borne inférieure de Godambe et Joshi est une borne *supérieure* pour la VA du meilleur estimateur linéaire sans biais d'un total de population, où la limite supérieure est dans l'espace de tous les ensembles de covariables. C'est pourquoi les plans optimaux assistés par un modèle constituent un choix raisonnable pour la prédiction fondée sur un modèle en cas d'incertitude sur la forme du modèle final, comme cela se produit souvent avant la réalisation de l'enquête. Les simulations confirment le résultat dans différentes situations, y compris quand la relation entre les variables cibles et auxiliaires est non linéaire et modélisée au moyen de splines. La VA est la plus basse par rapport à la borne quand une variable importante du plan de sondage n'est pas associée à la variable cible.

Mots-clés : Variance anticipée; inférence fondée sur un modèle; échantillonnage probabiliste; enquêtes-échantillons.

1 Introduction

L'inférence fondée sur un modèle pour des totaux de population finie repose sur un modèle hypothétique et ne fait généralement pas référence au plan d'échantillonnage. L'échantillonnage probabiliste, où toute unité i a une probabilité de sélection connue $\pi_i > 0$, n'est pas strictement nécessaire, mais il est tout de même souvent utilisé, parce qu'il « supprime le biais conscient et inconscient » (Valliant, Dever et Kreuter, 2013, page 310) et qu'il garantit le caractère non informatif de l'échantillonnage, qui est requis pour la plupart des procédures fondées sur un modèle (Chambers et Clark, 2012, page 12). Särndal, Swensson et Wretman (1992, page 534) font remarquer que « les partisans de l'inférence fondée sur un modèle préconisent la sélection aléatoire de l'échantillon comme mesure de protection contre le biais de sélection, mais les probabilités de randomisation ne jouent aucun rôle dans l'inférence ». Voir aussi Lohr (2010, page 263), Chambers et Clark (2012, page 92) et Scott, Brewer et Ho (1978), qui proposent des plans de sondage probabiliste pour les prévisions fondées sur un modèle. Pour un examen des approches fondées sur un modèle, voir aussi Valliant, Dorman et Royall (2000).

L'inférence assistée par un modèle (par exemple, Särndal et coll., 1992) est une autre méthode, dans laquelle les estimateurs sont sans biais par rapport au plan (du moins asymptotiquement), c'est-à-dire sans biais sur un échantillonnage probabiliste répété de toute population fixe. Sous cette contrainte, ils minimisent la variance anticipée (VA), qui est la variance sur les réalisations répétées de la population à

1. Robert Graham Clark, Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, Australie. Courriel : robert.clark@anu.edu.au.

partir d'un modèle et sur un échantillonnage probabiliste répété. Dans les modèles avec erreurs indépendantes, la VA la plus basse parmi ces estimateurs (pour tout plan de sondage probabiliste donné) est la borne inférieure de Godambe et Joshi (BIGJ) (Godambe et Joshi, 1965). La borne inférieure est atteinte asymptotiquement pour les modèles linéaires par l'estimateur par la régression généralisée, qui est bien connu.

Les plans assistés par un modèle sont des plans de sondage probabiliste qui visent à réduire au minimum la VA de l'estimateur par la régression généralisée (ou, d'une manière équivalente, à minimiser la BIGJ). Ces plans de sondage optimaux pour la VA ont été calculés pour l'inférence assistée par un modèle. En particulier, le plan de sondage qui réduit au minimum la BIGJ pour une taille d'échantillon attendue fixe pour les modèles avec indépendance a une probabilité proportionnelle à la racine carrée de la variance d'erreur du modèle de chaque unité (σ_i) (par exemple, Särndal et coll., 1992). Cela est ce qu'on appellera un plan PPσ. Le plan PPCσ est une généralisation qui permet des coûts unitaires inégaux (Steel et Clark, 2014). Il n'y a pas de résultats analogues sur l'échantillonnage probabiliste optimal pour une prédiction fondée sur un modèle, sauf dans des conditions fortes, une lacune que le présent article comble partiellement. Isaki et Fuller (1982) ont proposé une stratégie estimation-plan, qui consiste à utiliser le plan PPσ pour la prédiction fondée sur un modèle. Ils ont montré que ce plan est optimal quand les probabilités de sélection et leurs carrés se trouvent dans l'espace des colonnes de la matrice des covariables. Cette condition a un prix, comme le montrera la simulation réalisée dans l'article.

Des échantillons non probabilistes optimaux ont été dérivés pour les meilleurs prédicteurs linéaires sans biais (BLUP) fondés sur un modèle sous des modèles linéaires. Cela tend à donner des plans quelque peu extrêmes, où l'on choisit les unités ayant les valeurs les plus grandes, ou les valeurs les plus grandes et les plus petites, parmi les variables auxiliaires (par exemple, Royall, 1970). Des plans équilibrés robustes fondés sur un modèle ont été élaborés. Dans ces plans, un ou plusieurs moments d'échantillonnage des variables auxiliaires sont égaux aux moments de population correspondants (Royall et Herson, 1973), tandis que les « plans suréquilibrés » respectent une contrainte différente en matière de moments d'échantillonnage (Scott et coll., 1978). Un autre modèle équilibré a été proposé par Kott (1986). Ces modèles sont robustes pour les familles de solutions polynomiales autres qu'un modèle linéaire de travail. Il ne s'agit pas de plans probabilistes, bien que certains aient proposé des plans probabilistes pour respecter approximativement les contraintes d'équilibre (Valliant et coll., 2000, section 3.4). Des plans probabilistes équilibrés exactement ont également été proposés (Tillé, 2006). Le choix d'une stratégie d'équilibrage ou de suréquilibrage dépend de l'ensemble de solutions polynomiales qui est postulé. Dans une autre approche non probabiliste, Welsh et Wiens (2013) trouvent l'échantillon qui réduit au minimum la variance maximale fondée sur un modèle dans le voisinage d'un modèle de travail.

Le présent article calcule une borne supérieure asymptotique pour la VA du BLUP sous un échantillonnage probabiliste. La VA est la quantité la plus pertinente pour le plan de sondage probabiliste, même dans le cadre fondé sur un modèle, car le calcul de la moyenne sur tous les échantillons possibles convient avant la sélection de l'échantillon. La borne s'applique à tout plan de sondage probabiliste et se

situe dans l'espace des ensembles de covariables possibles. Cela est utile pour le plan de sondage en pratique, étant donné qu'on ne décide pas précisément du modèle qui sera utilisé avant la fin de la collecte des données. Par exemple, certaines variables du plan de sondage pourraient ne pas être incluses dans le modèle si les données de l'échantillon donnent à penser qu'elles sont peu pertinentes pour la variable dont le total est estimé, mais on ne peut pas le savoir avec certitude avant la réalisation de l'enquête. On peut aussi utiliser des splines, pour lesquelles le nombre et l'emplacement des nœuds seront déterminés en fonction des données de l'échantillon. Il apparaît que la borne supérieure est la BIGJ. Cela signifie que les plans assistés par un modèle, comme PPσ et PPCσ, sont des stratégies minimax pour l'estimation fondée sur un modèle. La borne supérieure est une égalité quand le modèle a une propriété particulière, qui est satisfaite si le modèle est suffisamment riche et comprend toutes les variables du plan de sondage.

D'autres chercheurs se sont penchés sur la relation entre le BLUP et l'estimateur par la régression généralisée assisté par un modèle, y compris des conditions dans lesquelles ces deux estimateurs sont identiques (par exemple, Isaki et Fuller, 1982; Tam, 1988) et des modifications apportées au BLUP de façon à ce qu'il soit équivalent à l'estimateur par la régression généralisée aux dépens de son optimalité selon le modèle (par exemple, Brewer, Hanif et Tam, 1988; Brewer, 1999; Nedyalkova et Tillé, 2008). Les résultats donnés ici sont nouveaux pour plusieurs raisons.

- Les résultats connus ne tiennent pas compte des situations où à la fois l'enquêteur veut utiliser le BLUP parce qu'il est optimal pour le modèle et où le BLUP et les estimateurs par la régression généralisée ne sont pas égaux. On montre ici que quand les deux estimateurs sont égaux, il s'agit, en un sens, du pire cas pour le BLUP.
- On calcule une expression de la VA du BLUP et on montre explicitement qu'elle est inférieure ou égale à la borne supérieure. Intuitivement, le résultat semble raisonnable, car la BIGJ est atteinte par l'estimateur par la régression généralisée convergent par rapport au plan de sondage, tandis que le BLUP n'est pas assujetti à la contrainte de convergence par rapport au plan de sondage. Toutefois, comme il n'est pas du tout évident d'après l'expression de la VA du BLUP que la borne supérieure s'applique, il est utile d'avoir un résultat explicite.
- Selon l'interprétation, la borne supérieure se situe dans l'espace des choix possibles pour les covariables $\mathbf{x}$ du modèle. Par conséquent, la borne supérieure est pertinente quand la personne concevant le plan de sondage ne connaît pas le modèle qui sera finalement adopté après la collecte des données.

La section 2 présente les principaux résultats théoriques. La section 3 confirme et illustre le résultat principal dans une étude par simulations avec une variable d'intérêt Y et deux variables auxiliaires : x_1 (continue) et x_2 (binaire). La valeur espérée de Y conditionnelle à ces variables est définie par un terme linéaire et un terme sinusoïdal en x_1. Elle ne dépend pas de x_2. Les probabilités de sélection sont une fonction de x_1 et x_2. Les BLUP sont calculés à partir du modèle ayant le critère d'information bayésien (CIB) le plus bas à partir d'un ensemble comprenant le modèle linéaire simple en x_1 et des splines en x_1

de différents degrés, avec et sans x_2. Le ratio de l'erreur quadratique moyenne (EQM) de la prédiction de la simulation du BLUP par rapport à la BIGJ est soit inférieur ou égal à 1, soit légèrement supérieur à 1 dans différentes situations. La section 4 est une discussion.

Une grande partie de la littérature comparant les estimateurs et les inférences assistés par un modèle et fondés sur un modèle s'est surtout intéressée au biais causé par des modèles mal spécifiés quand (a) la fonction moyenne est incorrecte ou (b) certaines variables du plan sont exclues de façon inappropriée. Voir par exemple Hansen, Madow et Tepping (1983) et le travail réalisé sur leur étude par simulations par Valliant et coll. (2000, section 3.4). La simulation de la section 3 prend en considération (a) et (b) dans une certaine mesure, mais ce n'est pas l'objet principal de l'article. Ici, l'objectif est de déterminer si un statisticien qui choisit un plan fondé sur un modèle peut utiliser la BIGJ comme borne supérieure pour la VA aux fins du plan de sondage, plutôt que de trancher entre une inférence fondée sur un modèle et une inférence fondée sur le plan. On suppose qu'un modèle suffisamment bon peut être trouvé au moyen des données de l'échantillon. Ce processus serait facilité si le plan minimise la VA maximale sur l'espace de tous les modèles linéaires. Il est recommandé d'utiliser un plan PPσ ou PPCσ en cas de grande incertitude sur la forme du modèle final.

2 Borne supérieure pour la VA du meilleur prédicteur linéaire sans biais (BLUP)

Soit $U = \{1, \ldots, N\}$ la population finie. Pour l'unité $i \in U$, la variable d'intérêt est y_i et le vecteur p des variables auxiliaires est $\mathbf{x}_i$. L'échantillon (de taille n) est s et l'ensemble sans échantillonnage est $r = U - s$. Les variables auxiliaires sont observées pour tous les $i \in U$ tandis que y_i est observé pour $i \in s$. L'objectif est de prédire $t_y = \sum_{i \in U} y_i$. Les probabilités de sélection sont $\pi_i = P[i \in s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N, y_1, \ldots, y_N] = P[i \in s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N] > 0$. On suppose qu'elles sont une fonction des valeurs de population de $\mathbf{x}_i$. Soit $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$ et $\mathbf{t}_{xr} = \sum_{i \in r} \mathbf{x}_i$.

La matrice n par p des valeurs d'échantillon de $\mathbf{x}$, qui a les lignes $\mathbf{x}_i^T$, est notée X_s. La matrice $N - n$ par p des valeurs sans échantillonnage de $\mathbf{x}$ est X_r. Le vecteur des valeurs d'échantillon de y est $\mathbf{y}_s$.

On suppose le modèle linéaire suivant M :

$$E_M[y_i] = \boldsymbol{\beta}^T \mathbf{x}_i \tag{2.1}$$

$$\text{var}_M[y_i] = \sigma_i^2 = \sigma^2 v_i \tag{2.2}$$

$$\text{cov}_M[y_i, y_j] = 0 \tag{2.3}$$

pour $i, j \in U$ avec $i \neq j$. Les indices M dans E_M, var_M et cov_M indiquent des distributions sur des réalisations répétées des valeurs de population à partir du modèle. On suppose généralement que les v_i sont connus, c'est-à-dire que les variances d'erreur sont connues à une constante de proportionnalité près.

Par exemple, dans les enquêtes-entreprises, v_i peut être une mesure de la taille de l'entreprise ou de la racine carrée de celle-ci. Les paramètres inconnus sont $\boldsymbol{\beta}$ et σ^2. Les valeurs de $\mathbf{x}_i$ sont considérées comme étant fixes.

Le meilleur prédicteur linéaire sans biais (BLUP) (noté $\hat{t}_y$) pour une généralisation du modèle M est donné au chapitre 2 de Valliant et coll. (2000). Sa variance de prédiction fondée sur un modèle est

$$\text{var}_M\left(\hat{t}_y - t_y\right) = \mathbf{t}_{xr}^T\left(\sum_{i\in s}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\mathbf{t}_{xr} + \sum_{i\in r}\sigma_i^2. \tag{2.4}$$

(On peut l'obtenir comme cas particulier du résultat 2.2.2 de la page 29 de Valliant et coll., 2000.)

La variance attendue est définie comme étant $\text{VA}\left(\hat{t}_y - t_y\right) = E_M E_p\left(\hat{t}_y - t_y\right)^2$ (Isaki et Fuller, 1982). Comme $\hat{t}_y$ est sans biais par rapport au modèle, sa VA est égale à

$$\text{VA} = E_p\text{var}_M\left(\hat{t}_y - t_y\right).$$

Le théorème 1 produira une approximation de cette VA. Le cadre asymptotique est basé sur les propriétés asymptotiques fondées sur le plan d'Isaki et Fuller (1982). On suppose une population infinie dénombrable $i = 1, 2, \ldots$. Une séquence de populations finies U_t est définie par $U_t = \{1, \ldots, N_t\}$ où $N_1 < N_2 < \ldots$. Pour chaque t, un échantillon s_t de taille n_t est sélectionné à partir de U_t par un plan de sondage probabiliste avec des probabilités de sélection $\pi_{i(t)} = P[i \in s_t]$. D'après (2.7) d'Isaki et Fuller (1982), on suppose que

$$0 < \lambda_1 < \pi_{i(t)} < \lambda_2 \tag{2.5}$$

pour certaines constantes λ_1 et λ_2. Isaki et Fuller (1982) constatent que les VA des estimateurs du total sont habituellement $O(n_t)$ (ce qui équivaut à $O(N_t)$) et que les totaux eux-mêmes sont également $O(n_t)$. Les moyennes de population seront désignées par $\bar{\mathbf{X}}_t = N_t^{-1}\sum_{U_t}\mathbf{x}_i$ et l'estimateur pondéré de la probabilité inverse de $\bar{\mathbf{X}}$ est $\hat{\bar{\mathbf{X}}}_\pi = N_t^{-1}\sum_{s_t}\pi_{i(t)}^{-1}\mathbf{x}_i$ (et de la même façon pour Y et les autres variables).

Deux nouvelles variables sont définies pour chaque unité i par $\mathbf{u}_{i(t)} = \pi_{i(t)}\mathbf{x}_i$ (un vecteur p) et $\mathbf{v}_{i(t)} = \pi_{i(t)}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T$ (une matrice p par p). Leurs moyennes de population sont $\bar{\mathbf{U}}_t$ et $\bar{\mathbf{V}}_t$ avec des estimateurs de probabilité inverse $\hat{\bar{\mathbf{U}}}_{t\pi}$ et $\hat{\bar{\mathbf{V}}}_{t\pi}$.

Théorème 1. *Supposons que*

$$\lim_{t\to\infty} E_p\left\{\hat{\bar{\mathbf{U}}}_{t\pi}^T\hat{\bar{\mathbf{V}}}_{t\pi}^{-1}\hat{\bar{\mathbf{U}}}_{t\pi} - \bar{\mathbf{U}}_t^T\bar{\mathbf{V}}_t^{-1}\bar{\mathbf{U}}_t\right\} = 0. \tag{2.6}$$

Alors

$$E_p var_M\left(\hat{t}_y - t_y\right) = VA + o(n_t). \tag{2.7}$$

où

$$VA = \sum_U(1-\pi_i)\mathbf{x}_i^T\left(\sum_U\pi_i\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\sum_U(1-\pi_i)\mathbf{x}_i + \sum_U(1-\pi_i)\sigma_i^2. \tag{2.8}$$

Remarques sur le théorème 1.

- L'hypothèse (2.6) rappelle le résultat (3.24) d'Isaki et Fuller (1982), mais elle comporte une différence importante. Dans Isaki et Fuller (1982), les variables d'unités dépendent seulement de i, alors qu'ici $\mathbf{u}_{i(t)}$ et $\mathbf{v}_{i(t)}$ dépendent à la fois de i et de t, car elles ont toutes les deux un facteur $\pi_{i(t)}$. Cependant, $\pi_{i(t)}$ sont bornés par (2.5), de sorte que la condition est plausible; elle ne le serait pas si $\pi_{i(t)}$ pouvait être arbitrairement proche de zéro.
- Il est clair que l'hypothèse (2.6) est satisfaite si $\hat{\bar{\mathbf{U}}}_{t\pi}$ et $\hat{\bar{\mathbf{V}}}_{t\pi}$ sont convergents selon la probabilité du plan pour $\bar{\mathbf{U}}_t$ et $\bar{\mathbf{V}}_t$, et si $\hat{\bar{\mathbf{V}}}_{t\pi}$ est inversible dans un voisinage de $\bar{\mathbf{V}}_t$. Comme l'indiquent Isaki et Fuller (1982) dans un commentaire sur leur condition (3.12), une exigence d'invertibilité de ce genre semble raisonnable « pour toute discussion sur l'estimation par la régression ».

Une borne supérieure pour la VA asymptotique sur tous les choix possibles du vecteur auxiliaire $\mathbf{x}_i$ sera maintenant calculée. Cela permet une incertitude quant aux variables auxiliaires qui seront finalement incluses dans le modèle, parce qu'en général, cette décision est prise seulement après la collecte des données. Par exemple, l'ensemble complet des variables utilisées dans le plan de sondage pourrait ou non se retrouver dans le modèle, ou les fonctions splines des covariables pourraient être incluses avec des nœuds fondés sur les données de l'échantillon. Le théorème 2 énonce la borne supérieure.

Théorème 2. *Soit VA la VA asymptotique définie par (2.8). Si $\sum_U \pi_i \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T$ est inversible et $\pi_i > 0$ pour tous les $i \in U$, alors*

$$VA \leq \sum_U \left(\pi_i^{-1} - 1\right) \sigma_i^2 \tag{2.9}$$

avec une égalité stricte si et seulement s'il existe un vecteur p $\boldsymbol{\lambda}$ tel que

$$\left(\pi_i^{-1} - 1\right) \sigma_i^2 = \boldsymbol{\lambda}^T \mathbf{x}_i \tag{2.10}$$

pour tous les $i \in U$.

Le deuxième membre de (2.9) est la borne inférieure bien connue de Godambe et Joshi (Godambe et Joshi, 1965) pour la VA des estimateurs sans biais par rapport au plan de sondage. Il s'agit ici d'une borne supérieure sur l'espace de modèle pour les BLUP fondés sur des modèles.

Supposons que le coût total d'exécution de l'enquête est $\sum_s C_i$ plus les coûts fixes, où C_i est le coût associé à l'unité d'enquête i. Le coût prévu est alors de $C_E = \sum_U C_i \pi_i$. Le plan de sondage qui minimise la borne supérieure dans (2.9) assujetti au coût fixe est le plan PPC σ qui a

$$\pi_i \propto \sigma_i / \sqrt{C_i} \tag{2.11}$$

(Steel et Clark, 2014, qui ont généralisé Särndal et coll., 1992, page 452) pour permettre des coûts inégaux. Le théorème 2 signifie que (2.11) est un plan minimax en cas d'incertitude sur la forme du modèle. Notons que seules les probabilités d'inclusion du premier ordre ont une incidence sur la VA et la

borne, mais elles ne spécifient pas entièrement le plan. On peut sélectionner les échantillons au moyen de ces probabilités d'inclusion de plusieurs manières (Tillé, 2006), notamment par échantillonnage probabiliste équilibré (Nedyalkova et Tillé, 2012), ce qui améliore la robustesse par rapport à la spécification erronée de modèle.

La condition d'égalité (2.10) équivaut à une condition bien connue pour que le BLUP soit égal à l'estimateur par la régression généralisée (formule 3 de Tam, 1988). Tam (1988) défend l'utilisation de plans de sondage permettant de satisfaire la condition (2.10), comme PPσ (à condition que le modèle comprenne une ordonnée à l'origine). En s'appuyant sur un résultat de Royall (1992), Nedyalkova et Tillé (2008) ont montré que PPσ est un plan optimal fondé sur un modèle en cas de coûts égaux quand v_i et $\sqrt{v_i}$ sont des fonctions linéaires de $\mathbf{x}_i$, soit une condition dite de *variances explicables*. Brewer et coll. (1988) ont constaté que la condition (2.10) peut également être satisfaite si le modèle d'estimation comprend une variable instrumentale, qui est une fonction appropriée des probabilités de sélection. Cependant, dans de nombreuses circonstances, la condition (2.10) n'est pas satisfaite, parce que certaines variables auxiliaires sont omises du modèle final, parce que plusieurs variables d'intérêt ont différentes structures de variance (ce qui écarte PPσ), parce que les coûts sont inégaux, ou parce que les variables instrumentales sont évitées en raison de la perte d'efficacité qu'elles impliquent. Le théorème 2 montre que la BIGJ est une borne supérieure dans ces circonstances qui ne sont pas traitées dans les résultats de ces auteurs. Le plan PPCσ dans (2.11) est un plan minimax dans ce contexte plus général.

3 Étude par simulations

On a mené une étude par simulations pour comparer la VA du BLUP et sa borne supérieure dans des situations où Y a une relation non linéaire avec une variable auxiliaire continue x_1. Une deuxième variable auxiliaire, x_2, est binaire et indépendante de x_1 et Y. Les probabilités de sélection dépendent de x_1 et x_2 de différentes manières. Pour chaque scénario, 5 000 populations et échantillons ont été produites, avec une population de 6 000 habitants et des tailles d'échantillon de 500 et 1 500, et avec une taille de population de 100 000 et une taille d'échantillon de 25 000. Tous les programmes sont disponibles au www.github.com/rgcstats/AVLB.

3.1 Simulation de populations

Les valeurs de population de x_1 étaient les quantiles $\{j/(101): j = 1, \ldots, 100\}$ d'une distribution lognormale avec une moyenne $-1/32$ et un écart-type de 0,25, avec des fréquences égales de chacune de ces 100 valeurs. Cela signifie que les x_1 sont positives et asymétriques à droite avec une moyenne de 1 et un écart de 0,54 à 1,74. On a fait en sorte que les valeurs de x_1 soient non stochastiques et discrétisées afin d'accélérer le calcul, de simplifier la génération de modèles lisses (voir ci-dessous) et de faciliter la comparaison des VA avec la BIGJ en la rendant constante d'une simulation à l'autre. Une deuxième variable binaire x_2 a pris les valeurs 0 et 1 avec une fréquence égale dans chaque valeur de x_1, de sorte que les deux covariables étaient orthogonales.

Conditionnelles à x_1 et x_2, les valeurs de population de Y ont été produites indépendamment comme étant

$$E_M Y_i = \mu(x_{1i}) + \varepsilon_i \tag{3.1}$$

où

$$\varepsilon_i \sim N(0;\ 0{,}25x_{1i}). \tag{3.2}$$

La fonction moyenne $\mu(.)$ était une fonction lisse mais non linéaire,

$$\mu(x_1) = 4x_1 + \sin(x_1\, 2\pi/h), \tag{3.3}$$

consistant en un terme linéaire et un terme sinusoïdal avec une période h. Quand h est grand, $\mu(\,)$ est presque linéaire sur la gamme de x_1 dans la population, tandis que si h est petit, il y a des cycles fréquents dans la fonction. La figure 3.1 montre la fonction moyenne pour les périodes utilisées dans la simulation (0,5; 1; 2; 5).

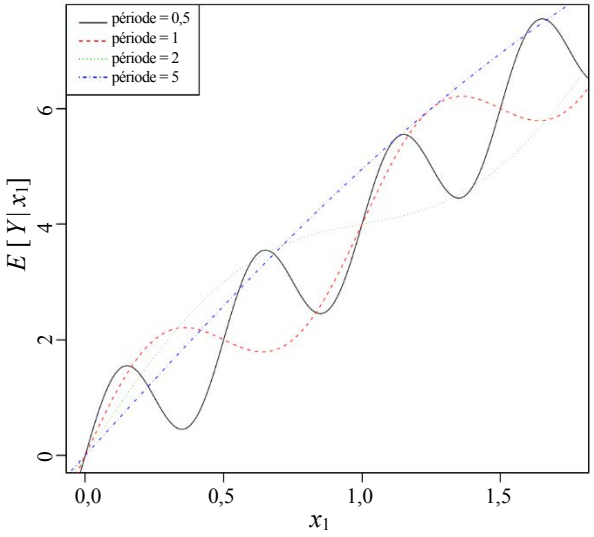

Figure 3.1 $\mu(x_1) = E[Y \mid x_1]$ pour les périodes utilisées dans l'étude par simulations.

3.2 Échantillon simulé

Les probabilités de sélection, π_i, ont été établies à

$$\pi_i \propto x_{1i}^b (1 + cx_{2i}) \tag{3.4}$$

où b était 0,5; 1 ou 2, soit une dépendance légère, moyenne ou élevée à x_{1i}. Les valeurs de c étaient 0; 0,5 ou 1,5, soit une dépendance nulle, moyenne ou élevée à x_{2i}. (Les autres valeurs de c ont également été utilisées aux fins de la figure 3.2 seulement.)

La deuxième variable auxiliaire, x_2, n'est pas liée à Y mais elle peut avoir une incidence sur les probabilités de sélection. On pourrait s'attendre à ce que le BLUP donne de meilleures résultats que la BIGJ quand les probabilités de sélection dépendent de x_2, étant donné que le BLUP peut être fondé sur un modèle omettant x_2, ce qui pourrait entraîner une variance plus faible. Bien entendu, il est aussi possible que le modèle de travail omette x_1, ce qui entraîne un biais du BLUP, mais cela ne s'est produit dans aucune des simulations. Pour examiner la robustesse de l'omission incorrecte de x_1, il serait intéressant d'examiner des relations plus faibles que celles présentées à la figure 3.1, mais cela dépasse l'objet du présent article.

Les probabilités d'inclusion sont forcées d'obéir à la proportionnalité dans (3.4), mais elles sont tronquées vers le haut à 1 et vers le bas à 1/40 et mises à l'échelle de sorte qu'elles s'ajoutent à la taille d'échantillon requise après troncature. Les échantillons sont sélectionnés par échantillonnage systématique à probabilités inégales avec ordonnancement aléatoire au moyen du progiciel d'`échantillonnage` de R (Tillé et Matei, 2016).

3.3 Estimation du total de la population de Y

Un modèle linéaire dans x_1 et des modèles splines dans x_1 comprenant entre 1 et 10 nœuds intérieurs sont ajustés à chaque échantillon. (À propos de l'utilisation de splines dans l'estimation d'enquête par sondage assistée par un modèle, voir par exemple Breidt, Claeskens et Opsomer, 2005.) On définit onze autres modèles en incluant également x_2 comme covariable additive. On utilise ensuite le modèle ayant le critère d'information bayésien (CIB) le plus bas pour calculer un BLUP de t_y. Cette étape de sélection du modèle devrait accroître la variabilité du prédicteur. Le BLUP fondé sur le modèle linéaire simple de x_1 est également calculé. On répète le processus avec des modèles de travail, y compris la spécification de variance correcte $\text{var}_M(y_i) \propto x_{1i}$ et une spécification erronée $\text{var}_M(y_i) \propto x_{1i}^2$.

3.4 Résultats de simulation

Les tableaux 3.1 à 3.4 montrent les ratios des EQM de prédiction de plusieurs estimateurs BLUP par rapport à la BIGJ (deuxième membre de l'équation 2.9) pour différents plans de sondage et choix de $\mu(x_1)$. Les EQM de prédiction sont les moyennes de toutes les simulations de $\left(\hat{t}_y - t_y\right)^2$ de sorte qu'elles se rapportent à la fois au modèle et au plan, tout comme les résultats des théroèmes 1 et 2.

Le tableau 3.1 a évalue le BLUP correspondant au modèle ayant le CIB le plus bas pour une taille d'échantillon de 500 avec des variances correctement spécifiées. Neuf plans de sondage sont présentés; ils correspondent à trois choix pour b (soit une dépendance faible, moyenne, ou élevée des probabilités de sélection à x_1) et c (soit une dépendance nulle, partielle, ou élevée des probabilités de sélection à x_2). La période h de la composante sinusoïdale de $\mu(x_1)$ est également indiquée (voir l'équation 3.3). Le tableau montre que :

- Le ratio est toujours inférieur ou égal à 1 ou légèrement supérieur à 1, ce qui concorde avec le théorème 2. Son écart va de 0,815 à 1,086.

- Le ratio diminue légèrement quand h augmente. Par conséquent, quand la valeur vraie de $E_M(y|x_1)$ est presque linéaire, le BLUP fondé sur un modèle a une EQM plus basse par rapport à la BIGJ, alors que pour les modèles davantage non linéaires, le ratio est plus proche de 1.
- Le ratio dépend de b dans une certaine mesure, bien que la tendance dépende des autres paramètres.
- Le ratio diminue considérablement quand c augmente, avec des réductions allant jusqu'à 20 % par rapport à $c = 0$. Cela montre que le BLUP donne des résultats nettement meilleurs que la BIGJ en présence d'une covariable x_2 qui est pertinente dans le plan, mais non pertinente pour Y si bien qu'elle peut être omise dans le modèle d'estimation.

Le tableau 3.1b présente les résultats pour un échantillon beaucoup plus grand de 25 000 personnes sur une population de 100 000. On a inclus ces résultats pour voir si les ratios sont inférieurs ou égaux à 1 pour les grandes valeurs n comme le prédit la théorie de la section 2. On a également utilisé un plus grand nombre de simulations pour ce panel (15 000 au lieu de 5 000). Les résultats sont uniquement présentés pour $c = 0$ car il s'agissait des plans ayant les ratios les plus élevés dans le tableau 3.1a. Les ratios du tableau 3.1b vont de 0,984 à 1,011. Les valeurs légèrement supérieures à 1 peuvent indiquer que le modèle spline de travail ne saisit pas parfaitement les fonctions sinusoïdales utilisées pour produire les données.

Le tableau 3.2 présente les mêmes situations que le tableau 3.1a, sauf que le BLUP est fondé sur un modèle de variance incorrectement spécifiée avec $\sigma_i^2 \propto x_i^2$ (le modèle générateur a $\sigma_i^2 \propto x_i$). Dans l'ensemble, les ratios de l'EQM du BLUP fondé sur un modèle au CIB le plus bas par rapport à la BIGJ sont légèrement supérieurs à ceux du tableau 3.1 (généralement de moins d'un point de pourcentage). Il reste que le ratio est presque toujours inférieur ou égal à 1, avec une valeur maximale de 1,089.

Le tableau 3.3 est identique au tableau 3.1a, sauf que la taille de l'échantillon est de 1 500 plutôt que de 500. Les ratios sont presque toujours inférieurs à ceux du tableau 3.1a. Le ratio maximal est de 1,030.

Le tableau 3.4 présente les résultats du BLUP fondé sur le modèle linéaire simple contenant seulement x_1 avec une variance incorrectement spécifiée. La taille de l'échantillon est de 500. Quand la période est 5, de sorte que le modèle vrai est presque linéaire dans x_1, ce BLUP fonctionne très bien. Les ratios sont alors toujours inférieurs à 1,1 et peuvent être aussi bas que 0,790. Quand la période est 2, il y a une courbure visible dans $E(y|x_1)$ (comme le montre la figure 3.1), mais le BLUP simple ne donne pas de bons résultats, tous les ratios étant inférieurs à 1,2. Cependant, en cas de périodes égales à 0,5 et 1, les ratios sont nettement supérieurs à 1, avec une valeur maximale de 3,4. Cela montre le biais important du BLUP en cas de spécification erronée du modèle.

La mesure dans laquelle les probabilités de sélection dépendent de x_2 est le principal facteur déterminant le ratio de l'EQM par rapport à la BIGJ, comme le montrent les tableaux 3.1 à 3.3. La

figure 3.2 montre ce phénomène de façon plus détaillée pour une variance correctement spécifiée. La taille de l'échantillon est 1 500 avec échantillonnage PPσ de sorte que $\pi_i \propto x_{1i}^{0,5}$. Les valeurs de c (0; 0,25; ..., 3) se trouvent sur l'axe des x et les résultats sont présentés pour différentes périodes h. La figure montre que le ratio est légèrement supérieur à 1 pour $c = 0$ et diminue de façon lissée avec c, à environ 0,6 quand $c = 3$. Des périodes plus élevées h (qui reflètent une relation plus lisse entre Y et x_1) sont également associées à des ratios plus faibles, mais les différences sont si faibles qu'il est presque impossible de les discerner à la figure 3.2.

Tableau 3.1
Ratios de l'EQM du BLUP fondé sur un modèle spline au plus bas CIB par rapport à la borne inférieure de Godambe et Joshi pour les tailles d'échantillon allant de 500 à 25 000 avec une variance correctement spécifiée. Les probabilités de sélection sont proportionnelles à $x_1^b(1 + cx_2)$. La période h contrôle le lissage de $E[Y \mid x_1]$

plan de sondage		période (h)			
b	c	0,5	1	2	5
(a) taille d'échantillon de 500					
0,5	0	1,086	1,064	1,052	1,057
0,5	0,5	0,990	0,963	0,951	0,956
0,5	1,5	0,840	0,827	0,808	0,815
1	0	1,033	1,015	0,997	1,006
1	0,5	1,006	0,992	0,973	0,985
1	1,5	0,877	0,859	0,854	0,858
2	0	1,080	1,063	1,035	1,081
2	0,5	1,046	1,021	0,996	1,039
2	1,5	0,870	0,853	0,839	0,856
(b) taille d'échantillon de 25 000					
0,5	0	1,011	1,011	1,011	1,010
1	0	1,007	1,006	1,006	1,006
2	0	0,985	0,985	0,984	0,984

Tableau 3.2
Ratios de l'EQM du BLUP fondé sur un modèle au plus bas CIB par rapport à la borne inférieure de Godambe et Joshi pour une taille d'échantillon de 500 avec spécification erronée de la variance. Les probabilités de sélection sont proportionnelles à $x_1^b(1 + cx_2)$. La période h contrôle le lissage de $E[Y \mid x_1]$

plan de sondage		période (h)			
b	c	0,5	1	2	5
0,5	0	1,089	1,068	1,055	1,061
0,5	0,5	0,996	0,967	0,959	0,962
0,5	1,5	0,852	0,830	0,819	0,825
1	0	1,033	1,012	0,998	1,001
1	0,5	1,006	0,992	0,978	0,988
1	1,5	0,886	0,863	0,854	0,862
2	0	1,078	1,061	1,035	1,047
2	0,5	1,048	1,017	0,997	1,010
2	1,5	0,878	0,857	0,842	0,856

Tableau 3.3
Ratios de l'EQM du BLUP fondé sur un modèle au plus bas CIB par rapport à la borne inférieure de Godambe et Joshi pour une taille d'échantillon de 1 500 avec une variance correctement spécifiée. Les probabilités de sélection sont proportionnelles à $x_1^b(1 + cx_2)$. La période h contrôle le lissage de $E[Y \mid x_1]$

plan de sondage		période (h)			
b	c	0,5	1	2	5
0,5	0	1,030	1,023	1,019	1,022
0,5	0,5	0,928	0,920	0,918	0,922
0,5	1,5	0,762	0,754	0,750	0,749
1	0	0,940	0,936	0,930	0,932
1	0,5	0,979	0,972	0,966	0,968
1	1,5	0,821	0,817	0,810	0,809
2	0	1,028	1,008	0,995	1,020
2	0,5	0,962	0,948	0,941	0,969
2	1,5	0,798	0,798	0,786	0,804

Tableau 3.4
Ratios de l'EQM du BLUP fondé sur un modèle linéaire simple par rapport à la borne inférieure de Godambe et Joshi pour une taille d'échantillon de 500 avec spécification erronée de la variance. Les probabilités de sélection sont proportionnelles à $x_1^b(1 + cx_2)$. La période h contrôle le lissage de $E[Y \mid x_1]$

plan de sondage		période (h)			
b	c	0,5	1	2	5
0,5	0	3,083	2,037	1,109	1,055
0,5	0,5	2,918	1,860	1,002	0,958
0,5	1,5	2,452	1,533	0,840	0,790
1	0	3,086	1,812	1,052	1,007
1	0,5	3,006	1,792	1,016	0,979
1	1,5	2,537	1,500	0,880	0,838
2	0	3,423	2,900	1,174	1,080
2	0,5	3,243	2,693	1,111	1,025
2	1,5	2,689	2,291	0,926	0,829

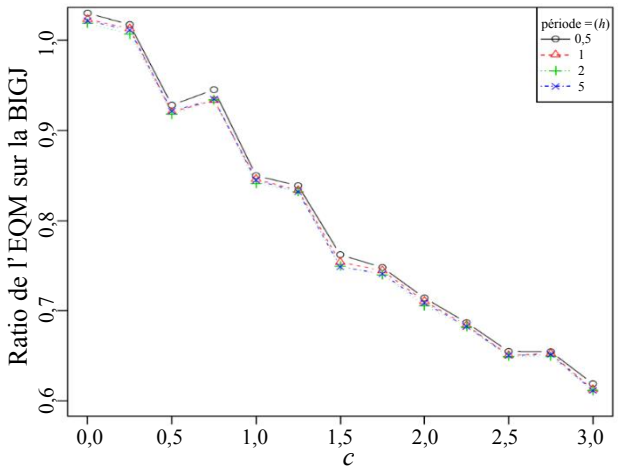

Figure 3.2 Ratios de l'EQM du BLUP fondé sur un modèle au plus bas CIB par rapport à la borne inférieure de Godambe et Joshi pour une taille d'échantillon de 1 500 avec une variance correctement spécifiée, comparativement à c, où les probabilités de sélection sont proportionnelles à $x_1^b(1 + cx_2)$ avec $b = 1$. La période (h) contrôle le lissage de $E[Y \mid x_1]$.

4 Discussion

La borne inférieure Godambe et Joshi est présentée ici comme une borne *supérieure* pour les variances anticipées de BLUP dans un modèle correct. Les EQM de simulation des BLUP fondés sur un modèle linéaire ou spline choisi adaptativement sont toujours inférieures à la BIGJ ou juste au-dessus, y compris en cas de spécification erronée des variances. Les EQM sont bien en dessous de la borne si une variable du plan importante ne figure pas dans le modèle.

Le résultat de la borne supérieure dépend de l'absence de biais du BLUP par rapport au modèle. Dans l'étude par simulations, les BLUP fondés sur un modèle incorrectement spécifié avaient des EQM nettement supérieures à la borne. Le choix d'un modèle de travail ayant un CIB minimal dans une classe comprenant des modèles splines a permis d'éviter ce problème.

Une fois que les données sont disponibles, les conditions d'inférence fondée sur un modèle pour l'échantillon sont sélectionnées. L'échantillonnage probabiliste présente néanmoins de nombreux avantages même s'il n'est pas la base de l'inférence (par exemple, Särndal et coll., 1992; Valliant et coll., 2013). À l'étape du plan, la variance anticipée est l'objectif le plus pertinent, parce qu'elle fait la moyenne sur tous les échantillons possibles qui peuvent alors être sélectionnés. La borne supérieure calculée ici est pertinente à l'étape du plan, parce qu'il y a généralement de grandes incertitudes sur la forme du modèle qui sera finalement adopté (il peut y avoir des exceptions quand on dispose de données historiques ou connexes appuyant la spécification d'un modèle ou quand on est prêt à croire que le vrai modèle se trouve dans une classe de modèles polynomiaux ou d'autres modèles spécifiques).

En pratique, il est judicieux d'adopter une stratégie consistant à :

i. établir π_i de façon à obtenir des valeurs faibles pour la borne supérieure $\sum_{i\in U}\left(\pi_i^{-1}-1\right)v_i$ où les variances du modèle sont proportionnelles à v_i (ou à une combinaison pondérée des bornes supérieures pour plusieurs variables d'intérêt), tout en respectant le coût et les considérations pratiques. S'il y a une seule variable d'intérêt et que les coûts unitaires sont proportionnels à C_i, alors $\pi_i \propto \sqrt{v_i/C_i}$ est recommandé, car il s'agit d'une stratégie minimax;

ii. une fois que l'échantillon est sélectionné et que les données sont disponibles, choisir un modèle de régression fondé sur ces données;

iii. estimer les totaux de population au moyen des BLUP sous le modèle sélectionné;

iv. cela peut conduire à ce que la condition (2.10) soit respectée ou pas, selon les coûts C_i et les variables auxiliaires sélectionnées dans le modèle final.

Tous les résultats optimaux du plan de sondage, qu'ils soient fondés sur un modèle, fondés sur un plan de sondage ou assistés par un modèle, reposent sur la connaissance des variances résiduelles relatives de l'unité ou de la strate. Cela semble inévitable. Le présent article est utile lorsque la forme du modèle moyen n'est pas connue à l'avance, car il donne une borne supérieure sur l'espace des modèles pour la moyenne. Il ne semble pas y avoir de borne aussi utile sur les modèles de variance possibles, si bien que la forme du modèle de variance doit être estimée ou supposée.

Remerciements

Je tiens à remercier Stephen Haslett de m'avoir encouragé à choisir de mener une recherche méthodologique malgré mes nombreuses autres priorités. Par ailleurs, l'article a considérablement été amélioré par la lecture minutieuse réalisée par deux examinateurs et un rédacteur adjoint.Cette recherche a été entreprise avec l'aide de ressources de la *National Computational Infrastructure* (NCI Australia), sous l'égide du NCRIS soutenu par le gouvernement australien.

Annexe

Démonstration du théorème 1

De (2.4),

$$E_p \mathrm{var}_M \left(\hat{t}_y - t_y\right) = E_p \left\{ \mathbf{t}_{xr}^T \left(\sum_{i \in s} \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{t}_{xr} \right\} + \sum_{i \in U} (1 - \pi_i)\, \sigma_i^2 . \tag{A.1}$$

À partir des définitions de $\hat{\bar{\mathbf{U}}}_\pi$ et $\hat{\bar{\mathbf{V}}}_\pi$ et de l'hypothèse (2.6), le premier terme de (A.1) devient

$$\begin{aligned} E_p \left\{ \mathbf{t}_{xr}^T \left(\sum_{i \in s} \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{t}_{xr} \right\} &= E_p \left\{ \left(N_t \bar{\mathbf{X}} - N_t \hat{\bar{\mathbf{U}}}_\pi \right)^T \left(N_t \hat{\bar{\mathbf{V}}}_\pi \right)^{-1} \left(N_t \bar{\mathbf{X}} - N_t \hat{\bar{\mathbf{U}}}_\pi \right) \right\} \\ &= N_t E_p \left\{ \left(\bar{\mathbf{X}} - \hat{\bar{\mathbf{U}}}_\pi \right)^T \hat{\bar{\mathbf{V}}}_\pi^{-1} \left(\bar{\mathbf{X}} - \hat{\bar{\mathbf{U}}}_\pi \right) \right\} \\ &= N_t \left\{ \left(\bar{\mathbf{X}} - \bar{\mathbf{U}} \right)^T \bar{\mathbf{V}}^{-1} \left(\bar{\mathbf{X}} - \bar{\mathbf{U}} \right) + o(1) \right\}. \end{aligned} \tag{A.2}$$

Le résultat découle immédiatement de (A.1) et (A.2).

Lemme 1 : Soit $a_1, \ldots, a_n$ et $b_1, \ldots, b_n$ des scalaires où $b_i > 0$ pour tous les i. On suppose que $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont des vecteurs p. Alors

$$\left(\sum_{i=1}^N a_i \mathbf{x}_i \right)^T \left(\sum_{i=1}^N b_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_{i=1}^N a_i \mathbf{x}_i \right) \le \sum_{i=1}^N a_i^2 / b_i \tag{A.3}$$

à condition que l'inverse de la matrice existe. L'égalité dans (A.3) est obtenue si et seulement si

$$a_i b_i^{-1} = \boldsymbol{\lambda}^T \mathbf{x}_i \tag{A.4}$$

pour tous les $i = 1, \ldots, n$ pour certains vecteurs p $\boldsymbol{\lambda}$.

Démonstration du lemme 1

Soit $b = \sum_{i=1}^n b_i$. Soit X une variable aléatoire discrète prenant les valeurs a_i / b_i. Soit $\mathbf{Y}$ une variable aléatoire discrète prenant les valeurs $\mathbf{x}_i$, pour $i = 1, \ldots, n$. Soit $P[Y = \mathbf{x}_i, X = a_i / b_i] = b_i / b$ pour $i = 1, \ldots, n$. Écrire $M_1 \le M_2$ si $M_1 - M_2$ est semi-définie négative pour toutes les matrices M_1 et M_2. Selon le théorème 1 de Tripathi (1999), pour tous les vecteurs aléatoires $\mathbf{X}$ et $\mathbf{Y}$,

$$E\left[\mathbf{X}\mathbf{Y}^T\right]\left\{E\left[\mathbf{Y}\mathbf{Y}^T\right]\right\}^{-1} E\left[\mathbf{Y}\mathbf{X}^T\right] \leq E\left[\mathbf{X}\mathbf{X}^T\right] \tag{A.5}$$

à condition que l'inverse de la matrice existe. Avec ma définition de X et $\mathbf{Y}$, (A.5) devient

$$\sum_{i=1}^{N} b_i b^{-1} a_i b_i^{-1} \mathbf{x}_i^T \left\{\sum_{i=1}^{N} b_i b^{-1} \mathbf{x}_i \mathbf{x}_i^T\right\}^{-1} \sum_{i=1}^{N} b_i b^{-1} a_i b_i^{-1} \mathbf{x}_i \leq \sum_{i=1}^{N} b_i b^{-1} a_i^2 b_i^{-2} \tag{A.6}$$

ce qui donne directement (A.3). Tripathi (1999) énonce que l'égalité est nette si

$$\mathbf{X}^T \boldsymbol{\lambda}_1 + \mathbf{Y}^T \boldsymbol{\lambda}_2 = 0 \tag{A.7}$$

avec une probabilité de 1 pour certaines valeurs λ_1 de la même dimension que $\mathbf{X}$ et λ_2 de la même dimension que $\mathbf{Y}$. Ici, (A.7) devient

$$a_i b_i^{-1} \lambda_1 = \mathbf{x}_i^T \boldsymbol{\lambda}_2$$

pour tous les i, ce qui équivaut à (A.4).

Démonstration du théorème 2

Soit $a_i = 1 - \pi_i$ et $b_i = \pi_i \sigma_i^{-2}$. Pour le lemme 1,

$$\begin{aligned}
\text{VA} &= \sum_U (1 - \pi_i)\, \mathbf{x}_i^T \left(\sum_U \pi_i \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_U (1 - \pi_i)\, \mathbf{x}_i^T + \sum_U (1 - \pi_i)\, \sigma_i^2 \\
&\leq \sum_U (1 - \pi_i)^2\, \pi_i^{-1} \sigma_i^2 + \sum_U (1 - \pi_i)\, \sigma_i^2 \\
&= \sum_U (1 - \pi_i)\, \pi_i^{-1} \sigma_i^2\, (1 - \pi_i + \pi_i) \\
&= \sum_U \left(\pi_i^{-1} - 1\right) \sigma_i^2
\end{aligned}$$

avec une égalité stricte si et seulement si

$$\boldsymbol{\lambda}^T \mathbf{x}_i = a_i b_i^{-1} = \left(\pi_i^{-1} - 1\right) \sigma_i^2$$

pour certains vecteurs $\boldsymbol{\lambda}$.

Bibliographie

Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4), 831-846.

Brewer, K.R.W. (1999). Le calage esthétique dans le cas de l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 25, 2, 231-239. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4883-fra.pdf.

Brewer, K.R.W., Hanif, M. et Tam, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.

Chambers, R., et Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press: Oxford.

Godambe, V.P., et Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

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

Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40(3), 202-204.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. Brooks-Cole: Boston, 2e édition.

Nedyalkova, D., et Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95(3), 521-537.

Nedyalkova, D., et Tillé, Y. (2012). Bias-robustness and efficiency of model-based inference in survey sampling. *Statistica Sinica*, 22(2), 777-794.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-387.

Royall, R.M. (1992). Robustesse et optimalité de plan dans des modèles de prédiction pour populations finies. *Techniques d'enquête*, 18, 2, 193-199. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1992002/article/14488-fra.pdf.

Royall, R.M., et Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68(344), 880-889.

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

Scott, A.J., Brewer, K.R.W. et Ho, E.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73(362), 359-361.

Steel, D.G., et Clark, R.G. (2014). Gains possibles lors de l'utilisation de l'information sur les coûts au niveau de l'unité dans un cadre assisté par modèle. *Techniques d'enquête*, 40, 2, 257-269. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14110-fra.pdf.

Tam, S.M. (1988). Asymptotically design-unbiased predictors in survey sampling. *Biometrika*, 75(1), 175-177.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Tillé, Y., et Matei, A. (2016). *Sampling: Survey Sampling*. R package version 2.8. https://CRAN.R-project.org/package=sampling.

Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters*, 63(1), 1-3.

Valliant, R., Dever, J.A. et Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

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

Welsh, A.H., et Wiens, D.P. (2013). Robust model-based sampling designs. *Statistics and Computing*, 23(6), 689-701, Novembre. https://doi.org/10.1007/s11222-012-9339-3.