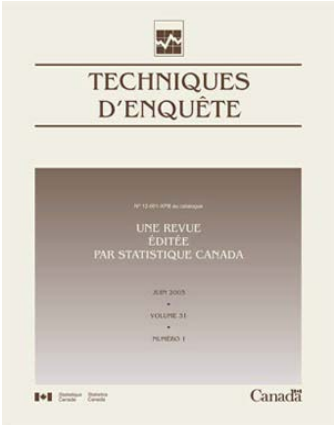
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

# Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif




par Jack Kuang Tsung Chen, Richard L. Valliant et Michael R. Elliott

Date de diffusion : le 21 juin 2018

Statistique Canada Statistics Canada

Canada

## Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

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

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

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

**Programme des services de dépôt**

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

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

## Note de reconnaissance

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

## Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

| | |
|---|---|
| . | indisponible pour toute période de référence |
| .. | indisponible pour une période de référence précise |
| ... | n'ayant pas lieu de figurer |
| 0 | zéro absolu ou valeur arrondie à zéro |
| $0^s$ | valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie |
| $^p$ | provisoire |
| $^r$ | révisé |
| x | confidentiel en vertu des dispositions de la *Loi sur la statistique* |
| $^E$ | à utiliser avec prudence |
| F | trop peu fiable pour être publié |
| * | valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05) |

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 125-155
Statistique Canada, nº 12-001-X au catalogue



# Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif

**Jack Kuang Tsung Chen, Richard L. Valliant et Michael R. Elliott[1]**

Résumé

Le cadre fondé sur l'échantillonnage probabiliste a joué un rôle dominant en recherche par sondage, parce qu'il fournit des outils mathématiques précis pour évaluer la variabilité d'échantillonnage. Toutefois, en raison de la hausse des coûts et de la baisse des taux de réponse, l'usage d'échantillons non probabilistes s'accroît, particulièrement dans le cas de populations générales, pour lesquelles le tirage d'échantillons à partir d'enquêtes en ligne devient de plus en plus économique et facile. Cependant, les échantillons non probabilistes posent un risque de biais de sélection dû à des différences d'accès et de degrés d'intérêt, ainsi qu'à d'autres facteurs. Le calage sur des totaux statistiques connus dans la population offre un moyen de réduire éventuellement l'effet du biais de sélection dans les échantillons non probabilistes. Ici, nous montrons que le calage assisté par un modèle en utilisant le LASSO adaptatif peut donner un estimateur convergent d'un total de population à condition qu'un sous-ensemble des variables explicatives réelles soit inclus dans le modèle de prédiction, permettant ainsi qu'un grand nombre de covariables possibles soit incluses sans risque de surajustement. Nous montrons que le calage assisté par un modèle en utilisant le LASSO adaptatif produit une meilleure estimation, pour ce qui est de l'erreur quadratique moyenne, que les méthodes concurrentes classiques, tels les estimateurs par la régression généralisée (GREG), quand un grand nombre de covariables sont nécessaires pour déterminer le modèle réel, sans vraiment qu'il y ait perte d'efficacité par rapport à la méthode GREG quand de plus petits modèles suffisent. Nous obtenons aussi des formules analytiques pour les estimateurs de variance des totaux de population, et comparons le comportement de ces estimateurs aux estimateurs bootstrap. Nous concluons par un exemple réel en utilisant des données provenant de la *National Health Interview Survey*.

**Mots-clés :** Estimateurs de type LASSO adaptatif; estimateur par la régression généralisée; échantillon non représentatif; surajustement; sélection des variables; propriété d'oracle.

# 1 Introduction

L'échantillonnage probabiliste a joué un rôle prépondérant en recherche par sondage pendant la plus grande partie du siècle dernier (Stephan, 1948; Frankel et Frankel, 1987). Étant donné des mesures complètes d'unités échantillonnées dont les probabilités de sélection sont connues, la théorie de la randomisation élimine le biais de sélection en générant des échantillons représentatifs de la population cible. Par ailleurs, les échantillons non probabilistes produits sans connaître les probabilités de sélection posent automatiquement un risque de biais de sélection, car les échantillons peuvent différer de la population cible en ce qui concerne des statistiques importantes (Groves, 2006). Les échecs bien documentés des sondages réalisés pendant les campagnes électorales présidentielles de 1936 et de 1948 mettent en relief les erreurs qui peuvent être commises lorsqu'on fait des inférences sur une population à partir d'échantillons non probabilistes (Mosteller, 1949).

Bien que le cadre d'échantillonnage probabiliste fournisse aux spécialistes des enquêtes des outils mathématiques précis pour évaluer et corriger les erreurs d'échantillonnage, la baisse des taux de réponse aux enquêtes s'appuyant sur les méthodes traditionnelles de collecte de données suscite des inquiétudes

1. Jack Kuang Tsung Chen, statisticien, Survey Monkey Inc., Palo Alto, CA. Courriel : jjkkttcc@gmail.com; Richard L. Valliant, professeur-chercheur, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. Courriel : valliant@umich.edu; Michael R. Elliott, professeur-chercheur, Survey Research Center, Institute for Social Research, et professeur, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI. Courriel : mrelliot@umich.edu.

quant au biais de non-réponse potentiellement élevé des échantillons probabilistes. Le *Pew Research Center* a indiqué que les taux de réponse à leurs sondages téléphoniques sont passés de 36 % en 1997 à 9 % en 2012 (Kohut, Keeter, Doherty, Dimock et Christian, 2012), ce qui fait penser que l'échantillonnage probabiliste par téléphone pourrait ne plus être une méthode viable pour les enquêtes auprès des populations générales. En outre, l'obtention de données sans exercer un grand contrôle sur l'ensemble d'unités auprès desquelles elles sont recueillies est souvent un exercice moins coûteux et plus rapide que l'échantillonnage probabiliste. Ces raisons font que l'échantillonnage non probabiliste connaît à l'heure actuelle une sorte de renaissance (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau, 2013; Elliott et Valliant, 2017). La collecte de données en ligne, une plateforme sans base de sondage universelle pour effectuer un échantillonnage probabiliste, représentait près de la moitié des dépenses en recherche par sondage aux États-Unis en 2012 (Terhanian et Bremer, 2012), et il est presque certain qu'elle a pris de l'essor depuis.

Pour de nombreux organismes d'enquête, l'ajustement des poids de sondage sur des données auxiliaires connues est l'étape finale et la plus cruciale du processus de construction des pondérations. Les approches classiques comprennent la poststratification, qui consiste à ajuster les poids de manière que la distribution des variables auxiliaires catégoriques dans l'échantillon pondéré concorde avec leur distribution dans la population, ainsi que son extension à l'estimation par la régression généralisée (GREG), qui fait en sorte que, pour chaque variable auxiliaire (continue ou catégorique), la somme des valeurs soit égale au total correspondant dans la population (Deville et Särndal, 1992). Le calage joue un rôle important en statistique officielle, parce qu'il peut produire des poids tels que les estimations démographiques pondérées obtenues d'après différentes enquêtes concordent.

Sous échantillonnage probabiliste, si les poids de sondage sont égaux à l'inverse des probabilités de sélection, les estimations pondérées des totaux dans l'échantillon sont des estimations sans biais sous le plan de sondage du total de population. Le calage ajuste les poids de sondage de façon minimale afin que les totaux pour les variables auxiliaires dans l'échantillon pondéré concordent avec leurs totaux de population connus (Särndal, Swensson et Wretman, 1992). Dans le contexte d'échantillonnage probabiliste, le calage est introduit pour réduire la variance ou corriger le biais grâce à un ajustement pour tenir compte du sous-dénombrement ou du surdénombrement de sous-groupes de l'échantillon. Dans le cas de grands échantillons, les poids calés finaux peuvent être appliqués à toutes les variables de l'enquête, parce qu'ils retiennent approximativement la propriété d'absence de biais des poids de sondage originaux. Par contre, dans le cas de l'échantillonnage non probabiliste, il n'existe pas de probabilités de sélection pour construire les poids de sondage initiaux qui peuvent produire des estimations sans biais. Donc, il n'est pas garanti que les poids calés habituels puissent fonctionner pour toutes les variables dans l'échantillon non probabiliste. Afin de faire des inférences à partir d'échantillons non probabilistes, une approche pratique consiste à construire un ensemble de poids qui peuvent réduire la racine carrée de l'erreur quadratique moyenne (REQM ou RMSE, de l'anglais *root mean square error*) des estimations pondérées pour un résultat d'intérêt particulier. Le calage assisté par un modèle fournit le cadre pour construire des poids calés ciblant une variable de résultat, étant donné un modèle pouvant approximer les valeurs espérées du résultat (Wu et

Sitter, 2001). La clé d'un calage assisté par un modèle réussi est un modèle possédant de fortes propriétés prédictives, c'est-à-dire que les paramètres du modèle estimés d'après un échantillon peuvent être utilisés pour prédire fiablement les valeurs dans un échantillon différent de la même population. Évidemment, de telles variables explicatives ne sont pas toujours disponibles; Tourangeau, Conrad et Couper (2013) donnent un exemple où le manque de covariables prédictives empêche que les ajustements des pondérations produisent de bons résultats. Cependant, Tourangeau et coll. (2013) s'intéressaient aux enquêtes auprès des ménages. Les variables explicatives peuvent être plus puissantes dans le cas d'enquêtes auprès d'établissements ou d'institutions, ou de certaines enquêtes spécialisées, comme les sondages électoraux. Par exemple, Wang, Rothschild, Goel et Gelman (2015) s'appuient sur l'affiliation politique et le candidat élu lors de l'élection précédente pour faire des prédictions exactes du résultat de l'élection présidentielle américaine de 2012 en se basant sur un échantillon non probabiliste dont la répartition différait considérablement de celle de l'ensemble des électeurs.

Manifestement, on pourrait donc s'attendre à ce que le calage assisté par un modèle soit le plus efficace lorsqu'il existe un ensemble relativement riche de covariables de population auxiliaires et, en conséquence, un très grand ensemble de modèles à prendre en considération. Dans ces conditions, il peut être difficile d'obtenir un équilibre entre la structure – pour minimiser l'erreur de spécification du modèle et donc le biais –, et la parcimonie – pour stabiliser les estimations et donc minimiser la variance. Le *Least Angle Shrinkage and Selection Operator*, ou LASSO, est une régression régularisée qui permet d'effectuer à la fois la sélection des variables et l'estimation des paramètres (Tibshirani, 1996). Une vaste gamme d'applications ont montré que le LASSO prévient bien le surajustement du modèle grâce à la sélection automatique de modèles plus précis et parcimonieux. Kamarianakis, Shen et Wynter (2012) ont réussi, à l'aide du LASSO, à prédire la vitesse de circulation moyenne en présence de multicollinéarités importantes dues à des variables explicatives agrégées au niveau de la région. Kohannim, Hibar, Stein, Jahanshad, Hua, Rajagopalan, Toga, Jack Jr, Weiner, de Zubicaray et McMahon (2012) ont appliqué la régression LASSO pour déterminer des sous-ensembles de polymorphismes de nucléotide unique (SNP) de grandes dimensions et corrélés qui sont associés aux mesures de la structure cérébrale. Dans le cadre d'une revue des défis liés à l'analyse écologique avec des covariables linéaires, Dormann, Elith, Bacher, Buchmann, Carl, Carre, Marquez, Gruber, Lafourcade, Leitao et Mnkemller (2013) ont constaté que le LASSO est l'une des méthodes qui produisent systématiquement de faibles valeurs de la racine carrée de l'erreur quadratique moyenne. Dans le domaine de la génétique et de la finance, le LASSO a été utilisé efficacement en modélisation prédictive comprenant des centaines ou des milliers de variables explicatives (Wu, Chen, Hastie, Sobel et Lange, 2009).

Des formes stabilisantes du calage classique sont prises en considération dans diverses publications. Park et Yang (2008) examinent pour un estimateur par la régression généralisée une forme de type régression ridge comprenant un terme de pénalité pour stabiliser les estimateurs par calage, qui s'avère converger sous le plan de sondage et réduire la variance dans des études en simulation. Goga, Muhammad-Shehzad et Vanheuverzwyn (2011), et Cardot, Goga et Shehzad (2017) ont examiné le calage sur les composantes

principales des totaux de population plutôt que les totaux de population proprement dit, ce qui permet de regrouper un grand nombre de variables auxiliaires en un sous-ensemble de dimension raisonnable. Aspect peut-être le plus pertinent pour les présents travaux, McConville (2011) et McConville, Breidt, Lee et Moisen (2017) ont élaboré, de nouveau sous calage classique, le cadre théorique pour montrer l'absence de biais et la convergence approximatives sous le plan de l'estimateur d'un total par calage avec le LASSO, sachant les estimations des paramètres de la régression LASSO. Bien que le calage assisté par un modèle LASSO soit très prometteur quant à la construction d'un ensemble de poids qui peut résulter en une petite REQM des estimations pondérées pour une variable de résultat dans un échantillon non probabiliste, il n'existe aucun cadre théorique établi pour les propriétés de biais et de convergence des estimateurs par calage assistés par un modèle LASSO pour un échantillon non probabiliste.

Donc, les principaux objectifs du présent article sont :

1) élaborer le cadre théorique pour le calage assisté par un modèle LASSO pour les variables de résultat continues ainsi que binaires, c'est-à-dire calculer l'estimation ponctuelle du total, son espérance asymptotique et l'estimation asymptotique de la variance théorique;
2) étudier la performance relative, en ce qui concerne la racine carrée de l'erreur quadratique moyenne, du calage avec le LASSO par rapport au calage classique sous différents types de résultats, de plans d'échantillonnage, de tailles d'échantillon et de structures de covariance des variables de calage.

Alors que notre développement de la théorie asymptotique repose sur l'hypothèse que les poids de sondage sont connus, une constatation clé est que le calage avec le LASSO donne des estimateurs convergents d'un total de population, que les poids de sondage soient ou non spécifiés correctement, à condition que le modèle de régression contienne tous les paramètres de superpopulation sous forme d'un sous-ensemble des paramètres dans le modèle. Donc, dans les études en simulation, nous nous concentrons sur l'estimation dans des conditions d'échantillonnage non probabiliste, où les poids de sondage initiaux sont considérés comme étant les mêmes que pour l'échantillonnage aléatoire simple (EAS), $d_i = N/n$ pour des tailles de population et d'échantillon $N$ et $n$, indépendamment de la façon dont les échantillons sont formés (qui, en pratique, serait inconnue). Nous appliquons également le calage avec le LASSO à l'estimation du nombre total d'adultes ayant reçu un diagnostic de cancer dans la population américaine, en utilisant des données sur l'incidence du cancer provenant de la *National Health Interview Survey* (NHIS) de 2013 et des données auxiliaires de population provenant de l'*American Community Survey* du *US Census Bureau*, en ne tenant pas compte des poids de sondage pour approximer un échantillon non probabiliste et comparer les résultats aux estimations entièrement pondérées (représentatives).

La présentation de l'article est la suivante. À la section 2, nous donnons la définition et la notation pour le calage et la régression LASSO. À la section 3, nous développons l'estimateur par calage avec le LASSO d'un total de population, son espérance sous le modèle et ses variances asymptotiques. À la section 4, nous décrivons la simulation et les résultats de l'évaluation des estimations de la racine carrée de l'erreur

quadratique moyenne et de la variance de l'estimateur calé par la méthode LASSO. À la section 5, nous présentons l'exemple de la NHIS. Enfin, à la section 6, nous résumons nos constatations en guise de conclusion.

# 2 Calage

## 2.1 Calage classique

Pour un échantillon analytique $s_A$ (l'échantillon qui nécessite un calage des poids) de taille $n$ tiré selon le plan de sondage $\mathcal{A}$ avec les poids de sondage $\underset{n\times 1}{\mathbf{d}}$ et la matrice diagonale des poids de sondage $\mathbf{D}$, les poids calés $\underset{n\times 1}{\mathbf{w}}$ minimisent une mesure de distance

$$E_{\mathcal{A}}\left[\sum_{i\in s_A} g(w_i, d_i)\Big/ q_i\right] \tag{2.1}$$

sous la contrainte :

$$\sum_{i\in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}} \tag{2.2}$$

où $E_{\mathcal{A}}$ est l'espérance sous le plan analytique (probabiliste), $g(w_i, d_i)$ est une fonction différentiable par rapport à $w_i$, strictement convexe sur un intervalle contenant $d_i$, et $g(d_i, d_i) = 0$, et où $\mathbf{T}^X$ est un vecteur ligne de totaux de population connus de variables de calage d'échantillon $\mathbf{X}$ (Deville et Särndal, 1992). La constante $q_i$ est indépendante du poids de sondage $d_i$. L'estimateur par la régression généralisée (GREG) utilisé habituellement fait appel à la distance du khi-deux : $g(w_i, d_i) = (w_i - d_i)^2 \big/ d_i$ avec $q_i = 1$. Sous cette mesure de distance :

$$\mathbf{w}^{\text{GREG}} = \mathbf{d} + \mathbf{DX}\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)^T. \tag{2.3}$$

L'estimation du total de population du résultat $\mathbf{y}$ est fondée sur les poids calés :

$$\begin{aligned}\hat{T}_y^{\text{GREG}} &= \mathbf{w}^{(\text{GREG})T}\mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy} \\ &= \hat{T}_y^{\text{HT}} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}\end{aligned} \tag{2.4}$$

où $\hat{T}_y^{\text{HT}} = \sum_{i\in s_A} d_i y_i$ est l'estimateur fondé sur le plan de sondage (pondéré) classique, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}$ est l'estimation des moindres carrés pondérés de la régression linéaire $E_\xi\left[y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}\right] = \mathbf{x}_i^T\boldsymbol{\beta}$, sachant les poids $\mathbf{D}$. (Cela correspond à l'estimateur poststratifié quand $\mathbf{X}$ consiste entièrement en totaux de cellule pour des variables catégoriques.) Les poids calés définis dans l'équation (2.3) ne dépendent d'aucune variable de résultat. Donc, le même ensemble de poids peut être appliqué à toutes les variables de l'enquête. Notons que la régression GREG suppose un modèle de travail linéaire. Même si $\hat{T}_y^{\text{GREG}}$ est

asymptotiquement sans biais sous le plan pour $T_y$, quand la relation entre $\mathbf{y}$ et $\mathbf{X}$ est non linéaire, comme cela est le cas quand $\mathbf{y}$ est binaire, la variance sous le plan de $\hat{T}_y^{\text{GREG}}$ peut être plus grande que la variance sous le plan de $\hat{T}_y^{\text{HT}}$.

## 2.2 Calage assisté par un modèle

Les estimateurs par calage assisté par un modèle peuvent présenter un avantage important par rapport à l'estimateur $\hat{T}_y^{\text{GREG}}$, parce que le calage assisté par un modèle permet d'utiliser des modèles non linéaires pour aider à construire les poids calés. Dans le calage assisté par un modèle (noté MC, de l'anglais *model-assisted calibration*), nous supposons qu'il existe une relation entre un résultat $\mathbf{y}$ et $\mathbf{X}$ par l'intermédiaire des deux premiers moments (Wu et Sitter, 2001) :

$$E_\xi\left(y_i \,|\, \mathbf{x}_i\right) = \mu\left(\mathbf{x}_i, \boldsymbol{\beta}\right), V_\xi\left(y_i \,|\, \mathbf{x}_i\right) = \nu_i^2\sigma^2 \tag{2.5}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ et $\sigma$ sont les paramètres de superpopulation inconnus, $\mu\left(x_i, \boldsymbol{\beta}\right)$ est une fonction connue de $\mathbf{x}_i$ et $\boldsymbol{\beta}$, et $\nu_i$ est une fonction connue de $\mathbf{x}_i$ ou $\mu\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$. $E_\xi$ et $V_\xi$ sont l'espérance et la variance sous le modèle $\xi$. Soit $\mathbf{B}$ l'estimation en population finie (ou recensement) de $\boldsymbol{\beta}$ (c'est-à-dire l'estimateur de la quasi-vraisemblance de $\boldsymbol{\beta}$ fondé sur la population finie entière), et $\hat{\mu}_i = \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right)$, où $\hat{\mathbf{B}}$ est l'estimation sur échantillon de $\mathbf{B}$. Les poids calés avec assistance d'un modèle $\mathbf{w}$ minimisent alors une mesure de distance $E_{\mathcal{A}}\left[\sum_{i\in s_A} g\left(w_i, d_i\right)/q_i\right]$ sous les contraintes $\sum_{i\in s_A} w_i = N$ et $\sum_{i\in s_A} w_i\hat{\mu}_i = \sum_i^N \hat{\mu}_i$. La principale différence conceptuelle entre le calage classique et le calage assisté par un modèle est que, dans ce dernier, les contraintes sont fondées sur deux quantités, à savoir 1) la taille de la population et 2) le total de population des valeurs prédites $\hat{\mu}_i$. Dans le calage classique, la contrainte est un vecteur de totaux de population de $\mathbf{X}$ (voir l'équation (2.2)). Sous la mesure de distance du khi-deux avec $q_i = 1$, les poids calés avec assistance d'un modèle sont :

$$\mathbf{w}^{\text{MC}} = \mathbf{d} + \mathbf{DM}\left(\mathbf{M}^T\mathbf{DM}\right)^{-1}\left(\mathbf{T}^M - \mathbf{d}^T\mathbf{M}\right)^T \tag{2.6}$$

où $\mathbf{T}^M = \left[N, \sum_i^N \hat{\mu}_i\right]$ et $\mathbf{M} = \left[\mathbf{d}, \left(\hat{\mu}_i\right)_{i\in s_A}\right]$. (Dans le contexte non probabiliste, le vecteur des poids de sondage $\mathbf{d}$ peut être remplacé par $(N/n)\,\mathbf{1}$.) L'estimation du total de population basée sur les poids calés avec assistance d'un modèle est alors :

$$\begin{aligned}\hat{T}_y^{\text{MC}} &= \left(\mathbf{w}^{\text{MC}}\right)^T\mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{M}} - \mathbf{d}^T\mathbf{M}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy} \\ &= \hat{T}_y^{\text{HT}} + \left(\sum_i^N \hat{\mu}_i - \sum_{i\in s_A} d_i\hat{\mu}_i\right)\hat{B}^{\text{MC}}\end{aligned} \tag{2.7}$$

où $\hat{B}^{\text{MC}}$ est la pente de calage qui satisfait les contraintes de calage (différente des estimations des paramètres du modèle $\hat{\mathbf{B}}$) :

$$\hat{B}^{\mathrm{MC}} = \frac{\sum_{i\in s_A} d_i \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)(y_i - \bar{y})}{\sum_{i\in s_A} d_i \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)^2}, \quad \hat{\bar{\mu}} = \sum_{i\in s_A} d_i \hat{\mu}_i \Big/ \sum_{i\in s_A} d_i, \quad \bar{y} = \sum_{i\in s_A} d_i y_i \Big/ \sum_{i\in s_A} d_i .$$

L'absence de biais et la petite variance de $\hat{T}_y^{\mathrm{MC}}$ dépendent toutes deux de la mesure dans laquelle les $\hat{\mu}_i$ sont de bonnes approximations de la vraie valeur espérée de $y_i$.

# 3 Sélection du modèle et calage robuste avec le LASSO adaptatif

## 3.1 Contexte du LASSO adaptatif

### 3.1.1 Définition et paramètres

Les coefficients de la régression de type LASSO adaptatif s'obtiennent en résolvant une équation de régression pénalisée. Pour la régression linéaire pénalisée par le LASSO adaptatif (Zou, 2006) :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left( \sum_{i\in s_A} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \mid \beta_j \mid \right) \tag{3.1}$$

où $\alpha_j^{\gamma}$ est un poids ajustable et $\lambda_n$ est une pénalité utilisée pour optimiser une mesure de l'adéquation du modèle. Similairement, pour la régression logistique pénalisée par le LASSO adaptatif :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left( \sum_{i\in s_A} [-y_i (\mathbf{x}_i^T \beta) + \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}))] + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \mid \beta_j \mid \right). \tag{3.2}$$

Sachant $\lambda_n$ et $\gamma$, nous pouvons calculer $\hat{\boldsymbol{\beta}}$ par des procédures itératives. Le package *glmnet* de R calcule les estimateurs LASSO adaptatif linéaire ainsi que logistique (Friedman, Hastie et Tibshirani, 2010).

Le rôle du paramètre de pondération, $\alpha_j$, est d'empêcher le LASSO de sélectionner des covariables dont la taille de l'effet est grande en faveur d'une réduction de l'erreur de prédiction quand la taille d'échantillon est petite. Donc, les poids sont inversement proportionnels aux tailles des effets des paramètres de régression : $\alpha_j \propto 1 / |\beta_j|$. Un choix fréquent de $\alpha_j$ est $1 / |\hat{\beta}_j^{\mathrm{EMV}}|$, où $\hat{\beta}_j^{\mathrm{EMV}}$ est l'estimation du maximum de vraisemblance de $\beta_j$. La puissance du paramètre de poids, $\gamma$, est une constante plus grande que 0 qui interagit avec $\alpha_j$ pour empêcher le LASSO de sélectionner ou d'exclure des paramètres. Par exemple, si nous souhaitons quand même que le LASSO favorise les covariables dont l'effet est grand quand la taille d'échantillon est petite, nous devons choisir une petite valeur de $\gamma$. Si nous voulons accorder encore moins d'importance aux tailles des effets, nous devons choisir une grande valeur de $\gamma$.

### 3.1.2 Propriété d'oracle

La mesure de la performance d'une méthode de sélection de modèles et d'estimation s'appuie sur un concept important appelé la « propriété d'oracle ». La méthode optimale sélectionne les bonnes variables et

fournit des estimations sans biais des paramètres choisis. Supposons que, dans un modèle de régression complet, les paramètres ont à la fois des composantes nulles et non nulles. Sans perte de généralité, considérons que les $p$ premières sont non nulles et que les $q$ dernières sont nulles :

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} = \mathbf{0} \end{pmatrix}.$$

Un modèle de régression possède la propriété d'oracle s'il satisfait les conditions suivantes (Fan et Li, 2001) :

- la probabilité d'obtenir l'estimation 0 pour des paramètres dont la valeur est nulle tend vers un : $\Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \to 1$ quand $n \to \infty$;
- les estimations des paramètres non nuls sont aussi bonnes que si le sous-modèle réel était connu : $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \to N(\mathbf{0}, \mathbf{C})$ où $\mathbf{C} = \Sigma\left(\boldsymbol{\beta}^{(1)}\right)$ est la matrice de covariance de $\boldsymbol{\beta}^{(1)}$ sous un modèle linéaire, et $\mathbf{C} = I^{-1}\left(\boldsymbol{\beta}^{(1)}\right)$ est l'inverse de la matrice d'information de Fisher de $\boldsymbol{\beta}^{(1)}$ sous le modèle linéaire généralisé.

Pour l'inférence en population finie, supposons que l'indice $\nu$ désigne une population de taille $N_\nu$, représentons par $\mathbf{B}_\nu$ les estimations de la quasi-vraisemblance de $\boldsymbol{\beta}$ dans la population $\nu$, et par $\hat{\mathbf{B}}_\nu$ l'estimation de $\mathbf{B}_\nu$ basée sur un échantillon de taille $n_\nu \le N_\nu$. Nous supposons que $N_\nu \to \infty$, $n_\nu \to \infty$ et $n_\nu / N_\nu \to 0$ quand $\nu \to \infty$. L'équivalent en population finie de la propriété d'oracle est alors :

$$\begin{aligned} \Pr\left(\hat{\mathbf{B}}^{(2)}_\nu = \mathbf{0}\right) &\to 1 \\ \sqrt{n_\nu}\left(\hat{\mathbf{B}}^{(1)}_\nu - \mathbf{B}^{(1)}_\nu\right) &\to N_\nu(\mathbf{0}, \mathbf{C}_\nu) \\ \mathbf{B}_\nu &\to \boldsymbol{\beta} \\ \text{quand } \nu \to \infty & \end{aligned}$$

où $\mathbf{C}_\nu = \Sigma\left(\mathbf{B}^{(1)}_\nu\right)$ est la matrice de covariance de $\mathbf{B}^{(1)}_\nu$ si le modèle est linéaire, et $\mathbf{C}_\nu = I^{-1}\left(\mathbf{B}^{(1)}_\nu\right)$ est l'inverse de la matrice d'information de Fisher de $\mathbf{B}^{(1)}_\nu$ sous le modèle linéaire généralisé.

Zou (2006) a montré que si $\lambda_n \Big/ \left(\sqrt{n}\Big/\left(\sqrt{n}\right)^\gamma\right) \to \infty$ et $\lambda_n / \sqrt{n} \to 0$, alors le LASSO adaptatif satisfait la propriété d'oracle. Les conditions requièrent que $\lambda_n$ augmente au moins à la vitesse de $\sqrt{n}\Big/\left(\sqrt{n}\right)^\gamma$, mais pas plus rapidement que $\sqrt{n}$. Le choix de $\lambda_n$ et $\gamma$, et le code R pour l'implémenter sont discutés en annexe.

## 3.2 Calage avec le LASSO

À la présente section, nous établissons les formules analytiques pour un estimateur LASSO du total, son espérance sous le modèle et les estimateurs asymptotiques des variances sous le plan de sondage. Nous faisons les hypothèses suivantes :

1. Les échantillons sont tirés selon un plan d'échantillonnage à un seul degré $\mathcal{A}$, en permettant des probabilités de sélection inégales. La probabilité de sélection de l'unité $i$ est notée $\pi_i^A$, et la probabilité de sélection conjointe des unités $i$ et $j$ est notée $\pi_{ij}^A$. Nous désignons le poids de sondage pour l'unité $i$ par $d_i^A = 1/\pi_i^A$, le vecteur des poids de sondage par $\mathbf{d}^A$, et la matrice diagonale des poids de sondage par $\mathbf{D}^A$.
2. Les données auxiliaires au niveau de la population sont connues et notées $\mathbf{X} = (\mathbf{x}_i^T)$, $i = 1, \ldots, N$.
3. Un modèle de superpopulation est considéré, comme il est décrit à la section 2.2 :

$$E_\xi(y_i \mid \mathbf{x}_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta})$$
$$V_\xi(y_i \mid \mathbf{x}_i) = v_i^2 \sigma^2.$$

4. Les paramètres de superpopulation réels sont un sous-ensemble du modèle de régression complet pour le LASSO :

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}.$$

5. La probabilité d'observer dans l'échantillon analytique l'étendue complète de $\mathbf{X}$ dans la population est non nulle.

### 3.2.1 Estimation ponctuelle : $\hat{T}_y^{\text{LASSO}}$

L'estimation par calage avec le LASSO du total peut être obtenue par les étapes suivantes :

1. Obtenir les coefficients de régression LASSO $\hat{\mathbf{B}}$ comme il est décrit à l'annexe.
2. Utiliser $\hat{\mathbf{B}}$ pour calculer $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$ dans la population.
3. Définir $\mathbf{T}^M = \left(N, \sum_i^N \hat{\mu}_i\right)$ et $\mathbf{M} = \left[\mathbf{d}^A, \sum_{i\in s_A} \hat{\mu}_i\right]$ sous la mesure de distance du khi-deux avec $q_i = 1$ :

$$\mathbf{w}^{\text{LASSO}} = \mathbf{d}^A + \mathbf{D}^A\mathbf{M}(\mathbf{M}^T\mathbf{D}^A\mathbf{M})^{-1}\left(\mathbf{T}^M - (\mathbf{d}^A)^T\mathbf{M}\right)^T. \tag{3.3}$$

4. Déterminer l'estimateur par calage avec le LASSO du total :

$$\begin{aligned} \hat{T}_y^{\text{LASSO}} &= (\mathbf{w}^{\text{LASSO}})^T\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{M}} - (\mathbf{d}^A)^T\mathbf{M}\right)(\mathbf{X}^T\mathbf{D}^A\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^A\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\sum_i^N \hat{\mu}_i - \sum_{i\in s_A} d_i^A\hat{\mu}_i\right)\hat{B}^{\text{MC}} \end{aligned} \tag{3.4}$$

où $\hat{B}^{\text{MC}}$ est la pente de calage qui satisfait les contraintes de calage :

$$\hat{B}^{\mathrm{MC}} = \frac{\sum_{i\in s_A} d_i^A \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)^2},\ \hat{\bar{\mu}} = \sum_{i\in s_A} d_i^A \hat{\mu}_i \Big/ \sum_{i\in s_A} d_i^A,\ \bar{y} = \sum_{i\in s_A} d_i^A y_i \Big/ \sum_{i\in s_A} d_i^A.$$

### 3.2.2 Comportement asymptotique de $\hat{T}_y^{\mathrm{LASSO}}$

Wu et Sitter (2001) ont établi les conditions pour calculer un estimateur asymptotique par calage assisté par un modèle. Nous énonçons les conditions ici en modifiant légèrement les notations par souci de cohérence avec les présents travaux de recherche. Soit $\boldsymbol{\beta}$ le paramètre de superpopulation réel pour le modèle défini en l'équation (2.5), et $\mathbf{B}$ l'estimateur de la quasi-vraisemblance en population finie de $\boldsymbol{\beta}$. Les conditions qui suivent sont utilisées pour obtenir estimateur asymptotique par calage avec le LASSO du total :

1. $\hat{\mathbf{B}} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$, où $\mathbf{B}$ est la pente de régression en population finie de $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$.
2. Pour chaque $\mathbf{x}_i$, $\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}$ est continue en $\mathbf{t}$, et $\max_i \left|\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}\right| \le h(\mathbf{x}_i, \boldsymbol{\beta})$ pour $\mathbf{t}$ dans un voisinage de $\boldsymbol{\beta}$, et $N^{-1}\sum_{i\in U} h(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.
3. Pour chaque $\mathbf{x}_i$, $\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}\partial\mathbf{t}^T$ est continue en $\mathbf{t}$, et $\max_{j,k} \left|\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial t_j \partial t_k\right| \le k(\mathbf{x}_i, \boldsymbol{\beta})$ pour $\mathbf{t}$ dans un voisinage de $\boldsymbol{\beta}$, et $N^{-1}\sum_{i\in U} k(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.
4. Les estimateurs de Horvitz-Thompson (HT) de certaines moyennes de population suivent asymptotiquement une loi normale (Fuller, 2009; pages 47 à 57).
5. $\lambda_n \Big/ \left(\sqrt{n}\Big/\left(\sqrt{n}\right)^{\gamma}\right) \to \infty$ et $\lambda_n/\sqrt{n} \to 0$.

**Lemme 1** : Supposons que le modèle de superpopulation (2.5) est vérifié. Soit $\mathbf{B}$ l'estimation de la quasi-vraisemblance en population finie de $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. Sous les conditions (1) à (5), l'estimateur asymptotique assisté par un modèle du total de population est :

$$\hat{T}_y^{\mathrm{MC}} = \sum_{i\in s_A} d_i^A \left(y_i - \mu_i B^{\mathrm{MC}}\right) + \sum_{i=1}^{N} \mu_i B^{\mathrm{MC}} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{3.5}$$

où

$$\mu_i = \mu(\mathbf{x}_i, \mathbf{B})$$

$$B^{\mathrm{MC}} = \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}.$$

*Preuve.* Voir l'annexe.

Étant donné le lemme 1, dans le théorème 1, nous obtenons $\hat{T}_y^{\mathrm{LASSO}}$, l'estimateur LASSO asymptotique du total. Dans le théorème 2, nous montrons que $\hat{T}_y^{\mathrm{LASSO}}$ est sans biais sous le modèle pour le total de population. Enfin, dans le théorème 3, nous déterminons les estimations de la variance pour l'estimateur d'un total par calage avec le LASSO.

**Théorème 1** : Supposons que les paramètres d'un modèle de régression complet comprennent des composantes nulles ainsi que non nulles. Sans perte de généralité, posons que les $p$ premières sont non nulles et que les $q$ dernières sont nulles :

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}, \quad \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} \quad \text{et} \quad \boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)},$$

sous les conditions (1) à (5), l'estimateur asymptotique du total par calage avec le LASSO est :

$$\hat{T}_y^{\text{LASSO}} = \sum_{i\in s_A} d_i^A \left(y_i - \mu_i B^{\text{MC}}\right) + \sum_{i=1}^{N} \mu_i B^{\text{MC}} + o_p\left(\frac{N}{\sqrt{n}}\right).$$

*Preuve.* Voir l'annexe.

**Théorème 2** : $\hat{T}_y^{\text{LASSO}}$ est sans biais sous le modèle, c'est-à-dire que $E_\xi(\hat{T}_y^{\text{LASSO}}) = T$.

*Preuve.* Voir l'annexe.

Donc, à condition que les paramètres de la régression LASSO incluent les paramètres de superpopulation, $\hat{T}_y^{\text{LASSO}}$ est sans biais sous le modèle, quels que soient les poids de sondage. (Notons qu'il s'agit d'une qualité que $\hat{T}_y^{\text{GREG}}$ partage avec $\hat{T}_y^{\text{LASSO}}$. Cependant, $\hat{T}_y^{\text{LASSO}}$ peut accepter des modèles contenant de beaucoup plus grands nombres de covariables que $\hat{T}_y^{\text{GREG}}$). Cette propriété est essentielle dans le cas des échantillons non probabilistes, pour lesquels il n'existe pas de poids de sondage initiaux pour garantir l'absence de biais.

**Théorème 3** : L'estimateur asymptotique de la variance de $\hat{T}_y^{\text{LASSO}}$ est donné par

$$\begin{aligned} v_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) &= \sum_{i\in s_A}\left(\frac{y_i - \hat{\mu}_i\hat{B}^{\text{MC}}}{\pi_i}\right)^2 (1-\pi_i) \\ &\quad + \sum_{i\in s_A}\sum_{j\neq i} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{\left(y_i - \hat{\mu}_i\hat{B}^{\text{MC}}\right)}{\pi_i} \frac{\left(y_j - \hat{\mu}_j\hat{B}^{\text{MC}}\right)}{\pi_j}. \end{aligned} \tag{3.6}$$

*Preuve.* La variance sous le plan théorique de l'estimateur LASSO est

$$\begin{aligned} V_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) &= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{\text{MC}}\right) + \sum_{i=1}^{N}\mu_i B^{\text{MC}}\right) \\ &= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{\text{MC}}\right)\right) \\ &= \sum_{i\in U}\left(\frac{y_i - \mu_i B^{\text{MC}}}{\pi_i}\right)^2 \pi_i(1-\pi_i) \\ &\quad + \sum_{i\in U}\sum_{j\neq i}(\pi_{ij} - \pi_i\pi_j)\frac{\left(y_i - \mu_i B^{\text{MC}}\right)}{\pi_i}\frac{\left(y_j - \mu_j B^{\text{MC}}\right)}{\pi_j} \end{aligned} \tag{3.7}$$

qui découle de l'équation (3.30) obtenue pour la variance de l'estimateur par calage avec le LASSO classique d'un total dans McConville (2011). L'équation (3.6) s'obtient en remplaçant les quantités de population par les estimations.

Une autre estimation de la variance, proposée par Särndal, Swensson et Wretman (1989), consiste à multiplier $(y_i - \hat{\mu}_i \hat{B}^{\text{MC}})$ par les poids $g$, qui sont les ratios des poids calés aux poids de sondage originaux :

$$\mathbf{g} = \mathbf{1}_{(n\times 1)} + \mathbf{M}\left(\mathbf{M}^T \mathbf{D}^A \mathbf{M}\right)^{-1}\left(\mathbf{T}^M - \left(\mathbf{d}^A\right)^T \mathbf{M}\right)^T$$

$$v.g_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) = \sum_{i\in s_A}\left(\frac{g_i\left(y_i - \hat{\mu}_i \hat{B}^{\text{MC}}\right)}{\pi_i}\right)^2 (1-\pi_i) + \sum_{i\in s_A}\sum_{j\neq i}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{g_i\left(y_i - \hat{\mu}_i \hat{B}^{\text{MC}}\right)}{\pi_i}\frac{g_j\left(y_j - \hat{\mu}_j \hat{B}^{\text{MC}}\right)}{\pi_j}. \quad (3.8)$$

Pour simplifier la notation, nous représentons $v_{\mathcal{A}}(\hat{T}_y^{\text{LASSO}})$ par $v^{\text{LASSO}}$ et $v.g_{\mathcal{A}}(\hat{T}_y^{\text{LASSO}})$ par $v_g^{\text{LASSO}}$.

# 4 Étude en simulation

Nous concevons une simulation pour évaluer les propriétés en échantillon fini de $\hat{T}_y^{\text{LASSO}}$ et les estimations asymptotiques de la variance de $\hat{T}_y^{\text{LASSO}}$, $v^{\text{LASSO}}$ et $v_g^{\text{LASSO}}$. Nous considérons aussi un estimateur bootstrap naïf $v_{\text{boot}}^{\text{LASSO}}$, obtenu en tirant avec remise 500 échantillons de chaque échantillon simulé, en tant qu'estimateur de rechange de la variance de $\hat{T}_y^{\text{LASSO}}$.

Afin de simuler des échantillons non probabilistes, nous générons des échantillons avec probabilités de sélection inégales, mais fixons les poids de sondage à $\mathbf{d}^A = N/n$. Nous considérons aussi $\hat{T}_y^{\text{GREG}}$ (estimateur par calage classique) et $\hat{T}_y^{\text{HT}}$ (estimateur de Horvitz-Thompson purement fondé sur le plan de sondage). Puisque $\hat{T}_y^{\text{LASSO}}$ effectue à la fois la sélection de variables et l'estimation, nous appliquons une sélection pas à pas descendante (*backward stepwise selection*) pour sélectionner le modèle de travail pour la régression GREG. Bien qu'il n'existe aucune justification théorique de l'utilisation de la sélection pas-à-pas des variables, Skinner et Silva (1997) ont montré que, étant donné deux variables auxiliaires, une procédure pas à pas peut aboutir à une plus grande efficacité de l'estimateur GREG. Nous voulons connaître la performance de chaque estimateur sous 1) des populations dont les rapports signal-bruit (RSB) diffèrent, 2) des plans d'échantillonnage indépendants, informatifs et biaisés, et 3) des petites et grandes tailles d'échantillon. Le ratio signal-bruit est calculé selon les définitions données dans Czanner, Sarma, Eden et Brown (2008). Nous prenons deux niveaux de corrélation (faible/forte) entre les covariables, recoupés par deux niveaux de taille de l'effet (petite/grande) des covariables. Nous configurons les populations faible/grande et forte/petite de manière qu'elles aient le même RSB, afin de comprendre l'influence de la corrélation et de la taille de l'effet sur la performance de l'estimateur, étant donné le même RSB. Trois plans d'échantillonnage sont utilisés pour tirer les échantillons, à savoir l'échantillonnage aléatoire simple sans

remise, l'échantillonnage aléatoire simple (EAS), l'échantillonnage de Poisson avec probabilités de sélection proportionnelles aux covariables, $\text{POI}(\text{X})$, et l'échantillonnage de Poisson avec probabilités de sélection proportionnelles aux covariables et au résultat, $\text{POI}(\text{X+Y})$. L'échantillonnage $\text{POI}(\text{X+Y})$ simule le biais d'autosélection des échantillons non probabilistes, où la propension d'un répondant à participer à une étude est reliée à la variable d'analyse. Nous considérons deux tailles d'échantillon, soit 250 et 1 000. Donc, nous avons un total de $2 \times 2 \times 3 \times 2 = 24$ groupes expérimentaux.

## 4.1 Population

Pour créer une collinéarité entre les covariables, nous suivons une structure d'autocorrélation décroissante fréquemment utilisée dans les simulations liées au LASSO (Tibshirani, 1996) : $\text{cor}(X_i, X_j) = \rho^{|i-j|}$, $i = 1, \ldots, p$. Nous générons une population de taille $N = 100\,000$ à partir d'une loi normale multivariée de moyenne $\mathbf{0}_{(p\times 1)}$ et de covariance $\Sigma^{\rho}$, $p = 40$. La variable de résultat continue est générée par le modèle de régression :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{40} x_{i40} + N(0, 3).$$

La variable de résultat binaire est générée par le modèle de régression logistique :

$$\begin{aligned} \phi_i &= \text{expit}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{40} x_{i40}), \qquad \text{expit}(u) = (1 + \exp(u))^{-1} \\ y_i &= \text{bernoulli}(\phi_i). \end{aligned}$$

Nous prenons $\rho = 0{,}15$ pour la population avec faible corrélation, et $\rho = 0{,}73$ pour la population avec forte corrélation. Pour les variables de résultat continues ainsi que binaires :

$$\begin{aligned} &\text{Petite taille de l'effet} \quad \boldsymbol{\beta}^{(1)} := \beta_{12} \ldots \beta_{19}, \beta_{32} \ldots \beta_{39} = 0{,}45 \\ &\text{Grande taille de l'effet} \quad \boldsymbol{\beta}^{(1)} := \beta_{12} \ldots \beta_{19}, \beta_{32} \ldots \beta_{39} = 0{,}74. \end{aligned}$$

Pour $\mathbf{y}$ continue : $\beta_0 = 1$, pour $\mathbf{y}$ binaire : $\beta_0 = 0{,}4$. Le reste des $\beta_i = 0$. Sur 41 paramètres de régression, 16 sont non nuls et 25 sont nuls.

## 4.2 Plans d'échantillonnage

Trois plans d'échantillonnage sont utilisés pour générer l'échantillon :

1. Échantillonnage aléatoire simple (EAS) : probabilités de sélection $= n/N$.
2. Échantillonnage de Poisson avec probabilités proportionnelles à $\mathbf{X}$, $\text{POI}(\text{X})$

$$\begin{cases} \mathbf{y} \text{ continue : } \pi_i \propto 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35} \\ \mathbf{y} \text{ binaire : } \text{logit}(\pi_i) = 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35}. \end{cases}$$

3. Échantillonnage de Poisson avec probabilités proportionnelles à $\mathbf{X}$ et $\mathbf{y}$, $\text{POI}(\text{X+Y})$

$$\begin{cases} \mathbf{y} \text{ continue : } \pi_i \propto 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35} + 0{,}5y_i \\ \mathbf{y} \text{ binaire : } \text{logit}(\pi_i) \propto 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35} + y_i. \end{cases}$$

## 4.3 Mesures d'évaluation

Nous évaluons le biais, la variance et la REQM empiriques pour chaque estimateur du total. Nous évaluons les estimations asymptotiques de la variance et les estimations bootstrap de la variance d'après la couverture nominale de 95 % et le biais en pourcentage par rapport à la variance empirique. Nous utilisons l'approximation normale pour générer les intervalles de confiance. Nous calculons le biais en pourcentage comme étant $\text{biais en \%} = 100[v - \text{var}(\hat{T}_y^{\text{LASSO}})] / \text{var}(\hat{T}_y^{\text{LASSO}})$, où $\text{var}(\hat{T}_y^{\text{LASSO}})$ est la variance empirique obtenue à partir des échantillons simulés.

## 4.4 Résultats des simulations

Les résultats des simulations portent sur $S = 1\,000$ échantillons simulés par groupe expérimental. Le tableau 4.1 donne les résultats numériques pour le biais, la variance et la racine carrée de l'erreur quadratique moyenne de chaque estimateur sous différents plans expérimentaux en vue d'estimer le total d'une variable de résultat continue. Le tableau 4.2 donne les résultats numériques pour l'estimation du total d'une variable de résultat binaire.

### 4.4.1 Racine carrée de l'erreur quadratique moyenne

Sous EAS, tous les estimateurs sont sans biais, et les estimateurs LASSO et GREG donnent des résultats approximativement aussi bon que l'estimateur HT. Les plans $\text{POI}(\text{X})$ et $\text{POI}(\text{X+Y})$ donnent des échantillons biaisés en raison de la sélection avec probabilités plus élevées de cas présentant de plus grandes valeurs de covariables. Sous $\text{POI}(\text{X+Y})$, la sélection favorise aussi les cas présentant de plus grandes valeurs de la variable de résultat. Le biais absolu de l'estimateur LASSO diminue relativement à celui de l'estimateur GREG à mesure qu'augmente le RSB. Cette amélioration est plus spectaculaire dans le cas binaire que dans le cas continu, surtout pour $\text{POI}(\text{X+Y})$. Pour ce qui est de la REQM, l'estimateur LASSO offre une amélioration insignifiante par rapport à l'estimateur GREG pour l'estimation des totaux de variables de résultat continues. L'amélioration est légèrement notable, environ 3 %, lorsque le modèle contient des variables explicatives fortement corrélées. Dans le cas de variables binaires, l'estimateur LASSO donne lieu à une amélioration importante de l'EQM par rapport à l'estimateur GREG à mesure qu'augmente le RSB, la réduction étant de 20 % pour le plan $\text{POI}(\text{X})$ et de près de 50 % pour le plan $\text{POI}(\text{X+Y})$ quand le RSB est grand. En particulier, sous les types de population faible/grande et forte/petite, le RSB est le même, de sorte que la différence de performance entre les estimateurs LASSO et GREG est attribuée à la corrélation ou à la taille de l'effet. Le LASSO donne de meilleurs résultats en ce qui concerne le biais et la REQM pour le type de population forte/petite, ce qui donne à penser que l'avantage de l'estimateur LASSO par rapport à l'estimateur GREG est plus important lorsque le modèle contient des variables explicatives fortement corrélées. Il semble donc qu'en présence d'une multicollinéarité, la capacité de sélectionner les variables du LASSO est meilleure que celle de la procédure pas-à-pas utilisée dans la méthode GREG.

**Tableau 4.1**
**Résumé des simulations pour un résultat continu : total, biais et REQM × $10^3$; variance × $10^6$**

| Population | n | Plan d'échantillonnage | HT | | | GREG | | | LASSO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | biais | var | reqm | biais | var | reqm | biais | var | reqm |
| Faible/petite T = 100,8 RSB = 0,47 | 250 | EAS | 0,5 | 546 | 23,3 | 0,9 | 425 | 20,6 | 0,9 | 428 | 20,7 |
| | | POI(X) | 12,4 | 525 | 26,0 | -0,6 | 446 | 21,1 | -0,4 | 441 | 21,0 |
| | | POI(X+Y) | 19,4 | 519 | 29,9 | 4,6 | 443 | 21,5 | 4,7 | 431 | 21,3 |
| | 1 000 | EAS | 0,2 | 129 | 11,4 | 0,3 | 94 | 9,6 | 0,3 | 94 | 9,7 |
| | | POI(X) | 12,6 | 129 | 17,0 | -0,1 | 91 | 9,5 | -0,2 | 92 | 9,6 |
| | | POI(X+Y) | 19,7 | 128 | 22,7 | 4,9 | 91 | 10,7 | 5,0 | 91 | 10,7 |
| Faible/grande T = 101,4 RSB = 1,26 | 250 | EAS | 0,4 | 849 | 29,1 | 0,9 | 415 | 20,4 | 1,0 | 417 | 20,4 |
| | | POI(X) | 21,1 | 818 | 35,6 | -1,3 | 434 | 20,9 | -1,0 | 432 | 20,8 |
| | | POI(X+Y) | 31,7 | 817 | 42,7 | 3,7 | 427 | 21,0 | 4,0 | 427 | 21,1 |
| | 1 000 | EAS | 0,0 | 200 | 14,1 | 0,3 | 94 | 10,0 | 0,3 | 93 | 9,7 |
| | | POI(X) | 21,1 | 199 | 25,4 | -0,1 | 91 | 9,6 | -0,2 | 90 | 9,6 |
| | | POI(X+Y) | 31,7 | 196 | 34,6 | 4,9 | 91 | 10,7 | 4,8 | 89 | 10,6 |
| Forte/petite T = 101,8 RSB = 1,26 | 250 | EAS | 0,1 | 941 | 30,7 | 1,0 | 421 | 20,6 | 1,0 | 399 | 20,0 |
| | | POI(X) | 50,2 | 895 | 58,5 | -0,7 | 434 | 20,8 | -1,6 | 402 | 20,1 |
| | | POI(X+Y) | 57,8 | 872 | 64,9 | 4,0 | 435 | 21,2 | 3,0 | 399 | 20,2 |
| | 1 000 | EAS | 0,0 | 218 | 14,8 | 0,3 | 94 | 9,7 | 0,3 | 93 | 9,6 |
| | | POI(X) | 50,6 | 210 | 53,0 | -0,1 | 93 | 9,7 | -0,5 | 91 | 9,6 |
| | | POI(X+Y) | 58,2 | 209 | 59,9 | 4,7 | 95 | 10,8 | 4,2 | 92 | 10,5 |
| Forte/grande T = 103,1 RSB = 3,41 | 250 | EAS | -0,4 | 1 897 | 43,6 | 0,8 | 436 | 20,9 | 1,0 | 407 | 20,2 |
| | | POI(X) | 83,3 | 1 826 | 93,7 | -0,8 | 435 | 20,9 | -1,5 | 406 | 20,2 |
| | | POI(X+Y) | 96,4 | 1 779 | 105,3 | 3,7 | 428 | 21,0 | 3,0 | 404 | 20,3 |
| | 1 000 | EAS | -0,2 | 444 | 21,0 | 0,3 | 93 | 9,7 | 0,3 | 93 | 9,7 |
| | | POI(X) | 83,6 | 424 | 86,1 | -0,2 | 93 | 9,7 | -0,5 | 91 | 9,6 |
| | | POI(X+Y) | 96,9 | 423 | 99,0 | 4,4 | 94 | 10,6 | 4,1 | 92 | 10,4 |

**Tableau 4.2**
**Résumé des simulations pour un résultat binaire : total, biais et REQM × $10^3$; variance × $10^6$**

| Population | n | Plan d'échantillonnage | HT | | | GREG | | | LASSO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | biais | var | reqm | biais | var | reqm | biais | var | reqm |
| Faible/petite T = 56,2 RSB = 0,51 | 250 | EAS | 0,0 | 10,2 | 3,2 | 0,0 | 7,2 | 2,7 | 0,0 | 7,0 | 2,7 |
| | | POI(X) | 2,6 | 10,0 | 4,1 | 0,2 | 8,0 | 2,8 | 0,1 | 7,8 | 2,8 |
| | | POI(X+Y) | 4,9 | 9,8 | 5,8 | 2,0 | 8,1 | 3,5 | 1,8 | 7,8 | 3,3 |
| | 1 000 | EAS | -0,0 | 2,7 | 1,6 | 0,0 | 1,7 | 1,3 | 0,0 | 1,6 | 1,3 |
| | | POI(X) | 2,5 | 2,4 | 2,9 | 0,0 | 1,8 | 1,3 | -0,0 | 1,7 | 1,3 |
| | | POI(X+Y) | 4,7 | 2,3 | 5,0 | 1,8 | 1,8 | 2,2 | 1,6 | 1,7 | 2,1 |
| Faible/grande T = 54,4 RSB = 1,10 | 250 | EAS | -0,0 | 10,8 | 3,3 | 0,0 | 6,1 | 2,5 | 0,1 | 5,4 | 2,3 |
| | | POI(X) | 3,0 | 10,2 | 4,4 | 0,1 | 6,1 | 2,5 | 0,1 | 5,8 | 2,4 |
| | | POI(X+Y) | 5,3 | 9,8 | 6,2 | 1,6 | 6,2 | 2,9 | 1,3 | 5,8 | 2,8 |
| | 1 000 | EAS | -0,0 | 2,7 | 1,6 | 0,0 | 1,3 | 1,1 | 0,0 | 1,1 | 1,0 |
| | | POI(X) | 2,9 | 2,4 | 3,3 | 0,0 | 1,4 | 1,2 | -0,1 | 1,2 | 1,1 |
| | | POI(X+Y) | 5,2 | 2,2 | 5,4 | 1,4 | 1,4 | 1,8 | 1,1 | 1,2 | 1,6 |
| Forte/petite T = 54,2 RSB = 1,10 | 250 | EAS | -0,0 | 10,3 | 3,2 | 0,0 | 5,8 | 2,4 | 0,1 | 4,9 | 2,2 |
| | | POI(X) | 6,6 | 9,6 | 7,3 | 0,3 | 6,2 | 2,5 | -0,2 | 4,8 | 2,2 |
| | | POI(X+Y) | 8,6 | 9,3 | 9,1 | 1,8 | 6,3 | 3,1 | 0,9 | 4,9 | 2,4 |
| | 1 000 | EAS | -0,0 | 2,5 | 1,6 | 0,0 | 1,2 | 1,1 | 0,0 | 1,0 | 1,0 |
| | | POI(X) | 6,6 | 2,2 | 6,7 | 0,2 | 1,4 | 1,2 | -0,2 | 1,1 | 1,1 |
| | | POI(X+Y) | 8,5 | 2,1 | 8,7 | 1,6 | 1,4 | 2,0 | 1,0 | 1,0 | 1,4 |
| Forte/grande T = 52,8 RSB = 2,75 | 250 | EAS | -0,1 | 10,2 | 3,1 | -0,0 | 5,2 | 2,3 | 0,1 | 3,8 | 1,9 |
| | | POI(X) | 7,1 | 9,8 | 7,8 | 0,3 | 5,7 | 2,4 | -0,2 | 3,6 | 1,9 |
| | | POI(X+Y) | 9,1 | 9,4 | 9,6 | 1,5 | 5,7 | 2,8 | 0,5 | 3,7 | 2,0 |
| | 1 000 | EAS | -0,1 | 2,5 | 1,6 | -0,0 | 1,1 | 1,0 | 0,0 | 0,6 | 0,8 |
| | | POI(X) | 7,1 | 2,2 | 7,2 | 0,2 | 1,3 | 1,1 | -0,2 | 0,7 | 0,9 |
| | | POI(X+Y) | 9,1 | 2,2 | 9,2 | 1,4 | 1,2 | 1,8 | 0,5 | 0,7 | 1,0 |

### 4.4.2 Estimations de la variance de l'estimateur LASSO

Les tableaux 4.3 et 4.4 donnent les résultats pour la couverture nominale de 95 % des IC et le biais en pourcentage pour les deux estimateurs de variance asymptotiques de forme analytique établis dans le cadre de la présente étude, ainsi que l'estimateur de variance bootstrap naïf pour l'estimateur par calage avec le LASSO.

Pour les variables de résultat continues, la couverture des variances bootstrap est systématiquement proche de 95 % sous les plans d'échantillonnage EAS et $\mathrm{POI}(\mathrm{X})$ pour les deux tailles d'échantillon. Sous le plan $\mathrm{POI}(\mathrm{X+Y})$, on note une sous-couverture très modeste dans le tableau 4.3. La couverture des variances de forme analytique est sensible à la taille d'échantillon ainsi qu'au plan d'échantillonnage, les petits échantillons ayant tendance à produire une sous-couverture, particulièrement sous le plan $\mathrm{POI}(\mathrm{X+Y})$. La différence de couverture des estimations de la variance entre les échantillons de petite et de grande taille est prévisible, puisque les estimations de la variance sont asymptotiques et s'améliorent pour les grands échantillons. Pour ce qui est du biais des estimateurs de variance, il apparaît qu'il diminue quand le RSB augmente. Pour un même RSB, tant les variances asymptotiques de forme analytique que les variances bootstrap ont un plus petit biais, étant donné des variables explicatives avec forte corrélation relativement à des variables explicatives avec grande taille de l'effet. Les variances de forme analytique ont tendance à sous-estimer la variance empirique, surtout quand la taille d'échantillon est petite. Dans l'ensemble, la différence est faible entre les deux estimations de variance de forme analytique. La variance bootstrap a tendance à surestimer la variance empirique, mais le biais absolu est généralement plus faible que celui des estimations de variance de forme analytique.

Pour les variables de résultat binaires, tant les estimations asymptotiques de variance de forme analytique que les estimations de variance bootstrap sont sensibles à la taille de l'échantillon, au plan d'échantillonnage et au RSB. Pour la variance bootstrap, la couverture est systématiquement proche de 95 % sous les plans EAS et $\mathrm{POI}(\mathrm{X})$ pour les deux tailles d'échantillon et tous les types de population, mais elle varie de 75 % à 94 % sous le plan $\mathrm{POI}(\mathrm{X+Y})$. Sous ce plan $\mathrm{POI}(\mathrm{X+Y})$, la couverture de la variance bootstrap est meilleure pour la taille d'échantillon de 250 que pour celle de 1 000 quand le biais représente une plus grande part de la REQM, et meilleure pour les populations avec forte corrélation que pour celles avec faible corrélation. Pour ce qui est de la couverture, les variances de forme analytique présentent une tendance similaire aux variances bootstrap sous le plan $\mathrm{POI}(\mathrm{X+Y})$, c'est-à-dire une meilleure couverture pour les petites que pour les grandes tailles d'échantillon, et une meilleure couverture pour les populations avec forte corrélation que pour les populations avec faible corrélation. Sous EAS et $\mathrm{POI}(\mathrm{X})$, la couverture pour les variances de forme analytique s'améliore à mesure qu'augmente la taille de l'échantillon. En ce qui concerne le biais, tant la variance bootstrap que les variances de forme analytique ont un biais plus faible pour les plus grandes tailles d'échantillon. Si la taille d'échantillon est maintenue constante, le biais des estimations de variance de forme analytique augmente quand le RSB augmente. La même tendance ne s'observe pas pour les estimations de variance bootstrap. Comme pour les variables de résultat continues, les estimateurs de variance de forme analytique ont tendance à sous-estimer la variance empirique, surtout quand la taille

d'échantillon est petite. Contrairement aux constatations pour les variables de résultat continues, il s'avère que les estimations de variance de forme analytique pondérées par les poids $g$ ont de meilleures propriétés de biais que les estimations de variance de forme analytique non pondérées. La variance bootstrap a tendance à surestimer la variance empirique. Cependant, les biais sont nettement plus petits que pour les estimations de variance de forme analytique.

**Tableau 4.3**
**Couverture nominale de 95 % et biais en % des estimations de variance pour le LASSO**

| Variable de résultat continue | | | Couverture | | | Biais en % | | |
|---|---|---|---|---|---|---|---|---|
| **Population** | **n** | **Plan** | $v^{\text{LASSO}}$ | $v_g^{\text{LASSO}}$ | $v_{\text{boot}}^{\text{LASSO}}$ | $v^{\text{LASSO}}$ | $v_g^{\text{LASSO}}$ | $v_{\text{boot}}^{\text{LASSO}}$ |
| Faible/petite | 250 | EAS | 91,7 % | 91,8 % | 95,4 % | -22,6 % | -22,3 % | 2,9 % |
| | | POI(X) | 91,2 % | 91,2 % | 96,1 % | -25,1 % | -24,5 % | 5,7 % |
| | | POI(X+Y) | 89,6 % | 89,9 % | 95,4 % | -23,5 % | -22,8 % | 7,9 % |
| | 1 000 | EAS | 93,2 % | 93,2 % | 93,8 % | -7,3 % | -7,2 % | -0,3 % |
| | | POI(X) | 94,0 % | 93,9 % | 95,5 % | -5,7 % | -5,3 % | 6,6 % |
| | | POI(X+Y) | 90,0 % | 90,1 % | 92,1 % | -4,9 % | -4,4 % | 7,9 % |
| Faible/grande | 250 | EAS | 91,5 % | 91,5 % | 95,7 % | -22,6 % | -22,3 % | 6,2 % |
| | | POI(X) | 90,9 % | 91,2 % | 96,4 % | -25,4 % | -24,9 % | 8,8 % |
| | | POI(X+Y) | 90,0 % | 90,2 % | 95,1 % | -24,5 % | -23,7 % | 9,9 % |
| | 1 000 | EAS | 93,4 % | 93,5 % | 94,3 % | -6,6 % | -6,5 % | -0,1 % |
| | | POI(X) | 94,1 % | 94,2 % | 95,9 % | -4,0 % | -3,5 % | 7,6 % |
| | | POI(X+Y) | 90,7 % | 90,7 % | 92,7 % | -2,9 % | -2,3 % | 9,6 % |
| Forte/petite | 250 | EAS | 92,3 % | 92,2 % | 95,4 % | -17,4 % | -17,1 % | 2,0 % |
| | | POI(X) | 92,5 % | 92,6 % | 95,8 % | -17,9 % | -16,1 % | 6,4 % |
| | | POI(X+Y) | 91,2 % | 91,8 % | 96,5 % | -17,4 % | -15,4 % | 7,1 % |
| | 1 000 | EAS | 93,5 % | 93,5 % | 94,4 % | -6,5 % | -6,4 % | -0,9 % |
| | | POI(X) | 94,1 % | 94,0 % | 95,4 % | -5,0 % | -3,1 % | 5,7 % |
| | | POI(X+Y) | 91,9 % | 92,3 % | 93,4 % | -6,0 % | -3,9 % | 5,0 % |
| Forte/grande | 250 | EAS | 92,3 % | 92,3 % | 95,2 % | -19,6 % | -19,3 % | 2,2 % |
| | | POI(X) | 92,0 % | 92,3 % | 96,1 % | -19,6 % | -17,8 % | 7,4 % |
| | | POI(X+Y) | 91,2 % | 91,8 % | 95,6 % | -19,1 % | -16,9 % | 8,3 % |
| | 1 000 | EAS | 93,4 % | 93,4 % | 94,5 % | -6,5 % | -6,4 % | -0,7 % |
| | | POI(X) | 94,0 % | 94,5 % | 95,6 % | -4,7 % | -2,8 % | 6,7 % |
| | | POI(X+Y) | 92,2 % | 92,4 % | 93,4 % | -5,6 % | -3,3 % | 6,1 % |

**Tableau 4.4**
**Couverture nominale de 95 % et biais en % des estimations de variance pour le LASSO**

| Variable de résultat binaire | | | Couverture | | | Biais en % | | |
|---|---|---|---|---|---|---|---|---|
| **Population** | **n** | **Plan** | $v^{\text{LASSO}}$ | $v_g^{\text{LASSO}}$ | $v_{\text{boot}}^{\text{LASSO}}$ | $v^{\text{LASSO}}$ | $v_g^{\text{LASSO}}$ | $v_{\text{boot}}^{\text{LASSO}}$ |
| Faible/petite | 250 | EAS | 89,8 % | 90,0 % | 95,9 % | -28,1 % | -27,8 % | 9,2 % |
| | | POI(X) | 88,1 % | 88,6 % | 96,7 % | -37,3 % | -35,3 % | 9,2 % |
| | | POI(X+Y) | 79,0 % | 79,9 % | 91,2 % | -38,7 % | -35,9 % | 8,0 % |
| | 1 000 | EAS | 92,8 % | 92,8 % | 93,5 % | -11,9 % | -11,8 % | -3,5 % |
| | | POI(X) | 92,0 % | 92,8 % | 95,7 % | -17,9 % | -15,5 % | 1,0 % |
| | | POI(X+Y) | 68,6 % | 69,6 % | 74,6 % | -18,5 % | -14,9 % | 0,5 % |
| Faible/grande | 250 | EAS | 86,8 % | 87,0 % | 94,9 % | -37,7 % | -37,3 % | 11,3 % |
| | | POI(X) | 85,4 % | 86,1 % | 95,5 % | -42,9 % | -41,2 % | 14,4 % |
| | | POI(X+Y) | 78,7 % | 80,1 % | 92,6 % | -44,0 % | -41,3 % | 14,4 % |
| | 1 000 | EAS | 94,4 % | 94,3 % | 95,2 % | -5,5 % | -5,4 % | 5,8 % |
| | | POI(X) | 91,8 % | 92,1 % | 94,9 % | -20,5 % | -18,6 % | -1,8 % |
| | | POI(X+Y) | 76,8 % | 77,8 % | 82,9 % | -20,4 % | -16,9 % | -1,3 % |

**Tableau 4.4 (suite)**
**Couverture nominale de 95 % et biais en % des estimations de variance pour le LASSO**

| Variable de résultat binaire | | | Couverture | | | Biais en % | | |
|---|---|---|---|---|---|---|---|---|
| **Population** | **n** | **Plan** | $v^{\text{LASSO}}$ | $v_g^{\text{LASSO}}$ | $v_{\text{boot}}^{\text{LASSO}}$ | $v^{\text{LASSO}}$ | $v_g^{\text{LASSO}}$ | $v_{\text{boot}}^{\text{LASSO}}$ |
| Forte/petite | 250 | EAS | 89,2 % | 89,1 % | 94,4 % | -28,5 % | -28,1 % | 0,4 % |
| | | POI(X) | 89,0 % | 90,1 % | 95,5 % | -31,9 % | -25,3 % | 12,7 % |
| | | POI(X+Y) | 85,7 % | 88,4 % | 93,8 % | -33,9 % | -25,4 % | 10,9 % |
| | 1 000 | EAS | 93,9 % | 93,9 % | 95,6 % | -6,3 % | -6,2 % | 3,5 % |
| | | POI(X) | 92,6 % | 93,4 % | 94,8 % | -16,5 % | -9,2 % | 1,9 % |
| | | POI(X+Y) | 83,3 % | 85,4 % | 88,1 % | -15,0 % | -5,0 % | 5,2 % |
| Forte/grande | 250 | EAS | 82,8 % | 82,8 % | 93,8 % | -44,6 % | -44,3 % | -6,4 % |
| | | POI(X) | 83,6 % | 85,5 % | 95,1 % | -44,3 % | -39,4 % | 3,8 % |
| | | POI(X+Y) | 82,9 % | 85,1 % | 93,8 % | -45,1 % | -38,4 % | 4,6 % |
| | 1 000 | EAS | 94,3 % | 94,4 % | 96,1 % | -7,8 % | -7,6 % | 6,3 % |
| | | POI(X) | 91,3 % | 92,2 % | 94,0 % | -20,0 % | -13,8 % | 0,2 % |
| | | POI(X+Y) | 86,3 % | 88,6 % | 91,5 % | -18,1 % | -9,2 % | 2,8 % |

# 5 Application à la National Health Interview Survey (NHIS)

## 5.1 Données de la NHIS et l'ACS

Nous appliquons maintenant le calage avec le LASSO à la *National Health Interview Survey* (NHIS) de 2013 pour estimer le nombre total d'adultes (de 18 ans et plus) ayant reçu un diagnostic de cancer dans la population. La NHIS est réalisée auprès d'un échantillon représentatif de la population nationale de ménages de civils ne vivant pas en établissement tiré selon un plan d'échantillonnage aréolaire probabiliste à plusieurs degrés (*Centers for Disease Control and Prevention*, 2005). Chaque mois, des données liées à la santé sont recueillies par interviews sur place auprès d'un échantillon transversal de personnes appartenant aux ménages sélectionnés. Les données fournissent des pseudo unités primaires d'échantillonnage (UPE), des pseudo strates et des poids d'échantillonnage qui permettent de produire des estimations pondérées sous un plan de sondage complexe. En plus de mesures liées à la santé, la NHIS recueille aussi des données démographiques. Notre objectif est d'évaluer notre estimateur LASSO en traitant l'échantillon non pondéré de la NHIS comme représentant un échantillon non probabiliste et d'étudier comment les estimateurs par calage GREG et LASSO se comparent à l'estimateur pondéré selon le plan de sondage.

Pour caler les données de la NHIS sur un ensemble de variables démographiques et liées au revenu, nous utilisons les microdonnées de l'*American Community Survey* (ACS) de 2013 comme données de référence. L'échantillon de l'ACS est constitué de ménages sélectionnés par échantillonnage aréolaire probabiliste à plusieurs degrés dans 3 143 comtés des États-Unis. L'ACS est conçue pour améliorer les estimations sur petits domaines produites entre les échantillons pour le questionnaire détaillé du recensement décennal. Environ trois millions de ménages sont sélectionnés chaque année, et des mesures sont recueillies sur les types de ménages, ainsi que les caractéristiques démographiques des membres des ménages. L'ACS recueille aussi des données sur les logements collectifs, qui sont exclues de la présente analyse. Pour l'ACS de 2013, la taille de l'échantillon d'adultes était de 2 317 301. La taille de l'échantillon de la NHIS de 2013 était de 34 201 après élimination de 242 cas pour lesquels des valeurs manquaient pour les variables

démographiques. Pour les besoins de la présente analyse, nous traitons les estimations pondérées provenant de l'ACS comme les totaux connus de population, une hypothèse raisonnable étant donné les différences de taille d'échantillon.

## 5.2 Estimateurs

La variable de résultat étudiée est celle de savoir si une personne a reçu ou non un diagnostic de cancer. Définissons l'indicateur binaire pour la variable de résultat :

$$y_i = \begin{cases} 1 : \text{ si la personne } i \text{ a reçu un diagnostic de cancer} \\ 0 : \text{ autrement.} \end{cases}$$

Nous commençons par utiliser les poids de sondage de la NHIS de 2013, $\mathbf{w}^{\text{NHIS}}$, et les variables du plan de sondage pour obtenir une estimation sans biais du total de population $T_y = \sum_{i=1}^{N} y_i$. Puis, nous supposons que l'échantillon de la NHIS de 2013 est tiré selon un plan d'échantillonnage aléatoire simple, avec les poids de sondage initiaux $\mathbf{d}^A = N/n$, où $N$ est le total de population obtenu d'après l'ACS, et $n$ est la taille d'échantillon de la NHIS. Nous calons les poids de sondage $\mathbf{d}^A$ sur un ensemble de variables démographiques et de variables de revenu par calage GREG classique et par calage avec le LASSO. Enfin, à titre de compromis entre les méthodes GREG et LASSO, nous considérons le calage assisté par un modèle sur un modèle linéaire pour $y_i$ au lieu du LASSO en utilisant (2.7); notons que, quand $\hat{\mu}_i$ est calculé en utilisant le même modèle linéaire que dans l'estimateur GREG, les estimations ponctuelles du total correspondent, même si les poids de calage diffèrent. Donc, nous produisons sept estimations :

1. $\hat{T}_y^{\text{NHIS}} = \sum_{i \in s_A} w_i^{\text{NHIS}} y_i$ : estimation obtenue avec les poids de la NHIS.
2. $\hat{T}_y^{\text{HTEAS}} = \sum_{i \in s_A} (N/n)\, y_i$ : estimation obtenue avec les poids $\mathbf{d}^A = N/n$.
3. $\hat{T}_y^{\text{GREG1}} = \sum_{i \in s_A} w_i^{\text{GREG1}} y_i$ : estimation obtenue par calage de $\mathbf{d}^A$ par la régression généralisée (GREG) en utilisant toutes les variables de calage.
4. $\hat{t}_y^{\text{GREG1MC}} = \sum_{i \in s_A} w_i^{\text{GREG1MC}} y_i$ : estimation obtenue par calage assisté par un modèle (MC) sur un modèle linéaire utilisant les variables explicatives dans GREG1.
5. $\hat{T}_y^{\text{GREG2}} = \sum_{i \in s_A} w_i^{\text{GREG2}} y_i$ : estimation obtenue par calage de $\mathbf{d}^A$ par la méthode GREG en utilisant uniquement les variables de calage choisies par sélection pas-à-pas descendante.
6. $\hat{t}_y^{\text{GREG2MC}} = \sum_{i \in s_A} w_i^{\text{GREG2MC}} y_i$ : estimation obtenue par calage assisté par un modèle (MC) sur un modèle linéaire utilisant les variables explicatives dans GREG2.
7. $\hat{T}_y^{\text{LASSO}} = \sum_{i \in s_A} w_i^{\text{LASSO}} y_i$ : estimation obtenue par calage assisté par un modèle en utilisant le LASSO.

La variance de $\hat{T}_y^{\text{NHIS}}$ est l'estimation de la variance du total par linéarisation, tenant compte de la strate d'échantillonnage, des unités primaires d'échantillonnage et des poids de sondage dans l'échantillon de la NHIS de 2013. Les variances des estimateurs HTEAS, GREG1 et GREG2 sont les estimations de la variance

par linéarisation avec les poids $\mathbf{d}^A$, $\mathbf{w}^{\text{GREG1}}$ et $\mathbf{w}^{\text{GREG2}}$, respectivement. Nous obtenons la variance de l'estimateur LASSO par le bootstrap naïf.

## 5.3 Modèles de travail

Le tableau 5.1 donne les noms, les étiquettes, les valeurs et les répartitions des variables de calage utilisées dans la présente analyse. La première colonne donne la répartition non pondérée des variables dans l'échantillon de la NHIS. La deuxième colonne contient les répartitions des variables dans l'échantillon de la NHIS, pondérées par les poids au niveau de la personne $\mathbf{w}^{\text{NHIS}}$. La troisième colonne donne la répartition des variables dans la population obtenue à partir des données de référence de l'ACS. La catégorie « Données manquantes » pour le revenu est incluse en tant que catégorie distincte pour saisir la différence entre les profils de valeurs manquantes dans la NHIS et l'ACS. L'ajout d'une catégorie « Données manquantes » nous permet aussi de maintenir la taille de l'échantillon analytique. Comparativement à l'ACS, l'échantillon de la NHIS non pondéré contient de plus fortes proportions de femmes et de personnes veuves/divorcées/séparées, et une plus faible proportion de personnes blanches non hispaniques. Après pondération, les répartitions de la NHIS selon le sexe et la race sont proches des données de référence, et seules les catégories de l'état matrimonial présentent certaines différences.

Nous utilisons un modèle linéaire non pondéré avec sélection pas-à-pas descendante des variables pour déterminer le modèle de travail pour GREG2. Les variables finales incluses dans le modèle pour GREG2 sont l'âge, le niveau d'études, la race, la situation d'emploi (oui/non) et le revenu familial. Pour le calage GREG classique et le calage avec le LASSO, nous utilisons toutes les variables disponibles.

**Tableau 5.1**
**Variables de calage**

| | | | NHIS | ACS |
|---|---|---|---|---|
| | | Pas de poids | Poids au niveau de la personne | Poids au niveau de la personne |
| Région | Nord-est | 16 % | 18 % | 18 % |
| | Midwest | 20 % | 23 % | 21 % |
| | Sud | 37 % | 37 % | 37 % |
| | Ouest | 26 % | 23 % | 23 % |
| Âge | 18 à 29 ans | 19 % | 21 % | 21 % |
| | 30 à 39 ans | 17 % | 17 % | 17 % |
| | 40 à 49 ans | 16 % | 18 % | 18 % |
| | 50 à 59 ans | 17 % | 18 % | 18 % |
| | 60 à 69 ans | 15 % | 14 % | 14 % |
| | 70 à 79 ans | 9 % | 8 % | 8 % |
| | 80 ans et plus | 6 % | 4 % | 5 % |
| Sexe | Hommes | 45 % | 48 % | 48 % |
| | Femmes | 55 % | 52 % | 52 % |
| Niveau d'études | Inférieur au diplôme d'études secondaires | 16 % | 14 % | 13 % |
| | Diplôme d'études secondaires ou moins élevé | 26 % | 26 % | 28 % |
| | Études collégiales partielles | 20 % | 20 % | 23 % |
| | Diplôme d'études collégiales | 29 % | 30 % | 25 % |
| | Études de cycle supérieur | 10 % | 10 % | 10 % |
| Race/ethnicité | Blanche non hispanique | 60 % | 66 % | 66 % |
| | Noire non hispanique | 15 % | 12 % | 12 % |
| | Hispanique | 17 % | 15 % | 15 % |
| | Autre | 8 % | 7 % | 7 % |

**Tableau 5.1 (suite)**
**Variables de calage**

| | | | NHIS | ACS |
|---|---|---|---|---|
| | | Pas de poids | Poids au niveau de la personne | Poids au niveau de la personne |
| État matrimonial | Marié(e)/partenariat | 49 % | 60 % | 52 % |
| | Veuf(ve)/divorcé(e)/séparé(e) | 27 % | 18 % | 20 % |
| | Jamais marié(e) | 24 % | 22 % | 28 % |
| A un emploi | Oui | 35 % | 33 % | 39 % |
| | Non | 65 % | 67 % | 61 % |
| Revenu | 1er quartile | 22 % | 15 % | 19 % |
| | 2e quartile | 20 % | 17 % | 20 % |
| | 3e quartile | 21 % | 22 % | 20 % |
| | 4e quartile | 21 % | 28 % | 19 % |
| | Données manquantes | 17 % | 19 % | 22 % |

## 5.4 Résultats

Le tableau 5.2 donne les estimations, les erreurs-types (e.-t.), la racine carrée de l'erreur quadratique moyenne en traitant la valeur de la NHIS correctement pondérée comme étant la vraie valeur (REQM), l'écart en pourcentage par rapport à l'estimation de la NHIS : écart en % = $100\,(\hat{T} - \hat{T}_y^{\text{NHIS}}) / \hat{T}_y^{\text{NHIS}}$ et l'écart-type, ainsi que le minimum et le maximum, des poids associés à un estimateur donné. Nous traitons l'estimation de la NHIS comme étant l'estimation sans biais parce qu'elle est calculée avec les poids de sondage probabilistes fournis par la NHIS. Sans ajustement de la pondération, l'estimateur HTEAS présente un biais positif de 5,9 %. L'estimateur GREG2 réduit ce biais, le faisant passer de 5,9 % à 2,0 %, l'estimateur GREG1 réduit le biais à 1,8 %, tandis que l'estimateur LASSO réduit le biais à 0,9 %. Par définition, l'emploi de l'estimateur assisté par un modèle utilisant des variables explicatives linéaires donnera le même estimateur que le modèle GREG; cependant, la variabilité est réduite considérablement. Dans la présente analyse, si l'échantillon de la NHIS était non probabiliste, sans ajustement de la pondération, nous aurions surestimé de 1,18 million le nombre d'adultes atteints d'un cancer. Sous le calage classique, l'erreur est réduite à une surestimation de 365 000 (sans sélection des variables) ou de 392 000 (avec sélection des variables). Le calage avec le LASSO réduit encore davantage la surestimation qui se chiffre à 175 000.

**Tableau 5.2**
**Résultats pour l'estimation du nombre total de personnes atteintes d'un cancer. L'écart en % est la différence par rapport à l'estimation de la NHIS divisée par l'estimation de la NHIS**

| Estimateur | $\hat{T}$ | e.-t. | REQM | Écart en % par rapport à la NHIS | É.-T. (min, max) des poids |
|---|---|---|---|---|---|
| NHIS | 19 889 327 | 492 263 | 492 263 | 0,00 % | 5 913 (168; 93 244) |
| HTEAS | 21 070 498 | 362 883 | 1 235 657 | 5,94 % | 0 (6 866; 6 866) |
| GREG1 | 20 254 449 | 375 064 | 523 438 | 1,84 % | 2 474 (-2 409; 16 679) |
| GREG1 MC | 20 254 449 | 349 100 | 505 158 | 1,84 % | 269 (6 181; 7 326) |
| GREG2 | 20 281 603 | 367 900 | 537 802 | 1,97 % | 2 039 (-626; 13 947) |
| GREG2 MC | 20 281 603 | 349 552 | 525 421 | 1,97 % | 260 (6 215; 7 291) |
| LASSO | 20 064 671 | 347 586 | 389 309 | 0,88 % | 323 (5 786; 7 168) |

Comme prévu, l'estimation de la NHIS présente la plus grande erreur-type, car elle intègre convenablement le plan de sondage complexe. Si le modèle de calage de travail saisit correctement la relation entre la variable de résultat et les variables de calage, nous nous attendons à ce que les erreurs-types de l'estimateur par calage soient plus petites que celles de l'estimateur HTEAS. Ce n'est le cas pour aucun des deux estimateurs GREG, pour lesquels l'erreur-type est plus grande que celle de l'estimateur HTEAS, quoique la REQM soit plus petite en raison de la réduction du biais. En outre, l'erreur-type de l'estimateur GREG classique est environ 2,0 % plus grande que celle de l'estimateur GREG avec sélection descendante des variables, caractéristique qui est compensée par la réduction estimée de 6,6 % du biais (bien que cela ne suffise pas à réduire la REQM); l'utilisation de l'estimateur GREG assisté par un modèle (MC) réduit l'erreur-type et la racine carrée de l'erreur quadratique moyenne d'environ 5 à 7 % et 2 à 3 %, respectivement, comparativement aux estimations GREG classiques. Pour le calage avec le LASSO, nous observons une plus petite erreur-type que pour l'estimateur HTEAS, même avec l'estimation bootstrap de la variance qui a tendance à produire une surestimation. Si l'on n'utilise pas les poids de sondage corrects, le calage avec le LASSO produit l'estimation la plus précise d'un total de population, tout en donnant l'erreur-type la plus petite parmi les estimateurs considérés dans la présente application. Et cela, malgré le fait que la variabilité des poids calés avec le LASSO n'était qu'environ le septième de celle des poids GREG, ce que reflète la plus petite erreur-type de l'estimateur proprement dit et la REQM fortement réduite.

# 6 Conclusion

Dans le présent article, nous avons élaboré un estimateur par calage avec le LASSO des totaux de population, $\hat{T}_y^{\text{LASSO}}$, étant donné des données auxiliaires de population. Nous avons également calculé des estimations de variance de forme analytique pour $\hat{T}_y^{\text{LASSO}}$. Les résultats de simulation montrent que les estimations ponctuelles sont approximativement sans biais sous échantillonnage aléatoire simple et sous échantillonnage informatif. Pour la sélection d'échantillons qui sont reliés aux variables d'analyse, le LASSO a permis de réduire de manière significative le biais d'échantillon, même sans utiliser les poids de sondage corrects. Le LASSO a tendance à donner de meilleurs résultats que les modèles de travail obtenus par sélection pas à pas quand les covariables sont fortement collinéaires. Pour l'analyse portant sur de nombreuses variables catégoriques, pour lesquelles il existe des corrélations naturelles entre les catégories, l'estimateur par calage avec le LASSO peut donner de meilleurs résultats que les estimateurs par calage classiques, même si les tailles des effets sont petites. L'amélioration est modeste dans le cas des variables continues, mais considérable quand le résultat d'intérêt est binaire, comme en témoignent les simulations et l'exemple portant sur les données de la NHIS. Nous avons montré théoriquement et au moyen de simulations que le calage avec le LASSO est très prometteur pour ce qui est de faire des inférences sans biais sur les totaux de population à partir d'échantillons non probabilistes. Même si les estimations asymptotiques de variance de forme analytique n'ont pas produit une couverture nominale très précise, le bootstrap naïf est une approche de rechange viable. Dans une application en vue d'estimer le total de population des personnes ayant reçu un diagnostic de cancer, sans utiliser les poids de sondage corrects, l'estimateur par calage avec le LASSO a produit l'estimation qui était la plus proche de l'estimation fondée sur les poids de sondage

corrects. De tous les estimateurs pris en considération, l'estimateur par calage avec le LASSO possédait aussi la plus petite erreur-type, quoique l'estimation bootstrap de la variance qui a été utilisée ne tienne pas compte entièrement de la mise en grappe dans la NHIS, ce qui accroît généralement les erreurs-types. L'application montre que le calage avec le LASSO peut générer une inférence pour la population dans le cas d'une variable de résultat particulière, et que l'inférence est à la fois plus exacte et précise que lorsqu'on utilise les estimateurs par calage classiques.

La question qui se pose est celle de savoir quand le calage assisté par un modèle LASSO doit être utilisé au lieu des méthodes de calage classiques telles que la méthode GREG. Les résultats tant théoriques qu'empiriques décrits dans le présent article donnent à penser qu'il n'y a pas grand-chose à perdre en termes d'efficacité statistique lorsqu'on utilise le calage assisté par un modèle LASSO, mais sa mise en œuvre requiert un effort supplémentaire de la part de l'analyste. Bien que nous ne puissions pas donner des seuils précis, notre analyse laisse entendre que cet effort vaut la peine quand a) il existe un grand nombre de variables de calage possibles, b) un bon nombre de ces variables de calage sont vraisemblablement fortement corrélées, et c) la variable de résultat est binaire plutôt que continue. Nous pensons qu'au moins les conditions a) et b) se présenteront de plus en plus probablement dans des contextes non probabilistes, où les ensembles de données administratives pourraient fournir ces types de variables de calage et que des sous-ensembles de données obtenus par divers moyens contiendront la variable d'intérêt.

Même si la mise en œuvre du LASSO est particulièrement commode et rapide, il existe évidemment d'autres méthodes de régression modernes qui pourraient être considérées en plus du LASSO pour élaborer des modèles de régression pénalisée pour la régression assistée par un modèle de grande dimensionnalité, y compris les approches telles que la régression ridge, l'analyse en composantes principales, ou les arbres de régression additifs bayésiens (Chipman, George et McCulloch, 2010). Ces approches offrent l'occasion de poursuivre la recherche dans ce domaine.

Enfin, nous soulignons que les présents travaux ne représentent qu'une partie d'une littérature plus vaste et qui s'enrichit rapidement sur l'inférence à partir de sondages non probabilistes. En plus des travaux de McConville et coll. (2017), l'approche « Mr. P » (*multi-level regression and poststratification* ou MRP) de Wang et coll. (2015) fait aussi appel à des covariables de grande dimensionnalité pour ajuster les échantillons non probabilistes, en utilisant un modèle hiérarchique plutôt qu'une régression pénalisée. La quasi randomisation (Elliott, 2009; Elliott, Resler, Flannagan et Rupp, 2010; Elliott et Valliant, 2017) et l'appariement d'échantillons (Rivers, 2007; Vavreck et Rivers, 2008) fournissent aussi des solutions de rechange s'appuyant sur des données provenant de quantités de population connues ou d'estimations issues d'un échantillonnage probabiliste pour traiter les problèmes de biais de sélection dans les échantillons non probabilistes. Chacune possède des forces et des faiblesses par rapport aux autres et par rapport à l'assistance par un modèle LASSO. L'approche MRP s'appuie sur des hypothèses de distribution qui pourraient améliorer l'efficacité, mais aussi réduire la robustesse, et elle n'est pas simple à mettre en œuvre dans sa forme entièrement bayésienne. La quasi randomisation perd le lien avec une variable de résultat particulière, ce qui rend les poids qu'elle produit d'usage plus général, mais vraisemblablement moins efficaces, tandis que l'appariement d'échantillons requiert une intervention à l'étape du plan de sondage pour échantillonner

des éléments provenant de la base non probabiliste qui concordent avec les éléments provenant de la population, comme l'échantillonnage par quotas. La décision d'utiliser le calage assisté par un modèle LASSO doit être prise dans le contexte de ces compromis.

# Remerciements

Les auteurs remercient le professeur Fred M. Feinberg et la chercheuse scientifique associée Sunghee Lee de leur révision et suggestions utiles, ainsi que le rédacteur, le rédacteur associé et trois examinateurs de leurs suggestions qui ont beaucoup contribué à l'amélioration du présent article.

# Annexe

## Détermination des estimations pour le LASSO adaptatif

En pratique, nous n'observons pas le taux théorique de croissance de $\lambda_n$, qui optimise les mesures d'adéquation du modèle telles que l'AIC ou le BIC, à moins que nous n'ayons obtenu de nombreux échantillons de la même population ayant diverses tailles. Étant donné un échantillon, les choix de $\lambda_n$ et $\gamma$ dépendent du modélisateur. Dans la mise en œuvre de la fonction *glmnet* en R (Friedman et coll., 2010), une gamme de valeurs de $\lambda_n$ est déterminée selon le schéma suivant :

1. Prendre $\gamma = 0$.
2. Déterminer $\lambda_n^{\max}$ en trouvant la plus petite valeur de $\lambda_n$ qui fixe tous les coefficients à 0.
3. Si la taille d'échantillon $n$ est plus grande que le nombre de paramètres dans le modèle de régression, prendre $\lambda_n^{\min} = 0{,}0001\lambda_n^{\max}$. Si la taille d'échantillon $n$ est plus petite que le nombre de paramètres, prendre $\lambda_n^{\min} = 0{,}01\lambda_n^{\max}$ (pour fixer les paramètres à 0 plus rapidement).
4. Générer une grille de valeurs de $\lambda_n$, habituellement 100 points espacés également entre $\lambda_n^{\min}$ et $\lambda_n^{\max}$.

La gamme initiale de valeurs de $\lambda_n$ est déterminée indépendamment de $\gamma$. Les choix de $\gamma$ sont moins dictés par les données. Certains modélisateurs choisissent l'une des valeurs de $\gamma = 0{,}1; 0{,}5; 1; 2$. Ici, nous déterminons $(\lambda_n, \gamma)$ par validation croisée comme il suit :

1. Obtenir $\alpha_j = 1 / |\hat{\beta}_j^{\text{EMV}}|$.
2. Déterminer 100 valeurs également espacées de $\lambda_n$ basées sur l'exécution de *glmnet* en R.
3. Pour chaque paire $(\lambda_n, \gamma)$, $\lambda_n$ provenant de l'étape 2, et $\gamma = 0{,}1; 0{,}5; 1; 2$, scinder les données en cinq parties (folds). Utiliser quatre parties pour obtenir $\hat{\boldsymbol{\beta}}$.
4. Appliquer $\hat{\boldsymbol{\beta}}$ à la dernière partie non utilisée pour estimer $\hat{\boldsymbol{\beta}}$ et calculer une mesure. Pour une variable $\mathbf{y}$ continue, nous calculons l'erreur absolue moyenne (EAM), $\sum_{i \in s_{A(k)}} |\hat{\mu}_i - y_i|$. Pour une variable $\mathbf{y}$ binaire, nous calculons l'aire sous la courbe (ASC) (au moyen de la fonction *glmnet* :: *auc* en R).

5. Calculer la moyenne des cinq mesures pour chaque paire $(\lambda_n, \gamma)$, et choisir la paire possédant la meilleure mesure moyenne, c'est-à-dire l'EAM minimum pour une variable $\mathbf{y}$ continue, l'ASC maximum pour une variable $\mathbf{y}$ binaire.

Les estimations des coefficients de la régression LASSO adaptatif sont alors obtenues en résolvant les équations (3.1) ou (3.2) de la section 3.1 sachant la paire $(\lambda_n, \gamma)$ sélectionnée. Le code R utilisé pour effectuer la validation croisée est fourni dans le matériel complémentaire en ligne.

## Absence de biais asymptotique et variance de l'estimateur par calage assisté par un modèle LASSO d'un total de population

**Lemme 1** : *Considérons le modèle de superpopulation :*

$$E_\xi(y_k \mid \mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta}), V_\xi(y_k \mid \mathbf{x}_k) = v_k^2\sigma^2.$$

*Soit* $\mathbf{B}$ *l'estimation de la quasi-vraisemblance en population finie de* $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. *Sous les conditions (1) à (5) de la section 3.2, l'estimateur asymptotique assisté par un modèle du total de population est :*

$$\hat{T}_y^{MC} = \sum_{i \in s_A} d_i^A (y_i - \mu_i B^{MC}) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right) \quad \text{(A.1)}$$

*où*

$$\mu_i = \mu(\mathbf{x}_i, \mathbf{B})$$

$$B^{MC} = \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}.$$

*Preuve. La preuve est adaptée et étendue en partant de la preuve du théorème 1 dans Wu et Sitter (2001), avec de légères modifications de la notation pour qu'elle concorde avec celle du présent article. Nous commençons par dériver l'estimateur asymptotique assisté par un modèle pour une moyenne de population,* $\hat{\bar{y}}^{\text{MC}} = N^{-1}\hat{T}_y^{\text{MC}}$ *(voir l'équation (2.7)). Étant donné les conditions (2) et (3), le développement en série de Taylor d'ordre deux de* $\mu(\mathbf{x}_i, \hat{\mathbf{B}})$ *autour de* $\mathbf{B}$ *est :*

$$\mu(\mathbf{x}_i, \hat{\mathbf{B}}) = \mu(\mathbf{x}_i, \mathbf{B}) + \left\{\frac{\partial \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}}\right\}^T (\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \left\{\frac{\partial^2 \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^T}\,|_{\mathbf{t}=\mathbf{B}^*}\right\}(\hat{\mathbf{B}} - \mathbf{B}) \quad \text{(A.2)}$$

*pour* $\mathbf{B}^* \in (\hat{\mathbf{B}}, \mathbf{B})$ *ou* $(\mathbf{B}, \hat{\mathbf{B}})$. *Soit*

$$\mathbf{h}(\mathbf{x}_i, \mathbf{B}) = \frac{\partial \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t}}\,|_{\mathbf{t}=\mathbf{B}}$$

$$\mathbf{k}(\mathbf{x}_i, \mathbf{B}^*) = \frac{\partial^2 \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^T}\bigg|_{\mathbf{t}=\mathbf{B}^*}$$

*Notons que* $\mathbf{h}$ *est un vecteur de longueur* $m$ *et* $\mathbf{k}$ *est une matrice de dimension* $m \times m$, *où* $m$ *est le nombre de paramètres dans* $\boldsymbol{\beta}$. *Étant donné les conditions (2) et (3),*

$$\max_i \left| \mathbf{h}(\mathbf{x}_i, \mathbf{B}) \right| \leq h(\mathbf{x}_i, \mathbf{B}) \tag{A.3}$$

$$\max_{k,\, j} \left| \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*) \right| \leq k(\mathbf{x}_i, \mathbf{B}^*). \tag{A.4}$$

*Les conditions (1) et (3) impliquent que*

$$\mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) = \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(1/\sqrt{n}\right) \tag{A.5}$$

$$\equiv \mu_i + O_p\left(1/\sqrt{n}\right). \tag{A.6}$$

*L'équation (2.2) de la section 2.1 et les conditions d'existence de bornes (boundedness) de (2) et (3) de la section 3.2.2 impliquent que*

$$\begin{aligned} N^{-1}\sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) \\ &\quad + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\right)\left(\hat{\mathbf{B}} - \mathbf{B}\right) \\ &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) + O_p\left(\frac{1}{n}\right). \end{aligned} \tag{A.7}$$

*Étant donné les conditions (1), (4) et l'équation (A.7) :*

$$\begin{aligned} N^{-1}\sum_{k=1}^{N} \mu\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) &- N^{-1}\sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) \\ &= N^{-1}\sum_{k=1}^{N} \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{A.8}$$

*En utilisant les conditions (1) et (3),*

$$\begin{aligned} \bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) \Big/ \sum_{i\in s_A} d_i^A \\ &= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right) + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right) \\ &= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right) + O_p(1/n) \\ &= \bar{\mu} + \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right) + O_p(1/n) \\ &\quad (\textit{selon les conditions } (1) \textit{ et } (18)) \\ &= \bar{\mu} + O_p\left(1/\sqrt{n}\right) + O_p(1/n) \\ &= \bar{\mu} + O_p\left(1/\sqrt{n}\right) \end{aligned} \tag{A.9}$$

*pour* $\bar{\mu} = \sum_{i\in s_A} d_i^A \mu_i \Big/ \sum_{i\in s_A} d_i^A$.

*Alors, partant de (A.2) et (A.9) et en utilisant les conditions (1) à (3), nous avons*

$$\begin{aligned}
N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right) &= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu(\mathbf{x}_i,\mathbf{B})+\mathbf{h}^T(\mathbf{x}_i,\mathbf{B})\left(\hat{\mathbf{B}}-\mathbf{B}\right)+\left(\hat{\mathbf{B}}-\mathbf{B}\right)^T\mathbf{k}\left(\mathbf{x}_i,\mathbf{B}^*\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)-\bar{\mu}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+N^{-1}\sum_{i\in s_A} d_i^A\mathbf{h}^T\left(\mathbf{x}_i,\mathbf{B}\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)\\
&\quad + N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mathbf{B}}-\mathbf{B}\right)^T\mathbf{k}\left(\mathbf{x}_i,\mathbf{B}^*\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)-O_p\left(1/\sqrt{n}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+O_p\left(1/\sqrt{n}\right)+O_p\left(1/n\right)-O_p\left(1/\sqrt{n}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+O_p\left(1/\sqrt{n}\right). \qquad \text{(A.10)}
\end{aligned}$$

*Similairement,*

$$N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2 = N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)^2+O_p\left(1/n\right). \qquad \text{(A.11)}$$

*Partant de (A.10) et (A.11), nous avons :*

$$\begin{aligned}
\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)\left(y_i-\bar{y}\right)}{\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2}=\frac{N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)\left(y_i-\bar{y}\right)}{N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2}\\
&= \frac{\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)\left(y_i-\bar{y}\right)+O_p\left(1/\sqrt{n}\right)}{\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)^2+O_p\left(1/n\right)}\\
&\to B^{MC} \qquad \textit{quand } n\to\infty. \qquad \text{(A.12)}
\end{aligned}$$

*Donc,* $\hat{B}^{MC} = B^{MC}+o_p(1)$, *et nous avons :*

$$\begin{aligned}
\hat{\bar{y}}^{MC} &= N^{-1}\hat{T}_y^{MC}\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)+\sum_{i\in s_A} N^{-1}d_i^A\mu\left(\mathbf{x}_i,\hat{\mathbf{B}}\right)\right)\hat{B}^{MC}\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A} d_i^A\mu\left(\mathbf{x}_i,\mathbf{B}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)\right)\left(B^{MC}+o_p(1)\right)\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A} d_i^A\mu\left(\mathbf{x}_i,\mathbf{B}\right)\right)B^{MC}+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

*Puisque* $N = O_p(N)$, *nous avons* $N\cdot o_P\left(1/\sqrt{n}\right) = O_p(N)\,o_p\left(1/\sqrt{n}\right) = o_p\left(N/\sqrt{n}\right)$. *Donc,*

$$\begin{aligned}\hat{T}_y^{MC} &= N\hat{\bar{y}}^{MC} = N\left(N^{-1}\mathbf{d}^A\mathbf{y} + \left(N^{-1}\sum_{k=1}^{N}\mu(\mathbf{x}_k,\mathbf{B}) - N^{-1}\sum_{i\in s_A}\mu(\mathbf{x}_i,\mathbf{B})\right)B^{MC} + o_p\left(\frac{1}{\sqrt{n}}\right)\right)\\ &= \mathbf{d}^A\mathbf{y} + \left(\sum_{k=1}^{N}\mu(\mathbf{x}_k,\mathbf{B}) - \sum_{i\in s_A}\mu(\mathbf{x}_i,\mathbf{B})\right)B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right)\\ &= \sum_{iins_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N}\mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right). \qquad \text{(A.13)}\end{aligned}$$

**Théorème 2** : *Supposons que les paramètres d'un modèle de régression complet possèdent à la fois des composantes nulles et non nulles, et sans perte de généralité, posons que les $p$ premières sont non nulles et que les $q$ dernières sont nulles :*

$$\boldsymbol{\beta}^F = \begin{pmatrix}\boldsymbol{\beta}^{(1)}_{(p\times 1)}\\ \boldsymbol{\beta}^{(2)}_{(q\times 1)}\end{pmatrix},\ \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} \quad et \quad \boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)}.$$

*Sous les conditions (1) à (5), l'estimateur asymptotique du total par calage avec le LASSO est :*

$$\hat{T}_y^{LASSO} = \sum_{i\in s_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N}\mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right). \qquad \text{(A.14)}$$

*Preuve. Sous la condition (5), la régression LASSO adaptatif satisfait la propriété d'oracle par l'entremise des théorèmes 1 et 4 dans Zou (2006) :*

$$\begin{aligned}Pr(\mathbf{B}^{(2)} = \mathbf{0}) &\to 1\\ \sqrt{n}\left(\hat{\mathbf{B}}^{(1)} - \mathbf{B}\right) &\to N(\mathbf{0},\mathbf{C})\\ \mathbf{B} &\to \boldsymbol{\beta}\end{aligned}$$

*où $\mathbf{C} = \Sigma(\mathbf{B})$ est la matrice de covariance de $\mathbf{B}^{(1)}$ sous le modèle linéaire, et $\mathbf{C} = I^{-1}(\mathbf{B})$ est l'inverse de la matrice d'information de Fisher de $\mathbf{B}^{(1)}$ sous le modèle linéaire généralisé. D'après le théorème de Slutsky, la propriété d'oracle implique que $\hat{\mathbf{B}}^{(1)} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$. Selon la condition (1) et le lemme 1 :*

$$\begin{aligned}\hat{T}_y^{LASSO} &\approx \hat{T}_y^{MC}\\ &= \sum_{iins_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N}\mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right).\end{aligned}$$

**Théorème 3** : *$\hat{T}_y^{LASSO}$ est sans biais sous le modèle.*

*Preuve. Sous l'hypothèse de notre cadre théorique, les paramètres de superpopulation sont un sous-ensemble des paramètres de la régression LASSO complète, et nous pouvons prouver l'absence de biais sous le modèle de $\hat{T}_y^{LASSO}$ en prenant les espérances par rapport au modèle $\xi$. Nous notons d'abord que :*

$$E_\xi[B^{MC}] = E_\xi\left[\frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}\right] = \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2} = 1.$$

*Donc,*

$$E_\xi\left[\hat{T}_y^{LASSO} - T\right] \approx E_\xi\left[\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{MC}\right) + \sum_{i=1}^{N}\mu_i B^{MC} - \sum_{i=1}^{N} y_i\right]$$

$$= \sum_{i\in s_A} d_i^A\left(\mu_i - \mu_i\right) + \sum_{i=1}^{N}\mu_i - \sum_{i=1}^{N}\mu_i \qquad \left(Puisque\ E_\xi\left[B^{MC}\right] = 1\right)$$

$$= 0.$$

Donc, à condition que les paramètres de la régression LASSO comprennent les paramètres de superpopulation, $\hat{T}_y^{\text{LASSO}}$ est sans biais sous le modèle quels que soient les poids de sondage. Cette propriété est essentielle en échantillonnage non probabiliste, où il n'existe pas de poids de sondage initiaux pour garantir l'absence de biais.

# Bibliographie


Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Chipman, H.A., George, E.I. et McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266-298.

Cardot, H., Goga, C. et Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Centers for Disease Control and Prevention (2005). *2004 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description*. National Center for Health Statistics: Hyattsville, Maryland. www.cdc.gov/nchs/data/nhis/srvydesc.pdf.

Czanner, G., Sarma, S.V., Eden, U.T. et Brown, E.N. (2008). A signal-to-noise ratio estimator for generalized linear model systems. *Proceedings of the World Congress on Engineering*, vol. 2.

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

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J. et Mnkemller, T. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecology*, 36, 27-46.

Elliott, M.R. (2009). Combining data from probability and nonprobability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R, Resler, A., Flannagan, C. et Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention*, 42, 530-539.

Elliott, M.R., et Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Fan, J., et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1-22.

Frankel, M.R., et Frankel, L.R. (1987). Fifty years of survey sampling in the United States. *Public Opinion Quarterly*, S127-S138.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Goga, C., Muhammad-Shehzad, A. et Vanheuverzwyn, A. (2011). Principal component regression with survey data: Application on the French media audience. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, 3847-3852.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Kamarianakis, Y., Shen, W. et Wynter, L. (2012). Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. *Applied Stochastic Models in Business and Industry*, 28, 297-315.

Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A., Jack Jr, C.R., Weiner, M.W., de Zubicaray, G.I. et McMahon, K.L. (2012). Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience*, 6, 115.

Kohut, A., Keeter, S., Doherty, C., Dimock, M. et Christian, L. (2012). Assessing the representativeness of public opinion surveys. *Pew Research Center for The People & The Press*. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Mosteller, F. (1949). *The Pre-Election Polls of 1948: The Report to the Committee on Analysis of Pre-Election Polls and Forecasts*, vol. 60, Social Science Research Council.

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

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

Park, M., et Yang, M. (2008). Ridge regression estimation for survey samples. *Communication in Statistics - Theory and Methods*, 37, 532-543.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*, American Statistical Association.

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

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

Skinner, C., et Silva, P. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 76-81.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association,* 43, 12-39.

Terhanian, G., et Bremer, J. (2012). A smarter way to select respondents for surveys? *International Journal of Market Research,* 54, 751-780.

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

Tourangeau, R., Conrad, F.G. et Couper, M.P. (2013). *The Science of Web Surveys.* Oxford University Press, Oxford, Royaume-Uni.

Vavreck, L., et Rivers, D. (2008). The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion, and Parties,* 355-366.

Wang, W., Rothschild, D., Goel, S. et Gelman, A. (2015). Forecasting elections with non-representative Polls. *International Journal of Forecasting*, 31, 980-991.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. et Lange, K. (2009). Genome-wide association analysis by LASSO penalized logistic regression. *Bioinformatics*, 25, 714-721.

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