Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Étalonner les échantillons non probabilistes à l'aide d'échantillons probabilistes utilisant la méthode LASSO

Michael R. Elliott, Jack K-T Chen et Richard L. Valliant[1]

Résumé

En raison de la diminution des taux de réponse aux enquêtes téléphoniques, il est devenu difficile pour les sondeurs électoraux de saisir les intentions de vote en temps opportun. Cela a donné lieu à une utilisation accrue des échantillons provenant d'enquêtes en ligne non probabilistes. Étant donné que les échantillons non probabilistes sont susceptibles de biais d'échantillonnage, nous élaborons une méthode d'étalonnage assistée par modèle à l'aide d'une régression adaptative LASSO – LASSO-estimation-contrôle (ECLASSO). Cette méthode peut produire un estimateur cohérent de totaux de population dans la mesure où un sous-ensemble des vrais prédicteurs est inclus dans le modèle de prévision, ce qui permet d'inclure un grand nombre de covariables possibles sans risque de surajustement. Nous appliquons le modèle ECLASSO pour prédire les résultats du vote de l'élection de mi-mandat aux États-Unis en 2014.

Mots-clés : Enquête probabiliste; pondération de propension; estimateur de régression général; méthode d'étalonnage assistée par modèle; sondage électoral.

1. Introduction

Les échantillons non probabilistes font de plus en plus partie intégrante de la vie de l'analyste d'enquête. Il y a plusieurs raisons à cela. La diminution du nombre de lignes téléphoniques terrestres et l'amélioration des technologies de contrôle téléphonique ont donné lieu à des problèmes majeurs quant à l'utilisation d'enquêtes téléphoniques pour saisir les intentions de vote de manière opportune (Kohut et coll. (2012), Sturgis et coll. (2016)). Des taux croissants de non-réponse (Dutwin et Lavrakas, 2016) et l'augmentation des coûts représentent également des défis. Sur le plan positif, les échantillons non probabilistes peuvent fournir des mesures détaillées de l'intérêt non présent dans les échantillons probabilistes, ainsi que de plus grandes tailles d'échantillons à un coût moindre, surtout dans les petits domaines. Cela donne la possibilité d'améliorer les inférences si les accroissements en précision ne sont pas accablés par un biais d'échantillonnage provenant de l'échantillon non probabiliste.

2. Un cadre pour l'inférence de l'échantillon non probabiliste

La résurgence de l'échantillonnage non probabiliste a amené les chercheurs dans les enquêtes à se pencher sur différentes méthodes d'ajustement pour les échantillons non probabilistes en recourant à des échantillons probabilistes. Elliott et Valliant (2017) examinent des travaux dans ce domaine, et divisent les méthodes entre les approches « quasi-probabilistes » (Schonlau et coll., 2004) et les approches de modélisation de « superpopulation » (Valliant et coll., 2000). Considérons la densité conjointe d'un vecteur de population de la variable d'analyse $\mathbf{Y} = (Y_1, Y_2, ..., Y_N)$ et des variables indicatrices 0-1 pour un échantillon *s*. Dans le contexte le plus général, cette densité peut être définie ainsi :

$$f(\mathbf{Y}, \boldsymbol{\delta}_s | \mathbf{X}; \Theta, \Phi) = f(\mathbf{Y} | \mathbf{X}; \Theta) f(\boldsymbol{\delta}_s | \mathbf{Y}, \mathbf{X}; \Phi)$$

[1]Michael R. Elliott, University of Michigan Institute for Social Research, 426 Thompson St., Ann Arbor, Michigan, USA, 48106 (mrelliot@umich.edu); Jack K-T Chen, Survey Monkey, Palo Alto, California, USA, 94301 (jjkkttcc@gmail.com); Richard L. Valliant, University of Michigan Institute for Social Research, 426 Thompson St., Ann Arbor, Michigan, USA, 48106 (valliant@umich.edu)

où **X** est une matrice $N \times p$ de covariables qui détermine **Y** par l'entremise du paramètre inconnu Θ, et le paramètre inconnu Φ détermine f ($\boldsymbol{\delta}_s$) par l'entremise de **Y** et **X** (Smith 1983; Rubin 1976; Little 1982).

Dans l'échantillonnage probabiliste, l'indicateur d'échantillonnage dépend seulement de **X** bien que les paramètres connus : $f(\boldsymbol{\delta}_s|\mathbf{Y},\mathbf{X};\Phi) = f(\boldsymbol{\delta}_s|\mathbf{X})$; les probabilités connues qui en résultent peuvent être utilisées pour produire des poids d'échantillonnage. Dans le cas de l'échantillonnage non probabiliste, $\boldsymbol{\delta}_s$ peut dépendre de **Y** et/ou de **Φ** en plus de **X**; si l'on pose l'hypothèse que suffisamment de **X** sont disponibles pour que $\boldsymbol{\delta}_s \perp \mathbf{Y} \mid \mathbf{X}$, $f(\boldsymbol{\delta}_s|\mathbf{X};\Phi)$ peut être modélisé et les probabilités estimées qui en résultent peuvent être inversées pour obtenir des « pseudo-poids », que l'on appelle parfois « score de propension pondéré » ou « quasi-randomisation ». Une autre approche, la modélisation de superpopulation, est centrée sur la relation sous-jacente entre **Y** et **X** en modélisant $f(\mathbf{Y}|\mathbf{X};\Theta)$. En répartissant Y en unités échantillonnées et non échantillonnées, $f(\mathbf{Y} \mid \mathbf{X};\boldsymbol{\Theta}) = f(\mathbf{Y}_s \mid \mathbf{Y}_{\bar{s}}, \mathbf{X};\boldsymbol{\Theta}) f(\mathbf{Y}_{\bar{s}} \mid \mathbf{X};\boldsymbol{\Theta})$, nous pouvons utiliser les estimations de modèle à partir de l'échantillon pour prédire les éléments non échantillonnés en posant comme hypothèse que $f(\mathbf{Y}_s \mid \mathbf{Y}_{\bar{s}}, \mathbf{X};\boldsymbol{\Theta}) = f(\mathbf{Y}_s \mid \mathbf{X};\boldsymbol{\Theta})$.

3. Étalonnage

3.1 Estimation par la régression généralisée

En partant des hypothèses qu'un échantillon probabiliste est utilisé et que l'espérance (modélisée) de **Y** est linéaire dans **X**

$$E_M(Y_i) = \mathbf{X}_i^T \boldsymbol{\beta} \tag{1}$$

pour le paramètre inconnu **β** de longueur *p*, nous pouvons trouver la valeur de **β** dans l'estimation d'équation à l'aide de données échantillonnées pour obtenir un estimateur par les moindres carrés

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T \mathbf{D} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{D} \mathbf{Y}_s$$

où $\mathbf{D} = diag(d_i)$ le poids de sondage d_i. Nous pouvons ensuite prédire les éléments non échantillonnés $\hat{\mathbf{Y}}_{\bar{s}}$ à l'aide de $\hat{\mathbf{Y}}_{\bar{s}} = \mathbf{X}_{\bar{s}} \hat{\boldsymbol{\beta}}$ où $\mathbf{X}_{\bar{s}}$ est la matrice $(N-n)\times p$ des variables auxiliaires pour les unités non échantillonnées. Un prédicteur du total de la population est ensuite donné par

$$\hat{T} = \sum_{i \in s} w_i Y_i = \sum_{i \in s} d_i Y_i + \left(\mathbf{T}_{Ux} - \hat{\mathbf{T}}_x\right) \hat{\boldsymbol{\beta}} \tag{2}$$

où $\mathbf{w} = \mathbf{d} + \left(\mathbf{T}_{Ux} - \mathbf{T}_{sx}\right)^T \left(\mathbf{X}_s^T \mathbf{D} \mathbf{X}_s\right)^{-1} \mathbf{X}^T \mathbf{D}$, $\mathbf{T}_{sx} = \sum_{i \in s} d_i \mathbf{x}_i$. Cela correspond à l'*estimateur de régression généralisée* ou (GREG) de Deville et Sarndal (1992). (Si **X** est une variable catégorique de niveau *H*, alors $\hat{T}$ correspond à l'estimateur stratifié : $\hat{T}^{ps} = \sum_{h=1}^{H} N_h \bar{y}_{sh}$, où N_h correspond à la taille de l'échantillon dans la h^{e} strate et $\bar{y}_{sh}$ à la moyenne d'échantillon dans la h^{e} strate. Cet estimateur est sans biais relativement au modèle sous (1) et approximativement sans biais par rapport au plan si un échantillon probabiliste est sélectionné avec les probabilités données par d_i^{-1}.

L'élaboration ci-dessus présume que les poids de sondage sont disponibles comme ils le seraient dans un échantillon probabiliste. Si l'échantillon est non probabiliste, alors **D** est omis, et le total de la population peut être estimé par la l'estimateur de prédiction

$$\hat{T} = \sum_{i \in s} Y_i + \left(\mathbf{T}_{Ux} - \mathbf{T}_{sx}\right)\hat{\boldsymbol{\beta}}$$

avec $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{Y}_s$ et $\mathbf{T}_{xs} = \sum_{i \in s} \mathbf{x}_i$. Cet estimateur est sans bais par rapport au modèle sous (1). Une autre solution consisterait à établir $\mathbf{D} = diag(N/n)$ et utiliser (2), comme cela est effectué ici dans la section 4.

Dans de nombreux cas, l'accessibilité des totaux de contrôle connus pour la population entière $\mathbf{T}_{Ux}$ peut être limitée, ce qui rend l'hypothèse $f(\mathbf{Y}_s \mid \mathbf{Y}_{\bar{s}}, \mathbf{X}; \boldsymbol{\Theta}) = f(\mathbf{Y}_s \mid \mathbf{X}; \boldsymbol{\Theta})$ plutôt forte. Dans ce cas, une enquête probabiliste de référence peut être accessible avec un ensemble plus riche de covariables de $\mathbf{X}$. Nous remplaçons $\mathbf{T}_{Ux}$ dans (2) par $\mathbf{T}_{Bx}$, le total de la population estimé à l'aide de l'enquête probabiliste (Dever et Valliant, 2010); nous faisons référence aux estimateurs qui en résultent en tant qu'estimateurs obtenus en estimant le modèle de régression généralisée de contrôle ou estimateurs ECGREG.

3.2 Méthode d'étalonnage assisté par modèle

Les poids w_i dans GREG peuvent être considérés (Deville et Särndal, 1992) comme les poids qui minimise $\sum_{i \in s}\left(w_i - d_i\right)^2 / d_i$ sous réserve de la contrainte que $\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{T}_{Ux}$ (pour l'étalonnage type pour les totaux de contrôle connus) ou $\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{T}_{Bx}$ (pour l'étalonnage aux totaux de contrôle de référence estimés). La *méthode d'étalonnage assistée par modèle* (Wu et Sitter, 2001) minimise $\sum_{i \in s}\left(w_i - d_i\right)^2 / d_i$ assujetti plutôt aux contraintes que $\sum_{i \in s} w_i = N$ et $\sum_{i \in s} w_i \hat{y}_i = \sum_{i \in U} \hat{y}_i$. Les poids d'étalonnage assistée par modèle sont donnés par $\mathbf{w}^{MC} = \mathbf{d} + \mathbf{DM}(\mathbf{M}^T\mathbf{DM})^{-1}(\mathbf{T} - \mathbf{d}^T\mathbf{M})^T$, où $\mathbf{T} = \left(N, \sum_{i \in U} \hat{Y}_i\right)$ et $\mathbf{M} = \left(\mathbf{1}_{n_s}, \left(\hat{Y}_i\right)_{i \in s}\right)$, et $\mathbf{1}_{n_s}$ est un vecteur de 1s de longueur égale à la taille de l'échantillon. Il peut être montré que l'estimateur assisté par modèle du total donné par la somme pondérée des poids assistés par modèle est donné par

$$\hat{T}^{MC} = \sum_{i \in s} w_i^{MC} Y_i = \sum_{i \in s} d_i Y_i + \left(\sum_{i \in U} \hat{Y}_i - \sum_{i \in s} d_i \hat{Y}_i\right)\hat{B} \tag{3}$$

où le $\hat{B}$ qui satisfait aux contraintes de l'étalonnage est donné par la corrélation pondérée sous le plan entre les valeurs observées et prédites :

$$\hat{B} = \frac{\sum_{i \in s} d_i \left(\hat{Y}_i - \bar{\hat{Y}}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i \in s} d_i \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}$$

où $\bar{\hat{Y}} = \sum_{i \in s} d_i \hat{Y}_i \Big/ \sum_{i \in s} d_i$ et $\bar{Y} = \sum_{i \in s} d_i Y_i \Big/ \sum_{i \in s} d_i$ sont les moyennes pondérées sous le plan des valeurs prédites et observées, respectivement. Pourvu que les poids de sondage d'origine produisent des estimations non biaisées, $\hat{T}^{MC}$ sera approximativement non biaisé sous le plan lorsque la taille de l'échantillon est grande; le « modèle assisté » donné par (1) améliore l'efficacité dans la mesure où il est exact. Tout comme pour les estimateurs ECGREG, nous pouvons remplacer les covariables obtenues à partir de la population par des covariables estimées à partir d'un échantillon de référence s_B (nous référons maintenant à notre échantillon analytique principal par s_A). Cela permet de remplacer les totaux de population $\mathbf{T}$ par $\hat{\mathbf{T}} = \left(\sum_{i \in s_B} d_i^B, \sum_{i \in s_B} d_i^B \hat{Y}_i\right)$, et notre estimateur du total de la population (3) est mis à jour en tant que

$$\hat{T}^{ECMC} = \sum_{i \in s} w_i^{ECMC} Y_i = \sum_{i \in s_A} d_i^A Y_i + \left(\sum_{i \in s_B} d_i^B \hat{Y}_i - \sum_{i \in s_A} d_i^A \hat{Y}_i \right) \hat{B} \quad (4)$$

que nous appelons estimateurs étalonnés à l'aide d'un modèle de contrôle estimé ou estimateurs ECEM.

3.3 ECLASSO

Ici, nous décrivons l'estimé de contrôle d'étalonnage LASSO ou ECLASSO. Dans de nombreux cas, nous pourrions vouloir utiliser un vecteur important des totaux de contrôle possibles, pour le modèle (1) particulièrement si nous les obtenons à partir d'une enquête probabiliste de référence. Or, l'utilisation d'un large nombre de covariables peut entraîner une prédiction instable; ainsi, plutôt que d'obtenir $\hat{\boldsymbol{\beta}}$ à l'aide de la méthode des moindres carrés comme dans ECEM, il faudra utiliser le LASSO adaptatif (Zhou, 2006), une procédure d'estimation plus robuste. McConville et coll. (2017) et Chen et coll. (2018) ont examiné l'utilisation du LASSO adaptatif pour la méthode d'étalonnage assistée par modèle dans la mise en œuvre à l'aide d'un échantillonnage probabiliste ou des valeurs connues de la population; ici nous examinons son utilisation lorsque l'échantillon analytique est un échantillon non probabiliste et que l'échantillon de référence est un échantillon probabiliste.

Les coefficients de régression du LASSO adaptatif sont obtenus en résolvant une équation de régression pénalisée. Pour la régression linéaire du LASSO adaptatif, il s'agit de

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left(\sum_{i \in s_A} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \left| \hat{\beta}_j^{MLE} \right|^{-\gamma} \right)$$

Pour la logistique du LASSO adaptatif, il s'agit de

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left(\sum_{i \in s_A} \left[-y_i \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) + \log\left(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right) \right] + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \left| \hat{\beta}_j^{MLE} \right|^{-\gamma} \right)$$

Le facteur $1 / \left| \hat{\beta}_j^{MLE} \right|^{\gamma}$ équilibre la sélection 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. Dès que les valeurs de λ et γ sont fixées, $\hat{\boldsymbol{\beta}}$ peut être calculé à l'aide de procédures itératives (Friedman et coll. 2010); ces algorithmes sont mis en œuvre dans le paquet R *glmnet*. Des valeurs de λ et γ peuvent être examinées dans une grille de valeurs et sélectionnées à l'aide de la validation croisée. Le LASSO adaptatif comporte une propriété de cohérence au modèle que l'on appelle « propriété d'oracle », qui stipule que, à condition qu'il y ait un modèle de régression dans lequel les paramètres comportent des composantes non nulles $\boldsymbol{\beta}^{[1]}$ et des composantes nulles $\boldsymbol{\beta}^{(2)}$, et que λ s'accroît au moins au taux de $\sqrt{n} / \left(\sqrt{n} \right)^{\gamma}$, mais avec moins de rapidité que $\sqrt{n}$, $P(\hat{\boldsymbol{\beta}}^{(2)} = 0) \to 1$, et $\sqrt{n} \left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)} \right) \to$ N (0,C) où $\mathbf{C} = I^{-1}(\boldsymbol{\beta}^{(1)})$ est l'inverse de la matrice d'information de Fisher de β. En termes moins techniques, sous réserve de conditions de régularités, le LASSO adaptatif convergera vers le vrai modèle pourvu que tous les prédicteurs linéaires réels soient dans le modèle.

Ayant obtenu $\hat{\boldsymbol{\beta}}$ par l'intermédiaire du LASSO adaptatif, nous obtenons $\hat{Y}_i = \mathbf{x}_i^T \boldsymbol{\beta}$ et calculons $\hat{T}^{ECLASSO}$ comme dans (4). Il peut être démontré qu'aussi longtemps que l'échantillon de référence a les bons poids de sondage, $\hat{T}^{ECLASSO}$ sera non biaisé asymptotiquement sous le plan et le modèle, avec une variance de plan asymptotique donnée par

$$v_{\mathrm{A}}(\hat{T}^{ECLASSO}) = \sum_{i\in s_A}\left(\frac{y_i-\hat{Y}_i\hat{B}}{\pi_i^A}\right)^2(1-\pi_i^A)+\sum_{i\in s_A}\sum_{j\neq i}\frac{\pi_{ij}^A-\pi_i^A\pi_j^A}{\pi_{ij}^A}\frac{(Y_i-\hat{Y}_i\hat{B})}{\pi_i^A}\frac{(Y_j-\hat{Y}_j\hat{B})}{\pi_j^A}+$$
$$\sum_{i\in s_B}\left(\frac{\hat{Y}_i\hat{B}}{\pi_i^B}\right)^2(1-\pi_i^B)+\sum_{i\in s_B}\sum_{j\neq i}\frac{\pi_{ij}^B-\pi_i^B\pi_j^B}{\pi_{ij}^B}\frac{\hat{Y}_i\hat{B}}{\pi_i^B}\frac{\hat{Y}_j\hat{B}}{\pi_j^B} \quad \backslash$$

Lorsque l'échantillon analytique A est un échantillon non probabiliste, on pose $\pi_i^A \equiv \frac{n_A}{N}$ et $\pi_{ij}^A = \frac{n_A(n_A-1)}{N(N-1)}$. Une autre approche consiste à utiliser un estimateur bootstrap, lequel peut être obtenu en établissant une population finie bootstrap de l'échantillon de référence et un échantillon aléatoire simple avec remise provenant de l'échantillon analytique, puis en calculant $\hat{T}^{ECLASSO}$ pour chaque échantillon bootstrap. D'après certaines études de simulation (non présentées), trouver la variance analytique tend à sous-estimer la variance réelle, alors qu'un estimateur bootstrap tend à être conservateur.

4. Prédiction des gagnants de 2014 aux postes de sénateurs et de gouverneurs des États-Unis

On a demandé à un échantillon aléatoire de 10 % des répondants ayant répondu à un sondage de SurveyMonkey en octobre 2014 de fournir leurs préférences en matière de vote dans le cadre des courses aux postes de sénateur et de gouverneur; environ 2 % à 3 % l'ont fait. Bien que l'échantillon eût été prélevé au hasard parmi les répondants au sondage, le taux de réponse était faible et, plus important encore, le bassin de répondants ayant répondu à un premier sondage de SurveyMonkey est non probabiliste et peut ne pas être représentative de la population des électeurs. Étant donné qu'imposer des conditions sur les électeurs probables améliore les prédictions électorales (Bolstein, 1991; Gutsche et coll., 2014), nous avons restreint notre analyse à ceux qui ont indiqué : 1) avoir déjà voté, 2) être absolument certains de voter ou 3) être enclins à voter, et qui, par la suite, ont indiqué qu'ils voteraient pour un candidat démocrate ou républicain, les deux principaux partis politiques aux États-Unis. Lorsque l'on a restreint l'enquête à 8 États pour les courses au poste de sénateur (GA, IL, MI, MN, NJ, NC, TX, VA) et à 11 États pour les courses au poste de gouverneur (AZ, CA, FL, GA, IL, MI, NY, OH, PA, TX, WI) dont les échantillons de référence pour l'enquête étaient de taille suffisante, la taille définitive des échantillons analytiques était de 33 199 pour la collecte sur les courses au poste de gouverneur et de 28 686 pour la collecte sur les courses au poste de sénateur.

L'échantillon probabiliste de référence d'électeurs probables a été obtenu par téléphone et par cellulaire pendant les mois de septembre et octobre 2014 par le Pew Research Center (http://www.pewresearch.org [en anglais]). L'échantillon de référence de personnes susceptibles de voter pour un gouverneur s'élevait à 1 094 et à 656 pour les personnes susceptibles de voter pour un sénateur. Les enquêtes comportaient un vaste ensemble de covariables communes : âge, sexe, race, études, religion, pratique religieuse, approbation d'Obama, préférence relative au parti.

La cible de l'inférence est la répartition des votes entre républicains et démocrates dans un État *r*, définie comme suit :

$$\hat{S}_{R(r)-D(r)} = \sum_{i\in s_{A(r)}} w_i y_i \Big/ \sum_{i\in s_{A(r)}} w_i - \sum_{i\in s_{A(r)}} w_i(1-y_i)\Big/ \sum_{i\in s_{A(r)}} w_i = 2\sum_{i\in s_{A(r)}} w_i y_i \Big/ \sum_{i\in s_{A(r)}} w_i - 1$$

où $s_{A(r)}$ est l'ensemble des répondants dans un État *A*, y_i est un indicateur de préférence pour le candidat républicain et w_i dépend de l'estimateur considéré. En particulier, nous considérons quatre estimateurs :

- UNWT : Non ajusté ($w_i = 1$).
- STATEWT : Étalonnage visant les mesures au niveau de l'État à partir d'une enquête probabiliste ($w_i = \hat{p}_i(\mathbf{x}_i)^{-1}$ où $\hat{p}_i(\mathbf{x}_i) = \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\alpha}}\Big/\left(1+\mathbf{x}_i^T\hat{\boldsymbol{\alpha}}\right)\right)$ et $\hat{\boldsymbol{\alpha}}$ est estimé à partir de la régression logistique d'un indicateur qui fait partie des covariables de l'enquête de référence).
- ECGREG : Méthode d'étalonnage assistée par modèle à l'aide de GREG (w_i tel que défini à la section 3.2).

- ECLASSO : Méthode d'étalonnage assistée par modèle à l'aide de LASSO (w_i tel que défini à la section 3.3).

Tous les intervalles de confiance ont été calculés à l'aide de bootstrap. Étant donné que les résultats véritables de l'élection sont connus, nous pouvons calculer les estimations du biais, la racine carrée de l'erreur quadratique moyenne (RCEQM) et la couverture. Compte tenu du nombre limité de répétitions (11 courses au poste de gouverneur et 8 courses au poste de sénateur), nous avons examiné un niveau α de 0,20 plutôt que 0,05 pour la couverture.

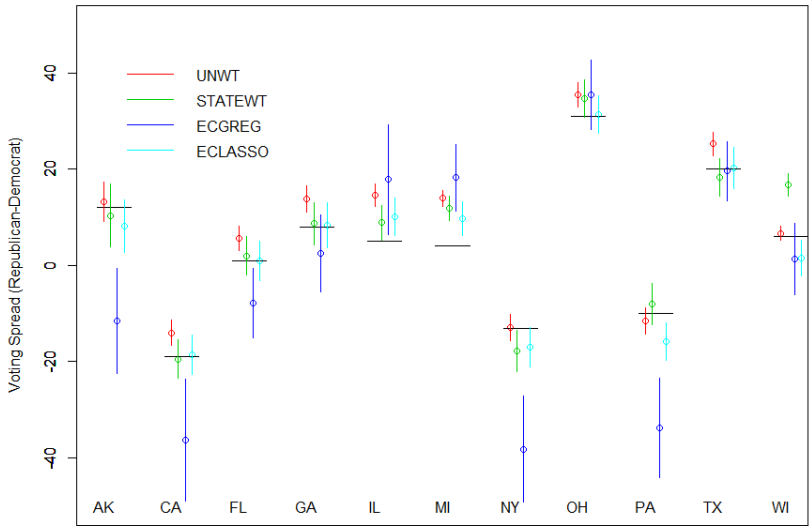
4.1 Résultats : courses au poste de gouverneur

La figure 1 montre les résultats pour les prédictions relatives à l'élection de 11 gouverneurs. Les estimateurs UNWT, STATEWT et ECLASSO ont prédit le parti politique gagnant pour tous les États observés dans l'analyse. L'estimateur ECGREG n'a pas correctement prédit le parti gagnant pour l'Arizona et la Floride. Cependant, sans les ajustements apportés à la pondération, l'échantillon comporte une surreprésentation républicaine, car 10 États sur 11 sont biaisés en faveur des candidats républicains. L'estimateur STATEWT a permis de réduire le biais pour la plupart des États, alors qu'ECGREG semble avoir donné lieu à un surajustement en faveur de l'orientation démocrate. ECLASSO a permis de réduire le biais d'échantillon non ajusté absolu à un maximum de 6 % de valeurs réelles dans les 11 États, par rapport à 10 % à 25 % pour les autres estimateurs. En moyenne, ECLASSO comporte également l'erreur relative la plus petite dans tous les États (0,5 % D contre 1,9 % R à 7,0 % D pour les autres estimateurs), ainsi que la RCEQM la plus petite (4,7 % contre 5,2 % à 15,0 % pour les autres estimateurs). Enfin, STATEWT et ECLASSO ont obtenu la meilleure couverture d'intervalle (7 sur 11 ou 64 % de l'IC de 80 %).

4.2 Résultats : courses au poste de sénateur

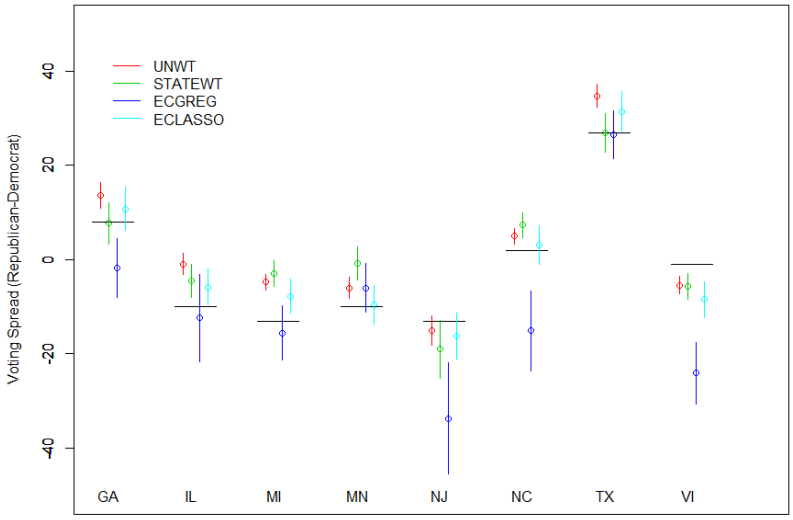
La figure 2 montre les résultats pour les prédictions relatives à l'élection de 8 sénateurs. Les estimateurs UNWT, STATEWT et ECLASSO ont prédit le parti politique gagnant pour tous les États observés dans l'analyse. L'estimateur ECGREG n'a pas correctement prédit le parti gagnant pour la Georgie et la Caroline du Nord. Semblablement à l'échantillon relatif aux courses pour le poste de gouverneur, l'échantillon relatif aux courses pour le poste de sénateur comportait davantage de votes républicains que la répartition réelle des votes avec 6 États sur 8 biaisés en faveur des candidats républicains. L'estimateur STATEWT a permis de réduire le biais pour la majorité des États, alors qu'ECGREG a donné lieu à un surajustement en faveur de l'orientation démocrate. ECLASSO a permis de réduire le biais d'échantillon non ajusté absolu à un maximum de 8 % de valeurs réelles dans les 8 États, par rapport à 9 % à 27 % pour les autres estimateurs. En moyenne, ECLASSO comporte également l'erreur relative la plus petite dans tous les États (1,0 % R contre 2,4 % R par rapport à 9,0 % D pour les autres estimateurs), ainsi que la RCEQM la plus petite (5,1 % contre 6,0 % à 12,2 % pour les autres estimateurs). Dans ce cas, ECLASSO a également obtenu la meilleure couverture d'intervalle (4 sur 8 ou 50 % de l'IC de 80 %), mais cette fois, avec ECGREG (dont les intervalles étaient beaucoup plus larges).

Figure 4.1-1
Résultats pour les courses au poste de gouverneur : estimations ponctuelles et intervalles de confiance de 80 %



Note : Les lignes noires indiquent la répartition réelle des votes.

Figure 4.2-1
Résultats pour les courses au poste de sénateur : estimations ponctuelles et intervalles de confiance de 80%.



Note : Les lignes noires indiquent la répartition réelle des votes.

5. Discussion et prochaines étapes

L'étalonnage est une méthode importante dont il faut tenir compte pour composer avec les biais sélection dans les échantillons non probabilistes. Nous avons élaboré un estimé de contrôle d'étalonnage LASSO ou ECLASSO, qui utilise le LASSO adaptatif pour optimiser un grand nombre de covariables potentielles dans les enquêtes probabilistes de référence pour ajuster les estimations obtenues à partir des enquêtes non probabilistes à l'aide de la méthode d'étalonnage assistée par modèle. Nous montrons qu'un échantillon de référence relativement petit peut être utilisé pour améliorer considérablement l'estimation des courses aux postes de gouverneurs et de sénateurs de 2014 aux États-Unis qui se sont fiés à une enquête non probabiliste d'utilisateurs de SurveyMonkey, par rapport à des estimations non ajustées ainsi que des estimations obtenues à l'aide d'autres méthodes comme l'ajustement du score de propension ou l'estimation de régression généralisée.

Combiner les données des échantillons probabilistes et non probabilistes demeure un domaine qui s'offre à la recherche : l'utilisation du score de propension, de la moyenne et du quantile d'appariement, les effets de mode, l'erreur de mesure, et l'harmonisation et la conformité des données sont tous des sujets importants. Nous espérons que l'application dont nous avons discuté ici donnera lieu à de tels travaux.

Bibliographie

Bolstein, R. (1991), « Predicting the likelihood to vote in pre-election polls », *The Statistician*, 27, p. 277–283.

Chen, J.K.T., R. L. Valliant, et M. R. Elliott (2018), « Model-assisted calibration of non-probability sample survey data using adaptive LASSO », *Survey Methodology*, 44, p. 117-144.

Dever, J., et R. L. Valliant (2010), « A comparison of variance estimators for poststratification to estimated control totals », *Survey Methodology*, 36, p. 45–56.

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

Dutwin, D., et P. Lavrakas (2016), « Trends in telephone outcomes », *Survey Practice*, 9(2), p. 1–9.

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

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

Gutsche, T., A. Kapteyn, E. Meijer, et B. Weerman (2014), « The RAND continuous 2012 presidential election poll », *Public Opinion Quarterly*, 78, p. 233–254.

Kohut, A., S. Keeter, C. Doherty, M. Dimock, et L. Christian (2012), *Assessing the representativeness of public opinion surveys.*

Little, R. J. A. (1982), « Models for nonresponse in sample surveys », *Journal of the American Statistical Association*, 77, p. 237–250.

McConville, K., F. Bredit, T. Lee, et G. Moisen (2017), « Model-assisted survey regression estimation with the lasso », *Journal of Survey Statistics and Methodology*, 5, p. 131–158.

Rubin, D. (1976), « Inference and missing data », *Biometrika*, 63, p. 581–592.

Schonlau, M., K. Zapert, L. Simon, K. Sanstad, S. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, et S. Berry (2004), « A comparison between responses from a propensity weighted web survey and an identical RDD survey », *Social Science Computer Review*, 22, p. 128–138.

Smith, T. M. F. (1983), « On the validity of inferences from non-random samples », *Journal of the Royal Statistical Society*, A146, p. 394–403.

Sturgis, P., N. Baker, M. Callegaro, S. Fisher, J. Green, W. Jennings, J. Kuha, B. Lauderdale, et P. Smith (2016), *Report of the Inquiry into the 2015 British general election opinion polls*.

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

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

Zou, H. (2006), « The adaptive lasso and its oracle properties », *Journal of the American Statistical Association*, 101, p. 1418–1429.