Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Rendement relatif des méthodes d'enquête fondées sur l'estimation par la régression assistée par un modèle

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

Date de diffusion : le 29 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Rendement relatif des méthodes d'enquête fondées sur l'estimation par la régression assistée par un modèle

Erin R. Lundy[1] et J.N.K. Rao[2]

Résumé

Le recours à des données auxiliaires pour améliorer l'efficacité d'estimateurs de totaux et de moyennes au moyen d'une procédure d'estimation d'enquête assistée par un modèle de régression a reçu une attention considérable ces dernières années. Des estimateurs par la régression généralisée (GREG), fondés sur un modèle de régression linéaire, sont actuellement utilisés dans le cadre d'enquêtes auprès d'établissements, à Statistique Canada et au sein de plusieurs autres organismes de statistiques. Les estimateurs GREG utilisent des poids d'enquête communs à toutes les variables d'étude et un calage aux totaux de population de variables auxiliaires. De plus en plus de variables auxiliaires sont disponibles et certaines peuvent être superflues. Cela mène à des poids GREG instables lorsque toutes les variables auxiliaires disponibles, y compris les interactions parmi les variables catégoriques, sont utilisées dans le modèle de régression linéaire. En revanche, de nouvelles méthodes d'apprentissage automatique, comme les arbres de régression et la méthode LASSO, sélectionnent automatiquement des variables auxiliaires significatives et mènent à des poids non négatifs stables et à d'éventuels gains d'efficacité par rapport à la méthode GREG. Dans cet article, une étude par simulations, fondée sur un ensemble de données-échantillon d'une enquête-entreprise réelle traité comme la population cible, est menée afin d'examiner le rendement relatif de la méthode GREG, d'arbres de régression et de la méthode LASSO sur le plan de l'efficacité des estimateurs.

Mots-clés : inférence assistée par modèle; estimation par calage; sélection du modèle; estimateur par la régression généralisée.

1. Introduction

Statistique Canada, comme plusieurs autres organismes de statistique, s'intéresse de plus en plus à l'utilisation des données auxiliaires, susceptibles de provenir de sources administratives, afin d'améliorer l'efficacité des estimateurs. Dans différentes disciplines, les techniques d'apprentissage automatique sont des outils populaires afin d'utiliser cette information auxiliaire. Souvent, ces méthodes n'exigent pas d'hypothèses de distribution comme les méthodes plus classiques et peuvent s'adapter à des relations non linéaires et non additives complexes entre les résultats et les variables auxiliaires.

Récemment, l'utilisation de techniques d'apprentissage automatique a été envisagée pour améliorer l'efficacité des estimateurs de totaux et de moyennes obtenus au moyen d'une procédure d'estimation d'enquête probabiliste assistée par un modèle de régression. Les estimateurs d'enquête des totaux de population finie assistés par un modèle de régression peuvent réduire la variabilité et entraîner des gains d'efficacité significatifs si les variables auxiliaires disponibles sont fortement associées à la variable d'intérêt de l'enquête. De plus en plus, de nombreuses variables auxiliaires sont disponibles et certaines peuvent être superflues. Dans ce cas, la sélection de variables suivie d'une procédure d'estimation par la régression fondée sur le modèle sélectionné peut améliorer l'efficacité des estimateurs d'enquête par la régression des totaux de population finie.

[1] Erin R. Lundy, Division des méthodes d'intégration statistique, Statistique Canada, Ottawa (Ontario) K1A 0T6
[2] J.N.K. Rao, École de mathématiques et de statistiques, Université Carleton, Ottawa (Ontario) K1S 5B6.

2. Estimation assistée par un modèle selon un échantillonnage probabiliste

2.1 Estimateurs GREG

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

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

où $d_i = \pi_i^{-1}$ désigne les poids d'échantillonnage.

Une stratégie pour utiliser des données auxiliaires dans la procédure d'estimation consiste à utiliser un estimateur de t_y assisté par un modèle en spécifiant un modèle de travail pour la moyenne de y étant donné $\boldsymbol{x}$ et à utiliser ce modèle pour prédire les valeurs de y. La spécification d'un modèle de travail de régression linéaire donne l'estimateur par la régression généralisée (GREG) (Cassel, Sarndal et Wretman, 1976). Ici, nous considérons l'estimateur GREG sous un modèle de travail de régression linéaire

$$y_i = \boldsymbol{x_i^T \beta} + \epsilon_i \qquad (2.1)$$

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

$$\hat{t}_{y,GREG} = \sum_{i \epsilon s} \frac{y_i - \boldsymbol{x_i^T \hat{\beta}_s}}{\pi_i} + \sum_{i \epsilon U} \boldsymbol{x_i^T \hat{\beta}_s} \quad (2.2)$$

où $\boldsymbol{\hat{\beta}_s}$ est un vecteur des coefficients de régression estimés.

L'estimateur GREG peut également être écrit comme 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 :

$$\hat{t}_{y,GREG} = \sum_{i \in s} \left[1 + (\boldsymbol{t_x} - \boldsymbol{\hat{t}_{x,HT}})^T \left(\sum_{k \in s} \boldsymbol{x_k x_k^T} d_k\right)^{-1} \boldsymbol{x_i}\right] d_i y_i = \sum_{i \in s} w_i \, y_i, \quad (2.3)$$

où $\boldsymbol{t_x}$ est le vecteur du total de population connu des covariables $\boldsymbol{x}$ et $\hat{\boldsymbol{t}}_{\boldsymbol{x,HT}}$ est le vecteur de l'estimateur de Horvitz-Thompson des totaux de la population de covariables $\boldsymbol{t_x} = \sum_{i \in U} \boldsymbol{x_i}$.

Si une procédure de sélection de variables, comme une procédure pas à pas, 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 pas à pas 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 avec de nombreuses variables auxiliaires, y compris les interactions des variables auxiliaires catégoriques, peut produire des poids GREG très variables w_i et gonfler considérablement la variance de l'estimateur GREG. 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 avec LASSO

Si le modèle de régression linéaire de (2.1) est parcimonieux, c.-à-d. que p est grand et, supposons, seuls p_0 des p coefficients de régression ne sont pas nuls, alors l'estimation des coefficients nuls entraîne une variation supplémentaire dans l'estimateur GREG (2.2). Dans ce cas, la sélection du modèle pour supprimer les variables superflues pourrait réduire la variance sous le plan globale de l'estimateur GREG, ce qui produirait des estimations plus efficaces des totaux de population finie. La méthode du LASSO (*Least Absolute Shrinkage and Selection Operator*, opérateur de sélection et réduction par moindres valeurs absolues), élaborée par Tibshirani (1996), effectue simultanément la sélection du modèle et l'estimation du coefficient en réduisant certains coefficients de régression à zéro. La méthode du LASSO estime les coefficients en minimisant la somme des résidus au carré 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 LASSO estimés à partir des poids d'enquête donnés par

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{s,L}} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y_s} - \boldsymbol{X_s}\boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{\Pi_s^{-1}}(\boldsymbol{Y_s} - \boldsymbol{X_s}\boldsymbol{\beta}) \; + \; \lambda \sum_{j=1}^{p} |\beta_j|,$$

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

$$\hat{t}_{y,LASSO} = \sum_{i \epsilon s} \frac{y_i - \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s,L}}}{\pi_i} + \sum_{i \epsilon U} \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s,L}}.$$

2.3 Estimateurs par calage de la méthode LASSO

Les méthodes de LASSO ne produisent pas directement de poids de régression, car les estimateurs ne peuvent pas être exprimés comme combinaisons pondérées des valeurs *y*. McConville et coll. (2017) ont élaboré des poids de régression de type LASSO à l'aide d'une approche par calage assisté par un modèle. Ces poids d'enquête de régression LASSO dépendent de la variable d'intérêt, *y*. L'estimateur par calage de type LASSO est calculé par régression de la variable d'intérêt, y_i, sur une ordonnée à l'origine et la fonction moyenne ajustée par la méthode LASSO $\boldsymbol{x_i^T \hat{\beta}_{s,L}}$. L'estimateur par calage du LASSO peut être écrit sous la même forme que (2.3), où $\boldsymbol{x_i}$ est remplacé par $\boldsymbol{x_i^*} = (1, \boldsymbol{x_i^T \hat{\beta}_{s,L}})^{\boldsymbol{T}}$:

$$\hat{t}_{y,CLASSO} = \sum_{i\in s}\left[1 + (\boldsymbol{t_{x^*}} - \boldsymbol{\hat{t}_{x^*,HT}})^T \left(\sum_{k\in s} \boldsymbol{x_k^* x_k^{*T}} d_k\right)^{-1} \boldsymbol{x_i^*}\right] d_i y_i.$$

Les poids des estimateurs par calage du LASSO sont calés sur la taille de population *N* et sur le total de la population des fonctions moyennes ajustées par la méthode LASSO.

2.4 Estimateur par arbre de régression

L'estimateur GREG peut aussi être exprimé comme suit :

$$\hat{t}_{y,r} = \sum_{i\epsilon s} \frac{y_i - \hat{h}_n(x_i)}{\pi_i} + \sum_{i\epsilon U} \hat{h}_n(\boldsymbol{x_i}) \text{ , (2.4)}$$

où $\hat{h}_n(\boldsymbol{x_i})$ est un estimateur de la fonction moyenne de Y_i étant donné $\boldsymbol{X_i} = \boldsymbol{x_i}$, $h(\boldsymbol{x_i}) = E(Y_i|\boldsymbol{X_i} = \boldsymbol{x_i})$, à partir des données de l'échantillon $(y_i, \boldsymbol{x_i}), i \in s$. Comme solution de rechange à un modèle de régression linéaire, McConville et Toth (2019) ont proposé d'estimer $h(\boldsymbol{x})$ au moyen d'un modèle d'arbre de régression. Le modèle de l'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.

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

$$\tilde{h}_n(\boldsymbol{x_i}) = \tilde{\#}(B_{nk})^{-1} \sum_{i\in s} \pi_i^{-1} y_i \, I(\boldsymbol{x_i} \in B_{nk}) = \tilde{\mu}_{nk}, \qquad (2.5)$$

où

$$\tilde{\#}(B_{nk}) = \sum_{i\in s} \pi_i^{-1} \, I(\boldsymbol{x_i} \in B_{nk})$$

est l'estimateur de HT de la taille de la population dans la case B_{nk}. On obtient l'estimateur par arbre de régression $\hat{t}_{y,TREE}$ en insérant l'équation (2.5) dans l'estimateur par la régression généralisée, donné dans l'équation (2.4), ce qui donne l'estimateur post-stratifié

$$\hat{t}_{y,TREE} = \sum_{k} N_k \tilde{\mu}_{nk},$$

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

Puisque $\tilde{h}_n(\boldsymbol{x_i})$ peut être écrit 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 post-stratifié, 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, on est certain que les poids de régression ne sont pas négatifs. Les poids produits par cette procédure d'estimation dépendent de la variable d'intérêt, *y*.

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 les performances des estimateurs d'enquête assistés par un modèle de régression à celles de l'estimateur de HT fondé uniquement sur le plan d'échantillonnage. 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. L'Enquête sur le financement et la croissance des petites et moyennes entreprises (EFCPME) est une enquête périodique auprès des entreprises, qui recueille des renseignements sur les types de financement utilisés par les entreprises.

3.1 Méthodologie de la simulation

Nous avons examiné des tailles d'échantillon de $n = \{200, 500, 1000\}$ à partir des 9 115 répondants de l'ensemble de données de l'EFCPME. Cet ensemble de données a été traité comme étant la population cible et des échantillons répétés ont été tirés au moyen d'un échantillonnage aléatoire simple stratifié. Nous avons supposé qu'il y avait deux strates : la strate A est composée d'unités dont le revenu est inférieur à 2,5 millions de dollars et la strate B 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 d'échantillonnage, les unités à revenu plus élevé sont surreprésentées, ce qui entraîne un plan de sondage à probabilités inégales.

Pour chaque échantillon, des modèles utilisant quatre variables *x* catégoriques, l'industrie (10 catégories), la taille de l'effectif (4 catégories), la région (6 catégories) et le revenu (8 catégories), ont servi à estimer le montant total des crédits commerciaux demandés. Les résultats ont ensuite été comparés au total réel. Pour chacune des trois tailles

d'échantillon différentes, 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 3.1-1 ci-dessous :

Tableau 3.1-1
Résumé des estimateurs assistés par un modèle pris en compte dans l'étude par simulations

Estimateur	Données auxiliaires	Poids de régression	Totaux de calage
GREG	Totaux marginaux	Indépendants de *y*	Toutes les variables auxiliaires
GREG avec sélection ascendante de variable (FSTEP)	Valeurs individuelles	Dépendants de *y*	Variables auxiliaires sélectionnées
Arbre de régression (TREE)	Valeurs individuelles	Dépendants de *y*, strictement positifs	Taille de la population de chaque case
LASSO (LASSO)	Valeurs individuelles		
LASSO calé (CLASSO)	Valeurs individuelles	Dépendants de *y*	Taille de la population et fonction moyenne ajustée par la méthode LASSO

3.2 Performance des estimateurs selon l'EQM sous le plan d'échantillonnage

Nous avons calculé l'erreur quadratique moyenne (EQM) sous le plan d'échantillonnage à partir des 5 000 estimations totales selon la taille de l'échantillon. Les résultats sont indiqués dans le tableau 3.2-1 Pour n=200, l'estimateur par arbre de régression et l'estimateur 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. Quand la taille de l'échantillon augmente, les gains d'efficacité des estimateurs d'enquête assistés par un modèle de régression, par rapport à l'estimateur de HT, sont essentiellement égaux. Pour les tailles d'échantillon plus grandes, il y a peu de gains d'efficacité quand on utilise les estimateurs assistés par un modèle par rapport à l'estimateur de HT, ce qui indique que les variables auxiliaires ne sont pas fortement liées à la variable d'intérêt.

Tableau 3.2-1
Ratio de l'EQM de chaque estimateur par rapport à l'EQM de l'estimateur de HT

	n=200	n=500	n=1000
GREG	1,084	0,959	0,954
FSTEP	1,040	0,945	0,958
TREE	0,983	0,963	0,949
LASSO (unidirectionnel)	1,009	0,946	0,947
CLASSO (unidirectionnel)	1,042	0,952	0,949
LASSO (bidirectionnel)	0,981	0,935	0,936
CLASSO (bidirectionnel)	1,045	0,959	0,950

Les gains d'efficacité potentiels des estimateurs assistés par un modèle dépendent de la puissance prédictive du modèle de travail. Nous avons examiné davantage les différences entre les différents estimateurs d'enquête assistés par un modèle en exécutant des simulations supplémentaires en utilisant différentes variables d'intérêt de l'enquête, générées selon les modèles suivants :

- modèle de LASSO avec effets principaux uniquement;
- modèle de LASSO avec effets principaux et interactions bidirectionnelles;
- modèle d'arbre de régression;
- modèle de régression linéaire sans effets principaux et avec quelques interactions bidirectionnelles;
- modèle de régression linéaire sans effets principaux et avec une seule interaction tridirectionnelle.

Le tableau 3.2-2 présente le ratio entre l'EQM sous le plan de chaque estimateur et celle de l'estimateur de HT selon les modèles de LASSO et d'arbre de régression générant la variable d'enquête d'intérêt pour une taille d'échantillon de *n=1000.* Comme prévu, l'estimateur basé sur 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 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. Dans tous les cas, on obtient des gains d'efficacité importants, par rapport à l'estimateur de HT fondé sur le plan de sondage.

Tableau 3.2-2
Ratio de l'EQM pour 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
GREG	0,749	0,855	0,878
FSTEP	0,749	0,855	0,876
TREE	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

Le tableau 3.2-3 montre le ratio entre l'EQM sous le plan d'échantillonnage des estimateurs et celle de l'estimateur de HT, où la variable d'enquête est générée à partir de modèles sans effets principaux pour les tailles d'échantillon de *n=200* et *n=1000*. Ici, les estimateurs LASSO avec interactions bidirectionnelles et l'estimateur par arbre de régression sont significativement plus efficaces que les estimateurs assistés par un modèle fondés sur des modèles avec effets principaux seulement pour les tailles d'échantillon plus grandes. Par rapport à l'estimateur GREG couramment utilisé, les gains d'efficacité des estimateurs LASSO avec interactions bidirectionnelles et de l'estimateur par arbre de régression sont significativement plus élevés en l'absence d'effets principaux. Ce résultat apparaît de façon évidente si l'on compare la colonne du LASSO bidirectionnel dans le tableau 3.2-2 à la colonne avec interaction bidirectionnelle du tableau 3.2-3. L'EQM relative est très semblable pour l'estimateur LASSO bidirectionnel et l'estimateur par arbre de régression, mais plus près de 1 pour l'estimateur GREG et l'estimateur LASSO unidirectionnel.

Tableau 3.2-3
Ratio de l'EQM pour chaque estimateur par rapport à l'EQM de HT selon des modèles sans effets principaux

	n=200		n=1000	
	Interactions bidirectionnelles	Une interaction tridirectionnelle	Interactions bidirectionnelles	Une interaction tridirectionnelle
GREG	1,045	1,044	0,935	0,911
FSTEP	1,042	1,013	0,935	0,910
TREE	1,015	0,975	0,824	0,796
LASSO (unidirectionnel)	0,982	0,959	0,930	0,899
CLASSO (unidirectionnel)	1,031	0,985	0,936	0,902
LASSO (bidirectionnel)	0,912	0,957	0,783	0,815
CLASSO (bidirectionnel)	0,990	1,010	0,795	0,818

4. Conclusions

Nous avons évalué les performances de plusieurs estimateurs d'enquête assistés par un modèle de régression, dans le contexte d'un échantillonnage probabiliste, au moyen d'une étude par simulations. Dans le contexte de nos données d'enquête sur les entreprises avec toutes les variables auxiliaires catégoriques, l'estimateur par arbre de régression et l'estimateur 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. Quand la taille de l'échantillon augmente, la différence d'efficacité entre les estimateurs d'enquête assistés par un modèle de régression 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 n'est pas très grande. C'est pourquoi nous avons généré des variables d'étude de sorte qu'il y ait une relation plus forte entre la variable d'étude et les variables auxiliaires catégoriques disponibles. 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 les tableaux 3.2-2 et 3.3-3 qui montrent que l'estimateur par arbre de régression et l'estimateur LASSO avec effets d'interaction sont plus efficaces que l'estimateur GREG couramment utilisé en présence d'interactions à deux facteurs. Dans l'ensemble, nous recommandons d'utiliser soit un estimateur LASSO (bidirectionnel) soit un estimateur par arbre de régression pour gagner en efficacité quand des interactions à deux facteurs sont susceptibles d'être présentes parmi les variables auxiliaires catégoriques. Même dans le cas des modèles qui ont seulement des effets principaux, les deux méthodes donnent de bons résultats par rapport à GREG en ce qui a trait à l'EQM parce que l'estimateur LASSO (bidirectionnel) réduit automatiquement à zéro les coefficients de régression associés aux interactions, tandis que l'estimateur par arbre de régression n'a pas besoin de spécification de la fonction moyenne.

Nos travaux actuels portent sur l'identification de scénarios où les méthodes d'apprentissage automatique aux fins de calage donnent de bien meilleurs résultats que les méthodes classiques. Nous étudions également l'estimation assistée par un modèle dans le cadre d'un échantillonnage non probabiliste.

Bibliographie

Cassel CM, CE Sarndal, et JH. Wretman (1976), « Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Population », *Biometrika*, 63, p. 615-620.

McConville KS, FJ Breidt, TCM Lee, et GG Moisen. (2017), « Model-assisted Survey Regression Estimation with the LASSO », *Journal of Survey Statistics and Methodology*, 5, p. 131-158.

McConville KS, et D. Toth (2019), « Automated Selection of Post-strata using a Model-assisted Regression Tree Estimator », *Scandinavian Journal of Statistics*, 46, p. 389-413.

Tibshirani R. (1996), « Regression Shrinkage and Selection via the LASSO », *Journal of the Royal Statistical Society, Series B,* 58, p. 267-288.