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

Techniques d'enquête

Efficacité relative des méthodes fondées sur l'estimation par régression d'enquête assistée par un modèle : une étude par simulations

par Erin R. Lundy et J.N.K. Rao

Date de diffusion : le 21 juin 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2022

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2022
Vol. 48, nº 1, pp. 53-78
Statistique Canada, nº 12-001-X au catalogue

Efficacité relative des méthodes fondées sur l'estimation par régression d'enquête assistée par un modèle : une étude par simulations

Erin R. Lundy et J.N.K. Rao[1]

Résumé

L'utilisation de données auxiliaires pour améliorer l'efficacité des estimateurs de totaux et de moyennes grâce à l'estimation par régression d'enquête assistée par un modèle a suscité un grand intérêt au cours de ces dernières années. Les estimateurs par la régression généralisée (ERG), basés sur un modèle de régression linéaire de travail, sont actuellement utilisés dans les enquêtes auprès d'établissements au sein de Statistique Canada et dans plusieurs autres organismes statistiques. Les estimateurs ERG utilisent des poids d'enquête communs à toutes les variables de l'étude et tiennent compte des totaux de population connus des variables auxiliaires. De plus en plus, de nombreuses variables auxiliaires sont disponibles, dont certaines peuvent être extérieures. Cela donne lieu des poids d'ERG fluctuants lorsque toutes les variables auxiliaires disponibles, y compris les interactions entre les variables catégorielles, sont utilisées dans le modèle de régression linéaire de travail. Par ailleurs, de nouvelles méthodes d'apprentissage automatique, comme les arbres de régression et les méthodes LASSO, sélectionnent automatiquement les variables auxiliaires significatives et donnent lieu à des poids non négatifs stables et à des gains d'efficacité possibles par rapport aux estimateurs ERG. Dans le présent article, une étude par simulations, basée sur un ensemble de données réelles d'une enquête réalisée auprès d'entreprises considérées comme la population cible, est menée pour étudier la performance relative des estimateurs ERG, des arbres de régression et des méthodes LASSO en matière d'efficacité des estimateurs et de propriétés des poids de régression connexes. Des scénarios d'échantillonnage probabiliste et d'échantillonnage non probabiliste sont étudiés.

Mots-clés : Inférence assistée par modèle; estimation par calage; sélection de modèles; estimateur par régression généralisée.

1. Introduction

Statistique Canada et plusieurs autres organismes statistiques cherchent de plus en plus à mettre à profit des données auxiliaires, provenant éventuellement de sources administratives, pour améliorer l'efficacité des estimateurs. Les techniques d'apprentissage automatique sont désormais largement utilisées dans différentes disciplines pour utiliser ces renseignements auxiliaires. Ces méthodes ne nécessitent souvent pas d'hypothèses de distribution des méthodes plus traditionnelles et peuvent s'adapter à des relations non linéaires et non additives complexes entre les résultats et les variables auxiliaires. Les méthodes d'apprentissage automatique ont été appliquées aux données d'enquête dans divers contextes, comme les plans réactifs ou les plans adaptatifs, le traitement des données, la correction de non-réponse et la pondération (Buskirk, Kirchner, Eck et Signorino, 2018; Kern, Klausch et Kreuter, 2019).

Récemment, l'utilisation de techniques d'apprentissage automatique pour améliorer l'efficacité des estimateurs de totaux et de moyennes à l'aide d'une estimation par régression d'enquête assistée par un modèle selon un échantillonnage probabiliste a été envisagée. Les estimateurs par la régression des totaux de population finie assistés par un modèle peuvent réduire la variabilité et donner lieu à des gains

1. Erin R. Lundy, Division des méthodes d'intégration statistique, Statistique Canada, Ottawa, Ontario, K1A 0T6. Courriel : erin.lundy@statcan.gc.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

d'efficacité importants si les variables auxiliaires disponibles sont étroitement associées à la variable d'intérêt de l'enquête. De plus en plus, de nombreuses variables auxiliaires sont disponibles, dont certaines peuvent être extérieures. Dans ce cas, la sélection des variables suivie d'une estimation par régression assistée par le modèle sélectionné peut améliorer l'efficacité des estimateurs par régression des totaux de population finie de l'enquête. Nous prenons en considération l'estimation de la population finie en utilisant l'estimateur par la régression généralisée (ERG) avec différents modèles linéaires de travail (Särndal, Swensson et Wretman, 1992). Des estimateurs assistés par un modèle, fondés sur les méthodes LASSO et LASSO adaptatif (McConville, Breidt, Lee et Moisen, 2017) et les arbres de régression (McConville et Toth, 2019), ont été appliqués aux données d'enquête. D'autres modèles non linéaires, comme les splines pénalisées et les réseaux neuronaux, ont été étudiés pour l'estimation assistée par un modèle. Voir Breidt et Opsomer (2017) pour consulter une étude de ces techniques.

Un autre champ de la recherche pour lequel l'utilisation d'estimateurs assistés par un modèle a été proposée est l'estimation à partir d'échantillons non probabilistes. L'augmentation des coûts et la diminution des taux de réponse suscitent un intérêt croissant pour l'utilisation d'échantillons non probabilistes. Le processus de génération d'un échantillon non probabiliste n'est toutefois pas connu et de tels échantillons sont assujettis à des biais d'échantillonnage. Deux approches couramment utilisées pour l'estimation à partir d'échantillons non probabilistes sont la quasi-randomisation et la modélisation de superpopulation. Dans la première, l'échantillon est considéré comme s'il était obtenu à partir d'un échantillonnage probabiliste, mais avec des probabilités de sélection inconnues. Les probabilités de pseudo-inclusion sont estimées grâce à un modèle de propension dans lequel les données de l'échantillon sont utilisées en combinaison avec un ensemble de données externes qui porte sur la population ciblée. Des techniques d'apprentissage automatique ont été employées dans l'estimation des probabilités de pseudo-inclusion ou de la même manière, dans la construction des pseudo-poids. Kern, Li et Wang (2020) ont étudié plusieurs techniques d'apprentissage automatique pour construire des pseudo-poids au moyen d'une pondération par noyau basée sur le score de propension pour des échantillons non probabilistes. Rafei, Flannagan et Elliott (2020) ont mis au point une approche de pseudo-pondération qui repose sur des arbres de régression additifs bayésiens.

Dans l'approche par superpopulation, les valeurs observées des variables d'intérêt sont supposées être générées par un certain modèle. Le modèle est estimé à partir des données et, avec des données externes de contrôle de la population, est utilisé pour projeter l'échantillon sur la population. Dans ce cadre, le calage sur des totaux de population connus de variables auxiliaires offre un moyen de réduire potentiellement l'effet du biais d'échantillonnage. Chen, Valliant et Elliott (2018) ont discuté de la mise en œuvre du calage du modèle en utilisant la méthode de type LASSO adaptatif pour les données basées sur un échantillonnage non probabiliste. Dans les scénarios où les totaux de la population sont estimés, Chen, Valliant et Elliott (2019) ont incorporé l'incertitude d'échantillonnage des données de référence, obtenues à partir d'une enquête par échantillonnage probabiliste, dans la composante de variance d'un estimateur par calage assisté par un modèle fondé sur une régression de type LASSO adaptatif. Par conséquent, contrairement à l'échantillonnage probabiliste où l'utilisation de l'estimation assistée par un

modèle vise à améliorer l'efficacité des estimateurs, l'utilisation de ces techniques dans un contexte d'échantillonnage non probabiliste a pour but de limiter l'incidence du biais d'échantillonnage.

Nous considérons plusieurs estimateurs fondés sur la méthode LASSO ainsi qu'un estimateur par arbre de régression et nous évaluons leur performance dans un contexte d'échantillonnage probabiliste et dans un contexte d'échantillonnage non probabiliste. Dans la section 2, les estimateurs assistés par un modèle considérés sont étudiés. La mise en place d'une étude par simulations selon un échantillonnage probabiliste est décrite dans la section 3. Les résultats de l'étude par simulations sur l'erreur quadratique moyenne des estimateurs, le biais relatif des estimateurs de variance et les propriétés des poids d'enquête sont présentés dans la section 4. À l'exception de l'estimateur ERG, tous les estimateurs assistés par un modèle considérés dans cette section comportent une sélection de variables et produisent, le cas échéant, des poids de régression qui dépendent de la variable d'intérêt de l'enquête, y. L'incidence de l'utilisation d'un seul ensemble de poids de régression pour de multiples variables de l'étude connexes est également étudiée dans cette section. Les résultats de l'étude par simulations reposant sur un scénario d'échantillonnage non probabiliste sont détaillés dans la section 5. Pour conclure, nous résumons nos conclusions dans la section 6.

2. Estimation assistée par un modèle selon un échantillonnage probabiliste

2.1 Estimateurs par la régression généralisée

Considérons l'estimation d'un total de population finie $t_y = \sum_{i \in U} y_i$, où $U = \{1, \ldots, N\}$ correspond à l'ensemble des unités de la population finie et y_i est la valeur de la variable d'intérêt de l'enquête pour l'unité $i \in U$. Soit $s \subset U$ un échantillon sélectionné selon un plan de sondage $p(.)$, où $p(s)$ correspond à la probabilité de sélectionner s. Pour $i \in U$, on désigne par $\pi_i = \Pr[i \in s]$ les probabilités d'inclusion de premier ordre du plan. Nous supposons $\pi_i > 0$ pour tous les $i \in U$. De plus, on suppose que d variables auxiliaires, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ sont connues pour chaque $i \in U$. Une méthode standard consiste à utiliser l'estimateur de Horvitz-Thompson (HT) :

$$\hat{t}_{y,\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i$$

où $d_i = \pi_i^{-1}$ désigne les poids de sondage. Dans ce cadre strictement basé sur le plan, les données auxiliaires n'ont pas d'incidence sur la forme de l'estimateur, mais elles peuvent avoir une incidence sur les poids de sondage, d_i, par la spécification du plan de sondage.

Une stratégie pour utiliser les données auxiliaires dans l'estimation consiste à employer un estimateur assisté par un modèle de t_y en spécifiant un modèle de travail pour la moyenne de y étant donné $\mathbf{x}$ et à utiliser ce modèle pour prédire les valeurs de y. La spécification d'un modèle de régression linéaire conduit à l'estimateur par la régression généralisée (ERG) (Cassel, Särndal et Wretman, 1976).

L'estimateur ERG se caractérise généralement par une variance plus faible que l'estimateur de Horvitz-Thompson si le modèle de travail a un certain pouvoir de prédiction pour y. Ici, nous considérons l'estimateur ERG sous un modèle de régression linéaire de travail

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{2.1}$$

avec $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$, ε_i étant indépendant et identiquement distribué avec une moyenne de zéro et une variance de σ^2 et $\mathbf{x}_i = \left(1, x_{i1}, \ldots, x_{ip}\right)^T$. L'estimateur ERG est obtenu au moyen de :

$$\hat{t}_{y,\mathrm{ERG}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s \tag{2.2}$$

avec les coefficients de régression $\boldsymbol{\beta}$ estimés

$$\hat{\boldsymbol{\beta}}_s = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right) = \left(\mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s, \tag{2.3}$$

où $\mathbf{X}_s$ est une matrice $n \times \left(p+1\right)$, Y_s est un vecteur n et Π_s est une matrice diagonale $n \times n$ de probabilités d'inclusion de premier ordre pour les unités échantillonnées.

L'estimateur ERG peut également s'écrire sous la forme d'une somme pondérée de la variable d'intérêt, y, ce qui donne des poids de régression qui sont indépendants de y et, par conséquent, peuvent être appliqués à toute variable de l'étude, y :

$$\hat{t}_{y,\mathrm{ERG}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x,\mathrm{HT}}\right)^T \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^T d_k \right)^{-1} \mathbf{x}_i \right] d_i y_i = \sum_{i \in s} w_i y_i \tag{2.4}$$

où $\mathbf{t}_x$ est le vecteur du total de population connu des covariables $\mathbf{x}$ et $\hat{\mathbf{t}}_{x,\mathrm{HT}}$ est le vecteur de l'estimateur de Horvitz-Thompson des totaux de population de covariables $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$. Les poids de régression, w_i, sont appelés « poids de calage », car ils satisfont à la contrainte de calage $\sum_{i \in s} w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$. Le poids de calage w_i ne dépend pas de la variable de l'étude y_i. Notons que l'estimateur ERG (2.4) peut être exprimé de manière alternative comme suit :

$$\hat{t}_{y,\mathrm{ERG}} = \hat{t}_{y,\mathrm{HT}} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x,\mathrm{HT}}\right)^T \hat{\boldsymbol{\beta}}_s$$

qui ne nécessite que des totaux de population connus $\mathbf{t}_x$. Pour l'estimateur ERG, les valeurs individuelles de la population $\mathbf{x}_i$, $i \in U$ ne sont pas nécessaires.

Si une procédure de sélection des variables, comme une procédure de sélection par étape ascendante, est mise en œuvre avant l'ajustement du modèle de régression linéaire, les poids de calage dépendent de y étant donné que les modèles sélectionnés peuvent varier selon les variables de l'étude. Ce type d'estimateur par la régression d'enquête avec procédure de sélection par étape est calé sur les variables auxiliaires sélectionnées par la procédure de sélection de variables pour une variable d'intérêt particulière, y.

L'utilisation d'un modèle de régression linéaire de travail comportant de nombreuses variables auxiliaires, y compris des interactions des variables auxiliaires catégorielles, peut produire des poids très

variables et gonfler considérablement la variance de l'estimateur ERG. De plus, certains poids de régression, w_i, $i \in s$, peuvent être négatifs, ce qui fait perdre l'interprétation d'un poids en tant que nombre d'unités de population représentées par l'unité échantillonnée.

2.2 Estimateur par la régression d'enquête selon la méthode LASSO

Si le modèle de régression linéaire dans (2.1) est épars, c'est-à-dire que p est grand, et que, supposons, seuls p_0 des coefficients de régression p ne sont pas nuls, alors l'estimation des coefficients nuls (2.3) entraîne une variation supplémentaire dans l'estimateur ERG (2.2). Dans ce cas, la sélection du modèle pour supprimer les variables externes pourrait réduire la variance sous le plan global de l'estimateur ERG, ce qui produirait des estimations plus efficaces des totaux de population finie. La méthode *Least Absolute Shrinkage and Selection Operator* (LASSO), élaborée par Tibshirani (1996), permet d'effectuer simultanément la sélection du modèle et l'estimation du coefficient en réduisant certains coefficients de régression à zéro. La méthode LASSO permet d'estimer les coefficients en minimisant la somme des résidus carrés sous contrainte de pénalité sur la somme de la valeur absolue des coefficients de régression.

McConville et coll. (2017) ont proposé d'utiliser des coefficients de régression estimés selon la méthode LASSO par poids d'enquête calculés de la manière suivante :

$$\hat{\boldsymbol{\beta}}_{s,L} = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

où $\lambda \geq 0$. L'estimateur par la régression d'enquête de type LASSO pour le total t_y est alors donné par :

$$\hat{t}_{y,\,\mathrm{LASSO}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,L}}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,L}.$$

La valeur du paramètre de pénalité λ doit être choisie avant d'obtenir les coefficients estimés. En général, ce processus de spécification des hyperparamètres avant l'ajustement du modèle final est appelé réglage des hyperparamètres. Plusieurs critères de sélection peuvent être utilisés pour choisir la valeur des hyperparamètres, notamment le critère d'information d'Akaike (AIC), le critère d'information bayésien (BIC) ou la validation croisée. Nous avons utilisé une version de la validation croisée qui incorpore les poids de sondage dans notre étude par simulations; voir McConville (2011) pour une analyse sur la sélection du paramètre de pénalité pour les estimations du coefficient selon la méthode LASSO pondérées par des poids d'enquête.

2.3 Estimateur par régression d'enquête selon la méthode LASSO adaptatif

En réduisant les coefficients de régression à zéro, l'utilisation du critère LASSO présente l'inconvénient de produire des estimations biaisées pour les coefficients de régression qui sont loin de zéro. Selon le critère LASSO adaptatif (Zou, 2006), les coefficients de la pénalité l_1 sont pondérés par

l'inverse d'un estimateur convergent n de $\boldsymbol{\beta}$. Par conséquent, le biais pour les grands coefficients est généralement plus petit.

McConville et coll. (2017) ont étudié un estimateur par régression d'enquête de type LASSO adaptatif (ALASSO) :

$$\hat{t}_{y,\,\mathrm{ALASSO}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}}}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}},$$

où

$$\hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}} = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \frac{\left|\beta_j\right|}{\left|\hat{\beta}_{sj}\right|}$$

et $\hat{\boldsymbol{\beta}}_s$ est donné par (2.3). La dépendance de la méthode LASSO adaptatif à l'égard des estimations des coefficients de régression linéaire pondérés standard, $\hat{\boldsymbol{\beta}}_s$, entraîne une perte d'efficacité lorsque p est important, car les estimations $\hat{\boldsymbol{\beta}}_s$ sont généralement très instables.

2.4 Estimateurs par calage de la méthode LASSO

Les méthodes LASSO et LASSO adaptatif ne produisent pas directement des poids de régression, car les estimateurs ne peuvent pas être exprimés comme des combinaisons pondérées des valeurs y. McConfille et coll. (2017) ont mis au point des poids de régression d'enquête LASSO en utilisant une approche de calage par modèle et une approximation de régression généralisée non linéaire. Ces poids de régression LASSO dépendent de la variable d'intérêt, y.

L'estimateur de calage de type LASSO (CLASSO) est calculé en régressant la variable d'intérêt, y_i, sur un point d'intersection et la fonction moyenne ajustée par LASSO $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,L}$. L'estimateur par calage de type LASSO peut s'écrire sous la même forme que (2.4), où $\mathbf{x}_i$ est remplacé par $\mathbf{x}_i^* = \left(1, \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,L}\right)^T$:

$$\hat{t}_{y,\,\mathrm{CLASSO}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_{x^*} - \hat{\mathbf{t}}_{x^*,\,\mathrm{HT}}\right)^T \left(\sum_{k \in s} \mathbf{x}_k^* \mathbf{x}_k^{*T} d_k \right)^{-1} \mathbf{x}_i^* \right] d_i y_i. \tag{2.5}$$

De même, l'estimateur par calage de type LASSO adaptatif (CALASSO) est donné par :

$$\hat{t}_{y,\,\mathrm{CALASSO}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_{x^{**}} - \hat{\mathbf{t}}_{x^{**},\,\mathrm{HT}}\right)^T \left(\sum_{k \in s} \mathbf{x}_k^{**} \mathbf{x}_k^{**T} d_k \right)^{-1} \mathbf{x}_i^{**} \right] d_i y_i,$$

où la moyenne ajustée au LASSO pour $\mathbf{x}_i^*$ dans (2.5) est remplacée par l'ajustement de type LASSO adaptatif, $\mathbf{x}_i^{**} = \left(1, \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}}\right)^T$. Les poids pour les estimateurs par calage de type LASSO sont calés sur la taille de la population N et sur le total de la population des fonctions moyennes ajustées par la méthode LASSO.

2.5 Estimateur par arbre de régression

L'estimateur ERG peut aussi être exprimé comme suit :

$$\hat{t}_{y,r} = \sum_{i \in s} \frac{y_i - \hat{h}_n(\mathbf{x}_i)}{\pi_i} + \sum_{i \in U} \hat{h}_n(\mathbf{x}_i), \tag{2.6}$$

où $\hat{h}_n(\mathbf{x}_i)$ est un estimateur de la fonction moyenne de Y_i étant donné $\mathbf{X}_i = \mathbf{x}_i$, $h(\mathbf{x}_i) = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i)$, sur la base des données de l'échantillon $(y_i, \mathbf{x}_i), i \in s.$ Comme solution de rechange à un modèle de régression linéaire, McConville et Toth (2019) ont proposé d'estimer $h(\mathbf{x})$ au moyen d'un modèle d'arbre de régression fondé sur l'algorithme suivant :

1. Où $k(n)$ est la taille minimale de la case et α correspond à un niveau de signification donné.
2. Si l'ensemble de données contient au moins $2k(n)$ observations, passer à l'étape 3; sinon, arrêter.
3. Parmi les variables auxiliaires $x_l, l = 1, \ldots, d,$ choisir une variable pour diviser les données. La variable choisie x_l est celle qui présente la différence la plus significative après avoir testé l'hypothèse nulle d'homogénéité de $E[y \mid x_l]$. Si aucune variable ne conduit à une différence significative, arrêter.
4. Diviser les données en deux ensembles, S_L et S_R, en les séparant en fonction de la valeur de la variable x_l sélectionnée qui entraîne la plus grande diminution de l'erreur quadratique moyenne estimée, tout en satisfaisant à l'exigence que chaque sous-ensemble contienne au moins $k(n)$ unités.
5. Pour chacun des sous-ensembles de données obtenus, retourner à l'étape 1.

Le modèle d'arbre de régression qui en résulte regroupe les catégories d'une variable auxiliaire en fonction de leur relation avec la variable d'intérêt et il n'inclut que les variables auxiliaires et les interactions associées à cette variable. Il est important de mentionner que l'inclusion d'une variable catégorique ne nécessite pas une division pour chaque catégorie, ce qui peut réduire considérablement la taille du modèle tout en saisissant quand même des interactions importantes.

Après l'ajustement d'un modèle d'arbre de régression, nous obtenons un ensemble de cases $Q_n = \{B_{n1}, B_{n2}, \ldots, B_{nq}\}$ qui partitionnent les données. Soit $I(\mathbf{x}_i \in B_{nk}) = 1$ si $\mathbf{x}_i \in B_{nk}$ et 0 autrement, pour $k = 1, \ldots, q.$ Cela signifie que $I(\mathbf{x}_i \in B_{nk}) = 1$ pour exactement une case $B_{nk} \in Q_n$ pour chaque $i \in s.$ Pour chaque $\mathbf{x}_i \in B_{nk}$, l'estimateur de $h(\mathbf{x}_i)$ est donné par :

$$\tilde{h}_n(\mathbf{x}_i) = \tilde{\#}(B_{nk})^{-1} \sum_{i \in s} \pi_i^{-1} y_i I(\mathbf{x}_i \in B_{nk}) = \tilde{\mu}_{nk}, \tag{2.7}$$

où

$$\tilde{\#}(B_{nk}) = \sum_{i \in s} \pi_i^{-1} I(\mathbf{x}_i \in B_{nk})$$

est l'estimateur de HT de la taille de la population dans la case B_{nk}. On obtient l'estimateur par arbre de régression $\hat{t}_{y,\,\text{ARBRE}}$ en insérant l'équation (2.7) dans l'estimateur par la régression généralisée, donné dans l'équation (2.6), ce qui donne l'estimateur poststratifié :

$$\hat{t}_{y,\,\text{ARBRE}} = \sum_k N_k \tilde{\mu}_{nk},$$

où N_k est le nombre d'unités dans U qui appartiennent à la case k.

Puisque $\tilde{h}_n(\mathbf{x}_i)$ peut s'écrire sous la forme d'un estimateur par la régression linéaire avec q covariables de fonction indicatrice, l'estimateur par arbre de régression est aussi un estimateur poststratifié, où chaque case B_{nk} représente une post-strate. Cela signifie que cet estimateur est calé sur le total de population de chaque case et fournit un mécanisme axé sur les données, dépendant de y, pour la sélection des post-strates et qui garantit qu'aucune d'entre elles n'est vide. Ainsi, nous sommes assurés que les poids de régression ne sont pas négatifs. Les poids produits par cette procédure d'estimation dépendent de la variable d'intérêt y. Par conséquent, contrairement à l'approche de l'estimateur ERG, il n'existe pas d'ensemble unique de poids génériques qui sont appliqués à toutes les variables de l'étude. Au lieu de cela, un ensemble de poids est produit pour chaque variable d'intérêt de l'étude.

2.6 Estimation de la variance dans le cadre d'un échantillonnage aléatoire simple stratifié

Dans le cadre d'un échantillonnage aléatoire simple stratifié, un estimateur de la variance des estimateurs par régression d'enquête assistés par un modèle décrit ci-dessus est obtenu par la méthode de linéarisation de Taylor et est donné par :

$$\hat{V}\left(\hat{t}_y\right) = \sum_h \frac{N_h\left(N_h - n_h\right)}{n_h} \frac{1}{n_h - 1} \sum_{i \in s_h} \left(e_{hi} - \bar{e}_h\right)^2, \tag{2.8}$$

où h indexe les strates, N_h est le nombre d'unités de population dans la strate h, n_h est le nombre d'unités s_h échantillonnées dans la strate h, $e_{hi} = y_{hi} - \hat{h}_n(\mathbf{x}_{hi})$ est le résidu de l'unité d'échantillonnage i dans la strate h sous le modèle de régression et $\bar{e}_h$ est le résidu moyen dans la strate h.

Les estimateurs de variance peuvent facilement être aussi utilisés pour des plans de sondage plus complexes, mais pour des raisons de simplicité, nous n'avons utilisé l'expression que pour l'échantillonnage aléatoire simple stratifié qui est utilisé dans l'étude par simulations de la section 3.

3. Étude par simulations fondée sur les données de l'Enquête sur le financement et la croissance des petites et moyennes entreprises

Nous allons maintenant décrire une étude par simulations utilisée pour comparer la performance des estimateurs par la régression d'une enquête assistés par un modèle à celles de l'estimateur de HT fondé

uniquement sur le plan de sondage. En prenant les données de l'Enquête sur le financement et la croissance des petites et moyennes entreprises comme population, nous comparons les estimateurs dans des échantillons répétés des données pour produire des estimations du montant total demandé pour un crédit commercial, qui est un type particulier de financement.

3.1 Population de simulation

L'Enquête sur le financement et la croissance des petites et moyennes entreprises (EFCPME) est une enquête-entreprise périodique qui a lieu environ tous les trois ans et qui vise à recueillir des renseignements sur les types de financement auxquels les entreprises ont recours. L'échantillon est stratifié par taille et défini par le nombre d'employés, l'âge de l'entreprise, l'industrie en fonction du code à deux chiffres du Système de classification des industries de l'Amérique du Nord et l'emplacement géographique. Un échantillon d'environ 17 000 entreprises a été sélectionné pour la version de 2017 de l'enquête.

Le Registre des entreprises (RE) est la principale source d'information auxiliaire pour les enquêtes-entreprises à Statistique Canada. La base de sondage sur laquelle repose l'EFCPME a été créée en sélectionnant dans le RE de Statistique Canada toutes les entreprises ayant entre 1 et 499 employés et un revenu brut minimum de 30 000 $. Les entreprises sans but lucratif ainsi que les entreprises appartenant à certains sous-groupes industriels ont été exclues de la population cible. Le RE contient des renseignements sur l'emplacement, le nombre d'employés, le secteur d'activité ainsi que le chiffre d'affaires de chaque entreprise de la population.

3.2 Méthodologie de la simulation

Nous avons mené une étude par simulations pour comparer la performance relative de plusieurs estimateurs par la régression d'une enquête assistés par un modèle, en utilisant trois et quatre variables auxiliaires catégoriques. Nous avons examiné des tailles d'échantillon de $n = \{200; 500; 1\,000\}$ à partir des 9 115 répondants de l'ensemble de données de l'EFCPME. Cet ensemble de données formait la population cible, et des échantillons répétés ont été tirés en utilisant un échantillonnage aléatoire simple stratifié, car il s'agit du plan couramment utilisé par les organismes statistiques pour les enquêtes-entreprises. Nous avons supposé qu'il y avait deux strates : la strate A qui est composée d'unités dont le revenu est inférieur à 2,5 millions de dollars et la strate B qui est composée d'unités dont le revenu est supérieur à 2,5 millions de dollars. Nous avons supposé des tailles d'échantillon égales dans chaque strate, mais la plupart des unités de la population, environ 70 %, appartiennent à la strate A. Selon ce plan de sondage, les unités à revenu plus élevé sont surreprésentées, ce qui entraîne un plan de sondage à probabilités inégales. Des simulations préliminaires fondées sur un plan d'échantillonnage aléatoire simple ont également été envisagées et ont donné des résultats similaires. La taille minimale envisagée pour l'échantillon était $n = 200$, car pour des tailles d'échantillon plus petites et 28 catégories de variables

x, nous obtenions souvent des catégories sans unité échantillonnée. Dans ce cas, il n'est pas possible de calibrer l'estimateur ERG pour tous les totaux marginaux prédéfinis.

Pour chaque échantillon, des modèles reposant sur trois variables x, à savoir l'industrie (10 catégories), la taille de l'effectif (4 catégories) et la région (6 catégories), ont servi à estimer le montant total des crédits commerciaux demandés. Les résultats ont ensuite été comparés au total réel. Nous avons également examiné une quatrième variable, les revenus, laquelle comporte 8 catégories. Pour chaque combinaison des trois tailles d'échantillon différentes et des deux ensembles de variables auxiliaires, avec 20 et 28 catégories d'effets principaux, nous avons tiré 5 000 échantillons aléatoires stratifiés répétés de la population cible. Pour chaque échantillon, nous avons mis en œuvre l'estimateur de HT et plusieurs estimateurs d'enquête assistés par un modèle, comme le résume le tableau 3.1 ci-dessous :

Tableau 3.1
Résumé des estimateurs assistés par un modèle pris en compte dans l'étude par simulations

Estimateur	Données auxiliaires	Poids de régression	Totaux de calage
ERG	Totaux marginaux Seuls les principaux effets sont pris en compte	Indépendants de y	Toutes les variables auxiliaires
ERG avec sélection ascendante de variables (FSTEP)	Valeurs individuelles Seuls les principaux effets sont pris en compte	Dépendants de y	Variables auxiliaires sélectionnées
Arbre de régression (ARBRE)	Valeurs individuelles	Dépendants de y, strictement positifs	Taille de la population de chaque case
LASSO (LASSO)	Valeurs individuelles Prise en compte des effets principaux (unidirectionnel) et des interactions bidirectionnelles (bidirectionnel)		
LASSO calé (CLASSO)	Valeurs individuelles Prise en compte des effets principaux (unidirectionnel) et des interactions bidirectionnelles (bidirectionnel)	Dépendants de y	Taille de la population et fonction moyenne ajustée par la méthode LASSO
LASSO adaptatif (ALASSO)	Valeurs individuelles Seuls les principaux effets sont pris en compte		
LASSO adaptatif calé (CALASSO)	Valeurs individuelles Seuls les principaux effets sont pris en compte	Dépendants de y	Taille de la population et fonction moyenne ajustée par la méthode LASSO

Dans un premier temps, nous avons également envisagé l'estimateur fondé sur la méthode LASSO adaptatif et l'estimateur par calage de type LASSO adaptatif en utilisant tous les effets principaux et des interactions bidirectionnelles, mais les estimations des coefficients avec le modèle linéaire de l'estimateur ERG, $\hat{\boldsymbol{\beta}}_s$, étaient très instables, ce qui a entraîné des problèmes de singularité.

Tous les calculs ont été effectués avec R (version 3.4.0, 2017). Les estimateurs de HT, ERG, par arbre de régression et LASSO ont été calculés à l'aide du progiciel **mase** (McConville, Tang, Zhu, Li, Cheung et Toth, 2018) et les coefficients LASSO adaptatifs ont été calculés à l'aide du progiciel **glmnet** (Friedman, Hastie, Simon, Qian et Tibshirani, 2017). La fonction `cv.glmnet` a été utilisée pour sélectionner la valeur du paramètre de pénalité pour les estimateurs LASSO. Nous avons utilisé une procédure de validation croisée à 10 reprises qui permet d'inclure les poids de sondage. Pour l'estimateur par arbre de régression, la taille minimale de la case $k(n)$ était de 25 et le niveau de signification α était de 0,05. Nous avons également envisagé une taille minimale de case de 10 unités. Pour les petites tailles d'échantillon, il y avait un petit gain d'efficacité par rapport à une taille minimale de case de 25. Pour des tailles d'échantillon de $n = 1\,000$, différentes tailles minimales de case ont donné des résultats similaires en matière d'erreur quadratique moyenne. La sélection par étape ascendante pour l'estimateur FSTEP était basée sur la réduction du critère d'information d'Akaike (AIC) et a été effectuée à l'aide de la fonction `stepAIC` du progiciel MASS (Ripley, Venables, Bates, Hornik, Gebhardt et Firth, 2017).

En évaluant par régression le montant du crédit commercial demandé pour l'ensemble de la population finie avec les 28 catégories marginales, le coefficient de détermination ajusté était d'environ $R^2 = 0{,}22$ lorsque les effets principaux et les effets d'interaction bidirectionnels étaient pris en compte. Pour le modèle de population avec effets principaux seulement, le nombre d'effets significatifs était de 15, et pour le modèle de population avec effets principaux et interactions bidirectionnelles, il y avait 2 effets principaux significatifs et 29 effets d'interaction significatifs. Ces résultats à l'échelle de la population indiquent que les modèles prédictifs utiles seront probablement peu nombreux et qu'il peut y avoir d'importantes interactions bidirectionnelles.

L'ajustement des modèles d'arbre de régression en fonction du montant du crédit commercial demandé a donné lieu à 25 divisions. La première division était basée sur le chiffre d'affaires, indiquant qu'il s'agit de la donnée auxiliaire la plus fortement liée au montant du crédit commercial demandé. Des divisions se sont produites en fonction des quatre variables auxiliaires examinées : le revenu, l'industrie, le nombre d'employés et l'emplacement. Ce résultat est cohérent avec les conclusions selon lesquelles les modèles prédictifs utiles devraient être peu nombreux, mais permettre des interactions d'ordre supérieur.

4. Résultats de l'étude par simulations

4.1 Performance des estimateurs pour ce qui est de l'erreur quadratique moyenne du plan de sondage

Nous avons calculé le biais du plan et l'erreur quadratique moyenne (EQM) du plan à partir des 5 000 estimations totales par taille d'échantillon et par nombre de catégories marginales. Le pourcentage du biais relatif absolu du plan était inférieur à 2 % pour tous les estimateurs et pour tous les scénarios. Comme prévu, pour tous les estimateurs, le biais diminue à mesure que la taille de l'échantillon augmente.

La figure 4.1 présente l'EQM des estimateurs HT, ERG, ERG avec sélection ascendante de variables, de l'estimateur par arbre de régression et de l'estimateur par calage de type LASSO selon la taille d'échantillon, en se basant sur les 5 000 échantillons simulés. Les valeurs de l'EQM sont similaires pour les versions adaptative et non calée des estimateurs LASSO. Pour tous les estimateurs, la diminution de l'EQM est beaucoup plus prononcée de $n = 200$ à $n = 500$ que de $n = 500$ à $n = 1\,000$. Ce phénomène est probablement attribuable à la petite taille de l'échantillon par rapport au nombre de catégories pour les variables auxiliaires. Il n'est peut-être pas possible d'étudier tous les effets potentiels, en particulier les effets d'ordre supérieur, alors que seulement 200 unités sont échantillonnées.

Le tableau 4.1 présente le ratio entre l'EQM du plan de chaque estimateur et l'EQM de l'estimateur de HT pour le montant total du crédit commercial demandé. Pour $n = 200$, l'estimateur par arbre de régression et l'estimateur fondé sur la méthode LASSO (bidirectionnel) avec effets d'interaction à deux facteurs sont les seuls estimateurs assistés par un modèle qui procurent des gains d'efficacité, par rapport à l'estimateur de HT, lorsque le nombre de catégories de variables auxiliaires utilisées est important. Quand la taille de l'échantillon augmente, les gains d'efficacité des estimateurs par la régression d'une enquête assistés par un modèle, par rapport à l'estimateur de HT, sont essentiellement égaux. L'utilisation de l'un des estimateurs assistés par un modèle lorsque $n = 1\,000$ entraîne un léger gain d'efficacité par rapport à l'estimateur de HT. Il y a peu de gains d'efficacité quand les estimateurs assistés par un modèle par rapport à l'estimateur de HT sont utilisés, ce qui indique que les variables auxiliaires ne sont pas fortement liées à la variable d'intérêt.

Figure 4.1 Comparaison de l'erreur quadratique moyenne de l'estimateur de HT, de l'estimateur ERG, de l'estimateur FSTEP, de l'estimateur arbre de régression et des estimateurs par calage de type LASSO (unidirectionnel et bidirectionnel) pour le montant total du crédit commercial demandé.

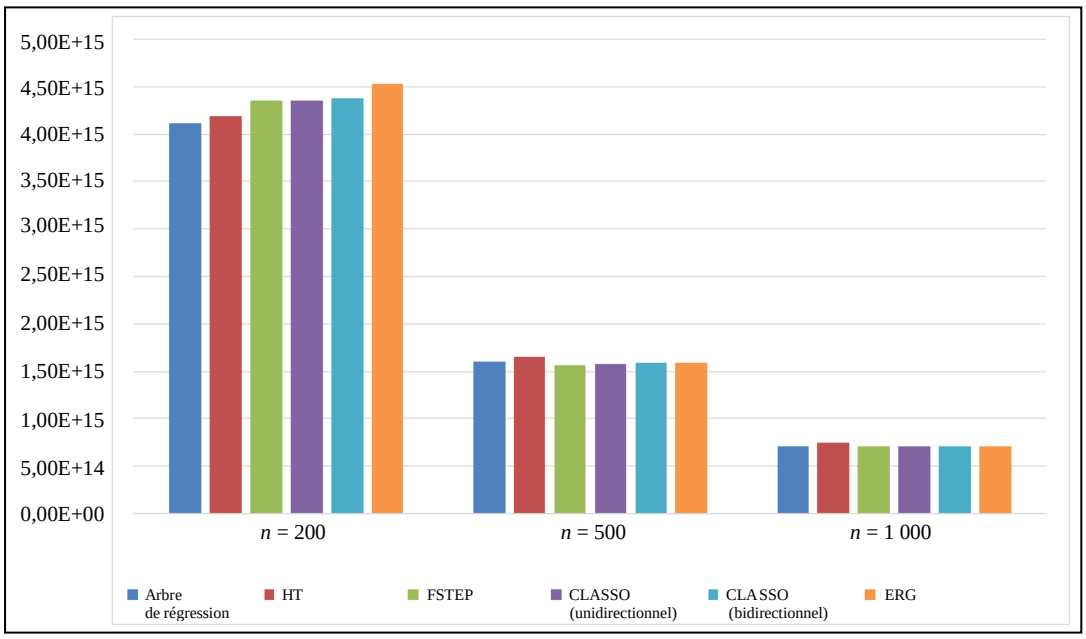

Tableau 4.1
Ratio de l'erreur quadratique moyenne (EQM) de chaque estimateur par rapport à l'EQM de l'estimateur de HT, avec 20 et 28 catégories marginales

	20 catégories			28 catégories		
	$n=200$	$n=500$	$n=1\,000$	$n=200$	$n=500$	$n=1\,000$
ERG	1,067	1,011	0,994	1,084	0,959	0,954
FSTEP	1,036	1,009	0,994	1,040	0,945	0,958
ARBRE	1,023	1,007	0,977	0,983	0,963	0,949
LASSO (unidirectionnel)	1,020	0,995	0,986	1,009	0,946	0,947
CLASSO (unidirectionnel)	1,047	1,004	0,990	1,042	0,952	0,949
LASSO (bidirectionnel)	0,999	0,995	0,952	0,981	0,935	0,936
CLASSO (bidirectionnel)	1,061	1,029	0,966	1,045	0,959	0,950
ALASSO	1,024	0,999	0,986	1,021	0,948	0,948
CALASSO	1,040	1,005	0,989	1,037	0,951	0,949

Les gains d'efficacité potentiels des estimateurs assistés par un modèle dépendent de la puissance prédictive du modèle de travail. Dans notre population de simulation, la force de la relation entre la variable d'intérêt et les variables auxiliaires disponibles est faible, ce qui n'entraîne que de légers gains d'efficacité par rapport à l'estimateur de HT purement fondé sur le plan. Par conséquent, pour examiner davantage les différences entre les différents estimateurs d'enquête assistés par un modèle, nous avons exécuté des simulations supplémentaires au moyen de différentes variables d'intérêt de l'enquête, générées selon les modèles suivants :

1. En supposant un modèle LASSO avec les effets principaux uniquement, nous avons obtenu les estimations du coefficient LASSO pour le montant du crédit commercial demandé, y_i, en utilisant les valeurs de population pour les variables auxiliaires $\mathbf{x}_i$, y compris le revenu.
2. Nous avons utilisé les estimations de coefficient $\hat{\beta}_L$ obtenues à l'étape 1 et les valeurs de population pour $\mathbf{x}_i$ afin de générer une nouvelle variable d'intérêt de l'enquête :

$$y_i^* = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_L + \mathbf{u}_i,$$

   où $\mathbf{u}_i$ est une variable aléatoire en distribution normale avec une moyenne de 0 et un écart-type σ choisi de telle sorte que le coefficient de détermination ajusté soit approximativement de $R^2 = 0{,}5$.
3. Nous avons tiré 5 000 échantillons répétés de la population cible et calculé l'erreur quadratique moyenne de chaque estimateur du t_{y^*} total.
4. Les étapes 1 à 3 ont été répétées en ajustant un modèle de régression LASSO avec des effets principaux et des interactions bidirectionnelles, et un modèle par arbre de régression en utilisant l'algorithme décrit dans la section 2.5.

Le tableau 4.2 présente le ratio entre l'EQM du plan de chaque estimateur et celle de l'estimateur de HT selon les trois différents modèles générant la variable d'intérêt de l'enquête pour une taille d'échantillon de $n = 1\,000$. Comme prévu, l'estimateur assisté par le modèle de travail correctement spécifié est le plus efficace. Dans le cas où le véritable modèle générateur ne contient que les effets

principaux, l'hypothèse d'un modèle de travail avec des interactions d'ordre supérieur entraîne une légère perte d'efficacité. En présence d'interactions bidirectionnelles ou d'ordre supérieur, les estimateurs par arbre de régression et fondés sur la méthode LASSO ajustés avec des interactions bidirectionnelles sont plus efficaces que les estimateurs assistés par un modèle basés sur des modèles de travail avec les effets principaux uniquement. Lorsque le modèle générateur est un arbre de régression, l'estimateur par arbre de régression offre des gains d'efficacité modestes par rapport aux estimateurs fondés sur la méthode LASSO bidirectionnels. Ces gains peuvent s'expliquer par le fait que le modèle par arbre de régression regroupe les catégories d'une variable auxiliaire en fonction de leur relation avec la variable d'intérêt et, par conséquent, réduit la taille du modèle. Dans tous les cas, nous obtenons des gains d'efficacité importants, par rapport à l'estimateur de HT fondé sur le plan de sondage.

Tableau 4.2
Ratio de l'erreur quadratique moyenne (EQM) de chaque estimateur par rapport à l'EQM de HT selon différents modèles générant une variable d'intérêt de l'enquête

	LASSO (unidirectionnel)	LASSO (bidirectionnel)	Arbre de régression
ERG	0,749	0,855	0,878
FSTEP	0,749	0,855	0,876
ARBRE	0,803	0,821	0,778
LASSO (unidirectionnel)	0,747	0,850	0,871
CLASSO (unidirectionnel)	0,747	0,851	0,873
LASSO (bidirectionnel)	0,763	0,761	0,826
CLASSO (bidirectionnel)	0,763	0,765	0,833
ALASSO	0,750	0,849	0,872
CALASSO	0,750	0,851	0,873

4.2 Performance dans d'autres scénarios

Nous avons également examiné la performance des estimateurs fondés sur la méthode LASSO et de l'estimateur par arbre de régression dans des scénarios où il n'y a pas d'effets principaux, mais seulement des interactions bidirectionnelles. Nous avons généré une quatrième variable d'intérêt de l'enquête en utilisant le modèle de régression fondé sur la méthode LASSO avec effets principaux et interactions bidirectionnelles, comme il décrit dans la procédure ci-dessus. Cependant, à l'étape 2, nous fixons toutes les estimations des coefficients correspondant aux effets principaux dont la valeur est égale à 0.

La première colonne du tableau 4.3 (appelée « absence de multicolinéarité ») montre le ratio entre l'EQM du plan de sondage des estimateurs et celui de l'estimateur de HT, où la variable d'enquête est générée à partir d'un modèle sans effets principaux pour des tailles d'échantillon de $n = 1\,000$. Dans ce scénario, les estimateurs fondés sur la méthode LASSO avec interactions bidirectionnelles et l'estimateur par arbre de régression sont nettement plus efficaces que les estimateurs assistés par un modèle fondés sur des modèles avec effets principaux seulement. Par rapport à l'estimateur ERG couramment utilisé, les gains d'efficacité des estimateurs fondés sur la méthode LASSO avec interactions bidirectionnelles et de l'estimateur par arbre de régression sont nettement plus élevés en l'absence d'effets principaux. Ce

résultat apparaît de façon évidente si nous comparons la colonne du LASSO bidirectionnel dans le tableau 4.2 à la première colonne du tableau 4.3. L'EQM relative est très semblable pour les estimateurs fondés sur la méthode LASSO bidirectionnel et l'estimateur par arbre de régression, mais plus près de 1 pour l'estimateur ERG et les estimateurs fondés sur méthode LASSO unidirectionnel.

Tableau 4.3
Ratio de l'erreur quadratique moyenne (EQM) de chaque estimateur par rapport à l'EQM de HT dans un modèle générateur sans effets principaux et en l'absence de multicolinéarité ou en présence de celle-ci

	Absence de multicolinéarité	Variable en double	Catégories regroupées
ERG	0,935	-	-
ARBRE	0,824	0,850	0,842
LASSO (unidirectionnel)	0,930	0,945	0,942
CLASSO (unidirectionnel)	0,936	0,953	0,951
LASSO (bidirectionnel)	0,783	0,795	0,773
CLASSO (bidirectionnel)	0,795	0,809	0,781

Pour les données administratives comportant de nombreuses variables, il n'est pas rare que certaines variables soient colinéaires ou presque colinéaires. Par exemple, il est fréquent que des renseignements sur le nombre total d'employés et sur le nombre d'employés équivalents temps plein soient disponibles. L'estimateur ERG, et par extension l'estimateur FSTEP et les estimateurs LASSO adaptatifs, échouent en présence de colinéarité, car la matrice de plan est unique. Nous avons étudié la performance de l'estimateur par arbre de régression et des estimateurs fondés sur la méthode LASSO en présence de multicolinéarité. Nous avons examiné deux types de multicolinéarité :

- Doublon d'une variable catégorique existante. Nous avons créé trois nouvelles variables indicatrices correspondant à la taille de l'effectif.
- Catégories regroupées de variables auxiliaires existantes : nous avons créé une nouvelle variable indicatrice correspondant aux trois catégories de revenus les plus élevées.

L'EQM, par rapport à l'estimateur de HT, pour $n = 1\,000$, est indiquée dans les colonnes 2 et 3 du tableau 4.3. Ces résultats sont très similaires à ceux de la première colonne du tableau 4.3 correspondant à l'absence de multicolinéarité. L'estimateur par arbre de régression et les estimateurs fondés sur la méthode LASSO fournissent un moyen automatique de supprimer les variables auxiliaires colinéaires sans avoir d'influence sur les gains d'efficacité potentiels. Il convient de mentionner que d'autres méthodes, comme l'analyse en composantes principales, peuvent être utilisées pour éliminer la colinéarité, mais elles nécessitent une certaine expertise.

4.3 Performance des estimateurs de variance en ce qui a trait au biais relatif

Des estimateurs de variance basés sur (2.8) ont été établis pour chaque estimateur. Le tableau 4.4 présente le biais relatif en pourcentage de chaque estimateur pour le montant total du crédit commercial demandé. À des fins de comparaison, l'estimateur de variance théoriquement sans biais de l'estimateur de

HT est compris dans ce tableau. Cet estimateur de variance équivaut à l'expression fournie dans (2.8) où $e_i = y_i - \bar{y}_s$. Les estimateurs de variance pour les estimateurs par la régression d'une enquête assistés par un modèle comportent un biais négatif important qui augmente avec le nombre de variables auxiliaires, p. L'ampleur du biais négatif est plus importante pour les estimateurs fondés sur la méthode LASSO ajustés avec des interactions bidirectionnelles. Pour les échantillons de petite taille, le biais négatif est plus faible pour l'estimateur par arbre de régression. De même, pour les échantillons de petite taille, il existe une différence substantielle dans les biais entre les estimateurs ERG et FSTEP. Le fait d'effectuer une sélection de variables avant de calculer l'estimateur ERG de calage standard semble réduire le biais de l'estimateur de variance dans ce cas. Le biais diminue pour tous les estimateurs par la régression d'une enquête assistés par un modèle à mesure que la taille de l'échantillon augmente.

Tableau 4.4
Biais relatif en pourcentage des estimateurs de variance

	20 catégories			28 catégories		
	$n = 200$	$n = 500$	$n = 1\,000$	$n = 200$	$n = 500$	$n = 1\,000$
ERG	-12,44	-4,16	-1,60	-22,23	-10,86	-6,99
FSTEP	-7,05	-3,60	-1,62	-14,07	-7,71	-6,73
ARBRE	-5,79	-5,53	-2,81	-8,45	-12,93	-10,83
LASSO (unidirectionnel)	-7,79	-2,96	-1,14	-12,42	-9,49	-6,44
CLASSO (unidirectionnel)	-10,08	-3,74	-1,61	-16,01	-9,84	-6,52
LASSO (bidirectionnel)	-11,94	-11,57	-7,62	-16,12	-15,14	-13,08
CLASSO (bidirectionnel)	-19,99	-15,09	-9,06	-25,87	-19,04	-15,14
ALASSO	-8,69	-3,61	-1,41	-14,52	-9,43	-6,38
CALASSO	-9,40	-3,78	-1,48	-15,80	-9,64	-6,46
HT	5,19	5,72	5,82	4,90	-0,11	1,66

Compte tenu du biais des estimateurs de variance examinés dans le cas présent, en particulier pour les échantillons de petite taille, la qualité de l'approximation du développement en série de Taylor du premier ordre peut être une source de préoccupation. Pour un grand nombre de variables auxiliaires catégoriques, les termes restants dans le développement en série de Taylor peuvent ne plus être négligeables pour les petites tailles d'échantillon. Un autre estimateur de variance pour les estimateurs fondés sur la méthode LASSO a été examiné par McConville et coll. (2017), mais ses résultats n'offraient que de légères améliorations en matière de réduction des biais. Une autre préoccupation est la prise en compte adéquate de la procédure intrinsèquement axée sur les données utilisée pour estimer l'arbre de régression et les modèles LASSO. Le modèle d'arbre de régression comporte des divisions, tandis que les modèles LASSO comportent un paramètre de pénalité, tous deux dépendant de l'échantillon.

4.4 Propriétés des poids d'enquête

Les poids de régression sont directement disponibles pour les estimateurs ERG, FSTEP, par arbre de régression, par calage de type LASSO (unidirectionnel et bidirectionnel) et par calage de type LASSO adaptatif. Nous avons étudié les propriétés des poids pour ces estimateurs dans nos simulations.

Une grande variation des valeurs des poids n'est pas souhaitable, car elle permet à certaines unités d'avoir beaucoup plus d'influence que d'autres. Les poids positifs sont privilégiés par les organismes statistiques nationaux, car un poids négatif ne permet plus d'interpréter le nombre d'unités de population représentées par l'unité échantillonnée.

Premièrement, nous avons calculé la moyenne, sur des échantillons répétés, de la variance empirique des poids de chaque échantillon :

$$\overline{\mathrm{var}}(\mathbf{w}) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n-1}\sum_{j\in s^{(r)}}\left(w_j^{(r)} - \overline{w}^{(r)}\right)^2,$$

où $s^{(r)}$ est l'échantillon simulé r^{e}, $\overline{w}^{(r)} = \frac{1}{n}\sum_{j\in s^{(r)}} w_j^{(r)}$ et $w_j^{(r)}$ représentent le poids de l'unité j^{e} dans l'échantillon simulé r^{e}. Nous avons également calculé le coefficient de variation (CV) moyen des poids :

$$\overline{\mathrm{CV}}(\mathbf{w}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\sqrt{\frac{1}{n-1}\sum_{j\in s^{(r)}}\left(w_j^{(r)} - \overline{w}^{(r)}\right)^2}}{\overline{w}^{(r)}}.$$

Le tableau 4.5 présente la variance moyenne et le CV moyen des poids pour l'ensemble des échantillons lorsque les revenus sont inclus comme variable auxiliaire. Les poids pour l'estimateur ERG et, dans une moindre mesure, l'estimateur FSTEP sont beaucoup plus variables que les poids pour l'estimateur par arbre de régression et les estimateurs fondés sur la méthode LASSO, en particulier pour les échantillons de petite taille. La variabilité des poids pour les trois approches LASSO est très similaire et est toujours légèrement inférieure à la variabilité des poids pour l'estimateur par arbre de régression.

Tableau 4.5
Variance moyenne (CV) des poids entre les échantillons

	$n = 200$	$n = 500$	$n = 1\,000$
ERG	728,18 (0,59)	77,14 (0,48)	16,41 (0,44)
FSTEP	462,81 (0,47)	67,45 (0,45)	15,90 (0,44)
ARBRE	374,43 (0,42)	59,35 (0,42)	14,70 (0,42)
CLASSO (unidirectionnel)	354,57 (0,41)	56,21 (0,41)	14,03 (0,41)
CLASSO (bidirectionnel)	361,83 (0,42)	56,60 (0,41)	14,06 (0,41)
CALASSO	354,29 (0,41)	56,28 (0,41)	14,03 (0,41)

Nous avons également calculé la proportion d'échantillons simulés lorsque les poids de régression contenaient des valeurs négatives. Comme mentionné dans la section 2.5, par construction, nous pouvons être assurés que les poids de l'estimateur par arbre de régression seront positifs. Lorsque la taille de l'échantillon était de 200, l'estimateur ERG calé sur 20 catégories marginales donnait des poids négatifs pour environ 3 % des échantillons répétés. Il n'y avait pas de poids négatifs lorsque la taille de l'échantillon était de 500 ou de 1 000. Pour l'estimateur ERG calé sur 28 catégories marginales, environ 27 % des échantillons répétés de taille 200 contenaient des poids négatifs, et moins de 0,5 % des échantillons répétés de taille 500 contenaient des poids négatifs. Les poids de l'estimateur ERG sont instables lorsque la taille de l'échantillon est petite, en particulier si l'estimateur ERG est calé sur des variables auxiliaires comportant un grand nombre de catégories. L'utilisation de la sélection par étape

ascendante des variables avec l'estimateur ERG a permis de réduire considérablement le nombre d'échantillons simulés avec des poids négatifs pour les échantillons de petite taille. L'estimateur FSTEP appliqué aux 28 catégories marginales a donné des poids négatifs dans environ 0,5 % des échantillons répétés avec une taille d'échantillon de 200. Aucun poids négatif n'a été observé pour l'estimateur par calage de type LASSO avec seulement des effets principaux ou l'estimateur par calage de type LASSO adaptatif. L'utilisation de l'estimateur par calage de type LASSO avec des interactions bidirectionnelles a donné lieu à des poids négatifs dans moins de 0,05 % des échantillons simulés.

4.5 Estimation basée sur un seul ensemble de poids

Un inconvénient majeur de la mise en œuvre des approches basées sur l'arbre de régression et par calage de type LASSO est que les procédures d'estimation produisent des poids propres aux variables. Nous avons effectué des simulations supplémentaires dans lesquelles un seul ensemble de poids propre à une variable était appliqué à d'autres variables d'intérêt connexes de l'enquête. Dans le contexte des données de notre enquête-entreprise, nous avons examiné quatre variables d'intérêt de l'enquête, le montant du crédit commercial demandé ainsi que le montant demandé pour trois autres types de financement : la marge de crédit, la carte de crédit commerciale et le financement par crédit-bail. Nous avons examiné l'incidence sur le biais et la perte d'efficacité de l'utilisation d'un seul ensemble de poids, déterminé par une variable primaire d'intérêt, pour estimer le montant total demandé pour les trois autres variables d'intérêt de l'enquête. Plus précisément, nous avons calculé le pourcentage du biais relatif absolu du plan pour les estimateurs du montant total demandé et les estimateurs de variance. Nous avons également calculé le ratio entre l'EQM pour l'arbre de régression et trois approches par calage de type LASSO en utilisant l'ensemble des poids correspondant à une variable primaire d'intérêt et l'EQM pour les estimateurs au moyen des poids propres aux variables. Par souci de concision, nous n'avons examiné que les contextes comportant 28 catégories marginales.

Le pourcentage du biais relatif absolu du plan était inférieur à 2 % pour tous les estimateurs et pour tous les scénarios. Pour tous les estimateurs et toutes les variables primaires d'intérêt, le biais diminue lorsque la taille de l'échantillon augmente.

Contrairement au biais des estimateurs de variance basés sur des poids propres aux variables, le biais des estimateurs de variance basés sur un seul ensemble de poids pour une variable primaire d'intérêt ne diminue pas nécessairement lorsque la taille de l'échantillon augmente. De même, le biais n'est pas strictement unidirectionnel et peut être positif ou négatif. Pour les approches basées sur l'arbre de régression et par calage de type LASSO, le biais des estimateurs de variance est considérablement plus important pour la variable primaire d'intérêt utilisée pour calculer l'ensemble unique de poids que pour les autres variables de l'étude. La nature axée sur les données de ces estimateurs fait en sorte que la variance estimée pour la variable primaire d'intérêt est sous-estimée, comme le montre le tableau 4.4.

Le tableau 4.6 présente le ratio entre l'EQM du plan de chaque estimateur dont les poids sont déterminés par une variable primaire d'intérêt, et l'EQM de l'estimateur dont les poids sont propres aux

variables, calculé séparément pour chacune des quatre variables de l'étude pour n égal à 200 et à 500. L'utilisation d'un ensemble unique de poids déterminés par une variable primaire d'intérêt donne un EQM similaire ou légèrement plus élevé que lorsque nous utilisons des poids propres aux variables. dans le cas présent, la perte d'efficacité est modeste, inférieure à 8 % dans tous les contextes considérés. Des résultats similaires ont été obtenus pour le cas $n = 1\,000$. Il n'y a pas de tendance claire en matière de perte d'efficacité et de taille d'échantillon.

Tableau 4.6
Ratio entre l'erreur quadratique moyenne (EQM) de chaque estimateur dont les poids sont déterminés par la variable primaire d'intérêt et l'EQM des estimateurs dont les poids sont propres à la variable

		Crédit commercial		Marge de crédit		Carte de crédit commerciale		Financement par crédit-bail	
	n	**200**	**500**	**200**	**500**	**200**	**500**	**200**	**500**
Variable primaire : crédit commercial	ARBRE	-	-	1,01	0,97	0,99	1,00	0,99	1,00
	CLASSO (unidirectionnel)	-	-	0,99	0,99	1,01	0,99	1,00	1,01
	CLASSO (bidirectionnel)	-	-	0,93	0,94	0,92	0,98	0,92	0,97
	CALASSO	-	-	0,97	0,99	1,01	0,99	0,96	1,00
Variable primaire : marge de crédit	ARBRE	1,06	0,97	-	-	0,98	1,00	0,98	0,97
	CLASSO (unidirectionnel)	0,96	0,98	-	-	0,99	1,01	0,99	0,99
	CLASSO (bidirectionnel)	0,95	0,96	-	-	0,92	0,98	0,93	0,96
	CALASSO	0,97	0,98	-	-	0,99	1,00	0,96	0,98
Variable primaire : carte de crédit d'entreprise	ARBRE	1,06	1,01	1,06	0,97	-	-	0,99	1,02
	CLASSO (unidirectionnel)	0,99	1,02	0,98	0,97	-	-	0,99	1,02
	CLASSO (bidirectionnel)	0,98	1,00	0,95	0,93	-	-	0,92	0,99
	CALASSO	1,00	1,02	0,97	0,97	-	-	1,00	1,01
Variable primaire : financement par crédit-bail	ARBRE	1,07	1,03	1,06	1,05	0,99	1,02	-	-
	CLASSO (unidirectionnel)	0,99	1,05	0,98	1,04	0,99	1,02	-	-
	CLASSO (bidirectionnel)	0,97	1,02	0,96	1,01	0,92	0,99	-	-
	CALASSO	1,00	1,05	0,98	1,05	1,00	1,01	-	-

5. Estimation selon un échantillonnage non probabiliste

Dans la section qui suit, nous étudions l'effet du biais d'échantillonnage sur les estimateurs par la régression des enquêtes selon un échantillonnage non probabiliste. À cette fin, nous avons étudié deux types de biais d'échantillonnage qui sont peut-être présents dans les échantillons non probabilistes. Plus particulièrement, nous avons considéré un scénario dans lequel la probabilité de sélection ne dépend que des données auxiliaires disponibles pour toutes les unités de la population, et un scénario où la probabilité de sélection dépend de la variable d'intérêt de l'enquête. Dans les deux scénarios, nous avons évalué le biais relatif absolu, $\left|\hat{t}_y - t_y\right| \big/ t_y$, pour chaque estimateur du total. À l'instar de Chen, Valliant et Elliott (2018), nous traitons l'échantillon non probabiliste comme un échantillon aléatoire simple et fixons les poids de sondage à $d_i = N/n$ pour l'estimation du total t_y, car le processus de sélection des échantillons non probabilistes est inconnu dans la pratique.

5.1 Les probabilités de sélection dépendent des données auxiliaires

Nous avons tiré des échantillons répétés en utilisant le même plan de sondage aléatoire simple stratifié que celui utilisé dans la section 4. Le tableau 5.1 présente le biais relatif absolu de chaque estimateur du

montant total du crédit commercial demandé en supposant que $d_i = N/n$, lorsque l'échantillon est en fait sélectionné en utilisant un échantillonnage aléatoire stratifié disproportionné.

Comme prévu, l'estimateur de HT entièrement fondé sur le plan présente le biais le plus important, et ce biais ne diminue pas lorsque la taille de l'échantillon augmente. Le biais relatif absolu des estimateurs assistés par un modèle diminue lorsque la taille de l'échantillon n augmente. L'estimateur ERG présente le biais le plus faible, en particulier pour les échantillons de petite taille. En outre, l'estimateur ERG est approximativement sans biais si les revenus sont inclus comme l'une des variables auxiliaires pour le calage. Cependant, si une sélection par étape de variables est utilisée, l'estimateur ERG n'est plus sans biais pour les échantillons de petite taille. En revanche, si les revenus ne sont pas inclus comme variable de calage, l'estimateur ERG est légèrement biaisé. Les estimateurs fondés sur la méthode LASSO et, dans une moindre mesure, l'estimateur par arbre de régression, présentent un biais lié à un petit échantillon pour $n = 200$ lorsque les revenus sont correctement inclus comme variable auxiliaire. Ceci est particulièrement évident pour les estimateurs fondés sur la méthode LASSO standard qui ne comportent pas de calage sur des totaux de population connus. Pour n égal à 500 ou à 1 000, le fait d'inclure les revenus comme variable auxiliaire diminue considérablement le biais pour l'arbre de régression et les estimateurs fondés sur la méthode LASSO avec calage, mais ne diminue que légèrement le biais pour les estimateurs fondés sur la méthode LASSO sans calage. Cela indique que l'étape supplémentaire de calage est importante pour diminuer l'effet du biais d'échantillonnage, surtout si la taille de l'échantillon est petite.

Tableau 5.1
Pourcentage de biais relatif absolu de chaque estimateur sous échantillonnage stratifié, revenus inclus et revenus non inclus comme variable auxiliaire

	Revenus inclus			Revenus non inclus		
	$n = 200$	$n = 500$	$n = 1\,000$	$n = 200$	$n = 500$	$n = 1\,000$
ERG	0,31	0,06	0,06	4,84	5,12	4,71
FSTEP	2,67	0,44	0,06	9,20	5,18	4,92
ARBRE	4,15	1,04	0,50	17,40	10,20	8,94
LASSO (unidirectionnel)	17,42	5,10	2,32	16,32	8,88	6,49
CLASSO (unidirectionnel)	7,99	0,83	0,20	9,04	5,22	4,59
LASSO (bidirectionnel)	25,36	14,28	8,40	26,31	15,16	9,89
CLASSO (bidirectionnel)	10,72	1,44	1,02	14,19	5,56	3,84
ALASSO	14,95	5,63	3,00	14,35	8,64	6,51
CALASSO	9,63	2,54	1,25	9,27	5,77	4,92
HT	49,45	48,84	48,81	49,08	49,29	48,60

Ces résultats indiquent que lorsque la probabilité de sélection dépend d'une variable auxiliaire connue, son inclusion dans le modèle de travail de l'estimateur ERG diminue efficacement l'effet du biais d'échantillonnage. Ce n'était pas le cas pour les estimateurs assistés par un modèle qui reposaient sur une sélection de variables. Le fait d'effectuer une sélection de variables peut augmenter le biais, car les variables auxiliaires qui sont prédictives en matière de probabilité de sélection peuvent ne pas être sélectionnées et correctement prises en compte. Les estimateurs fondés sur la méthode LASSO peuvent

être construits de telle sorte que les variables spécifiées par l'utilisateur soient toujours incluses dans le modèle de régression de travail. Ces variables spécifiées par l'utilisateur peuvent être ajoutées à x_i^* dans l'équation (2.5) pour forcer le calage sur les totaux de population correspondants. Malheureusement, le mécanisme de sélection sous-jacent est inconnu dans la pratique et, par conséquent, il est difficile de déterminer correctement les variables qui ont une incidence sur la probabilité de sélection.

5.2 Les probabilités de sélection dépendent de la variable de l'étude

Ensuite, nous avons tiré des échantillons répétés en utilisant l'échantillonnage de Poisson où les probabilités d'échantillonnage dépendent de la variable d'intérêt de l'enquête. Nous supposons que les probabilités d'échantillonnage de Poisson sont données par :

$$\text{logit}\left(p_i\right) = \beta_0 + \beta_1 y_i$$

où y_i est le montant du crédit commercial demandé en millions de dollars, $\beta_1 = 0{,}5$ et $\beta_0 = -3{,}80; -2{,}85; -2{,}10.$ Les valeurs de l'ordonnée à l'origine, β_0, ont été choisies de manière à obtenir des tailles d'échantillon d'environ 200, 500 et 1 000 unités, en moyenne pour les échantillons simulés. Dans le cadre de ce plan de sondage, les unités dont les montants demandés pour le crédit commercial sont plus importants ont une probabilité plus élevée d'être échantillonnées et sont donc surreprésentées. Le tableau 5.2 présente le biais relatif absolu de chaque estimateur du montant total du crédit commercial demandé en supposant que $d_i = N/n$, lorsque l'échantillon est sélectionné en utilisant l'échantillonnage de Poisson informatif ci-dessus. Dans le cas présent, tous les estimateurs sont fortement biaisés, car le modèle de population ne tient pas en raison de l'échantillonnage informatif. L'ampleur du biais est très similaire d'un estimateur à l'autre et ne diminue pas sensiblement lorsque la taille de l'échantillon augmente. L'inclusion ou l'exclusion des revenus en tant que variable auxiliaire n'a pas d'incidence sur le biais.

Tableau 5.2
Pourcentage de biais relatif absolu de chaque estimateur dans le cadre d'un échantillonnage de Poisson, revenus inclus et revenus non inclus comme variable auxiliaire

	Revenus inclus			Revenus non inclus		
	β_0 = -3,8	β_0 = -2,85	β_0 = -2,1	β_0 = -3,8	β_0 = -2,85	β_0 = -2,1
ERG	23,53	22,27	20,45	24,74	22,91	21,21
FSTEP	24,54	22,55	20,58	25,16	23,24	21,15
ARBRE	24,07	22,73	20,15	24,93	22,47	20,55
LASSO (unidirectionnel)	24,29	22,73	20,65	25,45	23,29	21,38
CLASSO (unidirectionnel)	23,02	22,30	20,47	24,74	22,99	21,23
LASSO (bidirectionnel)	23,15	22,06	20,17	24,66	22,73	20,62
CLASSO (bidirectionnel)	20,11	20,18	19,01	22,62	21,63	19,98
ALASSO	24,44	22,72	20,66	25,50	23,21	21,36
CALASSO	23,91	22,46	20,53	25,10	23,01	21,25
HT	29,12	27,95	25,57	29,36	27,53	25,45

6. Conclusions

Nous avons évalué les performances de plusieurs estimateurs par la régression d'enquête assistés par un modèle, dans le contexte d'un échantillonnage probabiliste et non probabiliste, au moyen d'une étude par simulations. Tout d'abord, nous traitons des conclusions générales de notre étude par simulations reposant sur des échantillons probabilistes au moyen d'un plan à échantillonnage aléatoire simple (EAS) stratifié. Dans le contexte de nos données d'enquête-entreprise avec toutes les variables auxiliaires catégoriques, l'estimateur par arbre de régression et l'estimateur fondé sur la méthode LASSO (bidirectionnel) avec effets d'interaction à deux facteurs sont les seuls estimateurs assistés par un modèle qui procurent des gains d'efficacité, par rapport à l'estimateur de HT, en présence d'une petite taille d'échantillon et d'un nombre élevé de catégories de variables auxiliaires utilisées. De même, l'estimateur de variance de l'estimateur par arbre de régression est le moins biaisé dans ce scénario. Lorsque la taille de l'échantillon augmente, la différence d'efficacité entre les estimateurs par la régression d'enquête assistés par un modèle devient négligeable et tous sont légèrement plus efficaces que l'estimateur de HT. En général, les gains d'efficacité potentiels des estimateurs assistés par un modèle par rapport à l'estimateur de HT dépendent de la puissance prédictive du modèle. Dans notre population de simulation, la force de la relation entre la variable de l'étude et les variables auxiliaires catégoriques disponibles est quelque peu faible, comme en témoigne le coefficient de détermination ajusté R^2 autour de 0,20. Nous avons donc généré des variables d'étude conduisant à des valeurs plus grandes de R^2 autour de 0,50 en rendant la variance d'erreur du modèle plus petite. Comme prévu, les estimateurs assistés par un modèle ont procuré des gains d'efficacité importants par rapport à l'estimateur de HT dans tous les cas, comme on le voit dans le tableau 4.2 qui montre que l'estimateur par arbre de régression et l'estimateur fondé sur la méthode LASSO avec effets d'interaction sont plus efficaces que l'estimateur ERG couramment utilisé en présence d'interactions à deux facteurs. De plus, les poids de régression pour l'estimateur par arbre et les poids de calage pour les estimateurs par calage de type LASSO sont beaucoup moins variables, notamment pour les échantillons de petite taille, que les poids pour l'estimateur ERG. Nous avons également examiné la performance des estimateurs par calage de type LASSO et par arbre de régression dans un scénario où il n'y a pas d'effets principaux et où seules des interactions à deux facteurs sont présentes et dans un autre scénario où il y a une multicolinéarité entre les variables auxiliaires. Dans ce dernier scénario, l'estimateur ERG n'est pas applicable, et nous montrons que l'estimateur par arbre de régression et les estimateurs fondés sur la méthode LASSO fournissent un moyen automatique de supprimer les variables auxiliaires colinéaires sans diminuer les gains d'efficacité potentiels. Dans l'ensemble, nous recommandons l'utilisation des estimateurs fondés sur la méthode LASSO (bidirectionnels) ou de l'estimateur par arbre de régression pour ce qui est de l'efficacité lorsque des interactions à deux facteurs sont susceptibles d'être présentes parmi les variables auxiliaires catégoriques. Même dans le cas de modèles ne comportant que des effets principaux, les deux méthodes obtiennent de bons résultats par rapport à l'estimateur ERG en ce qui a trait à l'EQM, car l'estimateur fondé sur la méthode LASSO (bidirectionnel) réduit automatiquement à zéro les coefficients de régression associés

aux interactions, tandis que l'estimateur par arbre de régression ne nécessite pas la spécification de la fonction moyenne. Dans d'autres contextes où il existe des preuves de relations complexes non linéaires et non additives entre la variable d'enquête d'intérêt et les variables auxiliaires, l'utilisation d'autres méthodes d'apprentissage automatique basées sur des arbres, comme XGBoost et les forêts aléatoires, devrait être étudiée.

Dans la section 4.3, nous avons étudié la performance des estimateurs de variance en matière de biais relatif et montré que tous les estimateurs de variance présentent une sous-estimation importante pour les tailles d'échantillon $n = 200$ et les 28 catégories x. Le biais relatif de l'estimateur de variance de l'arbre de régression n'a pas diminué lorsque la taille de l'échantillon a augmenté, contrairement aux autres cas, et cela pourrait être attribuable à un surajustement. Dans le contexte de la méthode des forêts aléatoires, Dagdoug, Goga et Haziza (2021) ont examiné une procédure basée sur la validation croisée qui a conduit à de petits biais relatifs et à de bons taux de couverture. Il serait intéressant d'étudier une procédure similaire pour l'estimation de la variance de l'estimateur par arbre de régression.

Un inconvénient majeur des méthodes basées sur la méthode LASSO et par arbre de régression est que les procédures d'estimation ne produisent pas un ensemble de poids génériques qui peuvent être appliqués à toutes les variables de l'étude, y. Une autre approche possible consiste à dériver des poids de régression basés sur une variable primaire d'intérêt et à appliquer cet ensemble de poids aux variables d'étude connexes. Dans le contexte de l'enquête considérée dans la présente étude, l'utilisation d'un seul ensemble de poids pour un groupe de variables connexes a entraîné une faible perte d'efficacité, par rapport à l'utilisation de poids propres aux variables. De même, le biais des estimateurs est resté négligeable. Cette approche permet de conserver les propriétés souhaitables des poids de régression, une faible variabilité et, dans le cas de l'estimateur par arbre de régression, des poids strictement positifs. Cependant, les propriétés asymptotiques des estimateurs fondés sur la méthode LASSO et par arbre de régression n'ont pas été calculées pour un seul ensemble de poids, appliqué à de multiples variables d'étude.

Nous avons également envisagé l'utilisation d'estimateurs par la régression d'enquête assistés par un modèle pour les données provenant d'un échantillonnage probabiliste mal défini, traité comme un échantillon non probabiliste. Lorsque la probabilité de sélection dépend d'une variable auxiliaire observée, le biais des estimateurs assistés par un modèle diminue lorsque la taille de l'échantillon augmente. L'inclusion de la variable auxiliaire appropriée dans le modèle de travail de l'estimateur ERG élimine efficacement le biais d'échantillonnage. Il est difficile d'y parvenir dans la pratique, car le processus de sélection est inconnu. Effectuer une sélection de variables peut augmenter le biais des estimateurs par la régression d'enquête assistés par un modèle, car les variables auxiliaires liées à la probabilité de sélection peuvent ne pas être incluses dans le modèle de régression. En effet, dans nos simulations, le fait d'inclure correctement les revenus comme variable auxiliaire potentielle n'a pas nécessairement diminué le biais des estimateurs LASSO.

Lorsque la probabilité de sélection dépend de la variable d'intérêt de l'enquête, tous les estimateurs sont fortement biaisés. L'ampleur du biais est très similaire d'un estimateur à l'autre et ne diminue pas de

façon importante lorsque la taille de l'échantillon augmente. Dans notre population de simulation, les variables auxiliaires ne sont pas hautement prédictives pour les variables d'enquête d'intérêt. L'examen de l'incidence de la force de la relation entre les variables auxiliaires et la variable d'intérêt en présence d'une sélection informative mérite une étude plus approfondie.

Le biais d'échantillonnage peut ne pas être réduit en utilisant uniquement un échantillon non probabiliste, comme l'a démontré notre étude par simulations. Les méthodes basées sur l'intégration d'un échantillon non probabiliste permettant d'observer les variables de l'étude et les variables auxiliaires associées avec un échantillon probabiliste permettant d'observer uniquement les mêmes variables auxiliaires ont le potentiel de réduire le biais d'échantillonnage en modélisant les probabilités de participation (Chen, Li et Wu, 2020). Les méthodes de sélection à deux bases de sondage sont également possibles lorsque la variable d'étude est observée dans les deux échantillons et que les unités de l'échantillon probabiliste appartenant à l'échantillon non probabiliste peuvent être déterminées sans erreur de couplage et sans qu'il soit nécessaire de modéliser les probabilités de participation (Kim et Tam, 2020; Rao, 2021 et Beaumont, 2020). Toutefois, la méthode à deux bases de sondage n'est efficace que lorsque la fraction de sondage de l'échantillon non probabiliste est importante. Nous étudions les méthodes ci-dessus dans le contexte des enquêtes-entreprises, par exemple en intégrant des données d'enquête à des données administratives incomplètes traitées comme un échantillon non probabiliste.

Remerciements

Nous remercions Wesley Yung pour avoir entamé ce travail et pour ses commentaires et ses suggestions constructifs. Nous remercions également les réviseurs, le rédacteur en chef et le rédacteur adjoint pour leurs commentaires constructifs et leurs suggestions.

Bibliographie

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

Breidt, F.J., et Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

Buskirk, T.D., Kirchner, A., Eck, A. et Signorino, C.S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1), 1-10.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63(3), 615-620.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(523), 2011-2021.

Chen, J.K.T., Valliant, R.L. et Elliott, M.R. (2018). Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif. *Techniques d'enquête*, 44, 1, 125-155. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54963-fra.pdf.

Chen, J.K.T., Valliant, R.L. et Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3), 657-681.

Dagdoug, M., Goga, C. et Haziza, D. (2021). Model-assisted estimation through random forests infinite population sampling. *Journal of the American Statistical Association* (à paraître).

Friedman, J., Hastie, T., Simon, N., Qian, J. et Tibshirani, R. (2017). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* R package version 2.0-13.

Kern, C., Klausch, T. et Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1), 73-93.

Kern, C., Li, Y. et Wang, L. (2020). Boosted kernel weighting-using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smaa028.

Kim, J.K., et Tam, S.M. (2020). Data integration combing big data and survey sample data for finite population inference. *Revue Internationale de Statistique*, 89(2), 382-401.

McConville, K.S. (2011). *Department of Statistics Improved Estimation for Complex Surveys Using Modern Regression Techniques*, thèse de doctorat non-publiée, Colorado State University.

McConville, K.S., et Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

McConville, K.S., Breidt, F.J., Lee, T.C.M. et Moisen, G.G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, 5(2), 131-158.

McConville, K.S., Tang, B., Zhu, G., Li, S., Cheung, S. et Toth, D. (2018). *mase: Model-Assisted Survey Estimators*. R package version 0.1.1.

Rafei, A., Flannagan, C.A. et Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8(1), 148-180.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83(1), 242-272 (publié en ligne en avril 2020).

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. et Firth, D. (2017). *MASS: Modern Applied Statistics with S.* R package version 7. 3-47.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B,* 58(1), 267-288.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.