ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 3, *pp.* 657–681

Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling

Jack Kuang Tsung Chen

SurveyMonkey, Palo Alto, USA

and Richard L. Valliant and Michael R. Elliott

University of Michigan, Ann Arbor, USA

[Received December 2017. Final revision October 2018]

Summary. Declining response rates and increasing costs have led to greater use of non-probability samples in election polling. But non-probability samples may suffer from selection bias due to differential access, degrees of interest and other factors. Here we estimate voting preference for 19 elections in the US 2014 midterm elections by using large non-probability surveys obtained from SurveyMonkey users, calibrated to estimated control totals using model-assisted calibration combined with adaptive LASSO regression, or the estimated controlled LASSO, ECLASSO. Comparing the bias and root-mean-square error of ECLASSO with traditional calibration methods shows that ECLASSO can be a powerful method for adjusting non-probability surveys even when only a small sample is available from a probability survey. The methodology proposed has potentially broad application across social science and health research, as response rates for probability samples decline and access to non-probability samples increases.

Keywords: Election polls; General regression estimator; Model-assisted calibration; Probability survey; Propensity weighting

1. Introduction

One of the most prominent applications of survey research is election polling. The timeframe to collect critical voting intention is short, typically spanning just the last few weeks before the election day. Due to declining land-line phone coverage and improved phone screening technology, it has become a significant challenge for election pollsters to capture voting intentions in a timely way (Kohut *et al.*, 2012; Sturgis *et al.*, 2016). This became very clear in the recent US presidential election, where election polls underestimated Donald Trump's support *versus* Hillary Clinton because of non-response bias, measurement error ('shy' Trump voters) and failure to predict likely voters, among other reasons (Mercer *et al.*, 2016). Further, declines in response rates (Dutwin and Lavrakas, 2016) and increasing costs for probability surveys have impacted the collection of data for scientific research throughout the social science and health fields as well. Hence there is an increasing push to use data from administrative sources, social media and other non-probability-based sources to substitute for probability samples across the spectrum of survey research.

Address for correspondence: Michael R. Elliott, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA.
E-mail: mrelliot@umich.edu

 0035–9254/19/68657

Recent research has shown the potential use of non-probability samples to predict election outcomes. Wang *et al*. (2015) performed multilevel regression and post-stratification on Xbox users to predict the US 2012 presidential election results accurately. Tumasjan *et al*. (2010) found success in analysing the frequency of candidates appearing in Twitter texts to estimate the support for political candidates in the 2009 German federal election. However, because non-probability samples lack a well-defined sampling frame, they can have extremely imbalanced sample composition relative to the general voting population. Wang *et al*. (2015) found, for example, that the Xbox sample was over 90% male, with 75% aged 18–44 years, compared with less than 50% male and 50% age 18–44 years in the 2008 presidential election exit polls. Yet by making post-survey adjustments to match Xbox sample characteristics to 2008 exit poll characteristics, they could correctly forecast the outcome of the 2012 presidential election. In addition to basic voter demographics, the 2008 exit poll contained political ideology, party identification and information on the support for presidential candidate Obama, making the exit poll a powerful source of benchmark data for the 2012 presidential election where President Obama ran for re-election. For most elections, however, no such large-scale benchmark exists before the election. Post-survey adjustments are limited to basic demographics such as age, gender, race and education from large-scale government surveys. As voter intentions are often associated with other factors such as religious beliefs, attitudes towards current political agenda and political party support (Krosnick, 1988; Abramowitz, 2008), post-survey adjustments only to basic demographics are unlikely to remove all bias in imbalanced non-probability samples. Hence there is need to rely on adjustment to factors that might only be available in small, high quality benchmark samples such as the Pew Research Center (`http://www.pewresearch.org`) probability sample polls.

The resurgence of non-probability sampling has prompted survey researchers to explore different adjustment methods for non-probability samples by using probability samples. Elliott and Valliant (2017) review work in this area, dividing methods into 'quasi-likelihood' approaches (Schonlau *et al*., 2004) *versus* 'superpopulation' modelling approaches (Valliant *et al*., 2000). The quasi-likelihood approaches include propensity score weighting, which combines probability and non-probability samples to generate pseudo-selection-weights for non-probability sample respondents. Superpopulation modelling includes calibration adjustments, which adjust the non-probability sample so that the weighted sample totals of the calibration variables equal their benchmark totals. Here we undertake an approach that combines both quasi-likelihood and modelling approaches by utilizing a probability-based benchmark sample that is similar to the probability-based reference sample that is used for propensity score weighting. We then use an *assisting model* to predict an outcome of interest, given a set of calibration variables that exists in both probability and non-probability samples. The outcome variable in the non-probability sample is then calibrated to the predicted outcome total in the probability sample, given the probability sampling weights in the benchmark data. In addition, although a general theme in the literature is to include all variables that can be used for calibration, in practice this can lead to instability and overfitting, especially if, as is often the case, the probability sample is much smaller than the non-probability sample. Thus we employ the least angle shrinkage and selection operator LASSO (Tibshirani, 1996) to assist in the construction of weights for a specific outcome variable. LASSO performs both variable selection and parameter estimation, which can serve as a powerful assisting model by determining the most accurate and parsimonious model. We choose one variant of LASSO, the adaptive LASSO (Zou, 2006), as the assisting model, because the adaptive LASSO has shown to have model consistency properties under mild conditions (i.e. it can select the correct model, and provide asymptotically unbiased parameter estimates). We extend LASSO calibration to estimated control LASSO calibration, ECLASSO,

for incorporating sampling uncertainties of the benchmark data into the variance component of model-assisted calibration estimators. Although our focus is on adjusting non-probability samples by using benchmark probability sample surveys, we develop our framework in a setting that allows for adjustment of probability samples to benchmark data as well.

The organization of this paper is as follows. Section 2 provides background and notation for traditional post-survey weighting schemes that are used for non-probability samples. Section 3 provides background and notation for model-assisted calibration and formulates the ECLASSO estimator for a population total of continuous and binary outcome variables, $\hat{T}_y^{\text{ECLASSO}}$. Section 4 applies ECLASSO to predict the voting spread (the proportion of Democratic votes minus the proportion of Republican votes, $D - R$) for 11 gubernatorial elections and eight Senate elections in the US 2014 midterm election. Section 5 describes the simulation that was used to evaluate $\hat{T}_y^{\text{ECLASSO}}$ and the asymptotic linearized variance estimates. We summarize our findings in Section 6.

The data that are used for the simulation study and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`

2. Post-survey weighting schemes for non-probability samples

2.1. Propensity score weighting

Suppose that a non-probability sample and a probability-based reference sample are available, with a common set of measures, $\mathbf{X}$. Pooling the data from these studies, let $Z_i = 1$ if respondent i is a non-probability sample respondent and $Z_i = 0$ otherwise, with the propensity to be in the non-probability sample given by $p_i = \Pr(Z_i = 1|\mathbf{X})$. The propensity score weights are simply the inverse of propensity scores, $w_i^{\text{PSCORE}} = 1/p_i$. For an outcome of interest Y, the weighted estimate of Y based on w_i^{PSCORE} is unbiased only when we have conditional independence between Y and Z given $\mathbf{X}$: $P(Z = 1|\mathbf{X}, Y) = P(Z = 1|\mathbf{X})$. (This can be tested by considering the distribution of Y given Z conditional on p, either by comparing the distribution of Y given Z within categories of p, or by regressing Y on Z and comparing it with the regression of Y on Z and p simultaneously.) In practice, p_i must be estimated, typically via logistic regression. Estimators of totals based on propensity score weights are given by

$$\hat{T}_y^{\text{PSCORE}} = \sum_{i \in s_{\text{A}}} w_i^{\text{PSCORE}} y_i \tag{1}$$

where s_{A} is the non-probability sample, and y_i is a variable measured on unit i.

2.2. Traditional calibration

Define the analytic sample s_{A} of size n_{A} to be the data set containing the targeted data for analysis, Y. We consider the general setting where this could be either a probability sample with known design weights $\mathbf{d}_{n_{\text{A}} \times 1}^{\text{A}}$, or a non-probability sample, where, in the absence of true design information, d_i^{A} is typically set to N/n for all i, which is equivalent to assuming a simple random-sample design. Defining the diagonal matrix of design or pseudodesign weights as $\mathbf{D}^{\text{A}}$, the calibrated weights $\mathbf{w}_{n_{\text{A}} \times 1}$ minimize an expected distance measure with respect to the design of A, $\mathscr{A}$ (Deville and Sarndal, 1992):

$$E_{\mathscr{A}}\left[\sum_{i \in s_{\text{A}}} g(w_i, d_i^{\text{A}})/q_i\right] \tag{2}$$

under the constraint $\Sigma_{i\in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}}$ where $\mathbf{T}^{\mathbf{X}}$ is a row vector of known population totals of $\mathbf{X}$ from a population of size N and $g(w_i, d_i^A)$ is a differentiable function with respect to w_i, strictly convex on an interval containing d_i^A, and $g(d_i^A, d_i^A) = 0$. The χ^2-distance measure $g(w_i, d_i^A) = (w_i - d_i^A)^2/d_i^A$ with $q_i = 1$ yields the generalized regression estimator, GREG:

$$\mathbf{w}^{\mathrm{GREG}} = \mathbf{d}^A + \mathbf{D}^A\mathbf{X}(\mathbf{X}^T\mathbf{D}^A\mathbf{X})^{-1}(\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X})^T. \tag{3}$$

The estimate of population total of outcome $\mathbf{y}$ based on GREG-calibrated weights is

$$\begin{aligned}\hat{T}_y^{\mathrm{GREG}} &= (\mathbf{w}^{\mathrm{GREG}})^T\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + (\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X})\hat{\boldsymbol{\beta}}\end{aligned} \tag{4}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{D}^A\mathbf{X})^{-1}\mathbf{X}\mathbf{D}^A\mathbf{y}$ is the weighted least square estimate of the linear regression $\mathbf{y}$ on $\mathbf{X}$, given weights $\mathbf{D}^A$. (Again, in the non-probability setting, $\mathbf{d}^A = (N/n)\mathbf{1}$ and $\mathbf{D}^A = (N/n)\mathbf{I}$.) The calibrated weights that are defined in equation (3) do not rely on any outcome variable. Thus the same set of weights can be applied to all variables in the survey.

To incorporate uncertainties from benchmark totals, Dever and Valliant (2010) introduced estimated control calibration. The framework replaces known population totals $\mathbf{T}^{\mathbf{X}}$ in equation (3) by estimated totals from the benchmark $\hat{\mathbf{T}}^{\mathbf{X}}$:

$$\mathbf{w}^{\mathrm{ECGREG}} = \mathbf{d}^A + \mathbf{D}^A\mathbf{X}(\mathbf{X}^T\mathbf{D}^A\mathbf{X})^{-1}(\hat{\mathbf{T}}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X})^T. \tag{5}$$

The resulting estimator of population total is

$$\begin{aligned}\hat{T}_y^{\mathrm{ECGREG}} &= (\mathbf{w}^{\mathrm{ECGREG}})^T\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + (\hat{\mathbf{T}}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X})\hat{\boldsymbol{\beta}}.\end{aligned} \tag{6}$$

The estimate control calibration estimator ECGREG has the same general form as GREG; thus we use the notation $\mathbf{w}^{\mathrm{ECGREG}}$ and $\hat{T}_y^{\mathrm{ECGREG}}$ to denote weights and estimator based on the estimated control calibration.

2.3. Model-assisted calibration

Model-assisted calibration assumes a model between an outcome $\mathbf{y}$ and $\mathbf{X}$ through the first two moments:

$$\begin{aligned}E_\xi[y_k|\mathbf{x}_k] &= \mu(\mathbf{x}_k, \boldsymbol{\beta}), \\ V_\xi(y_k|\mathbf{x}_k) &= \nu_k^2\sigma^2\end{aligned} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and σ are unknown superpopulation parameters, $\mu(\mathbf{x}_k, \boldsymbol{\beta})$ is a known function of $\mathbf{x}_k$ and $\boldsymbol{\beta}$, and ν_k is a known function of $\mathbf{x}_k$ or $\mu(\mathbf{x}_k, \boldsymbol{\beta})$. E_ξ and V_ξ are the expectation and variance with respect to the model ξ. Let $\mathbf{B}$ be the finite population parameter of $\boldsymbol{\beta}$ that solves the population score equation $\Sigma_i^N\{y_i - \mu(\mathbf{x}_i, \mathbf{B})\} = 0$, and $\hat{\mathbf{B}}$ be the quasilikelihood estimator of $\mathbf{B}$ given by $\Sigma_{i\in s_A} d_i\{y_i - \mu(\mathbf{x}_i, \hat{\mathbf{B}})\} = 0$. The model-assisted calibrated weights $\mathbf{w}$ minimize a distance measure $E_{\mathscr{D}}[\Sigma_{i\in s_A} g(w_i, d_i^A)/q_i]$ under the constraints $\Sigma_{i\in s_A} w_i = N$ and $\Sigma_{i\in s_A} w_i\hat{\mu}_i = \Sigma_i^N\hat{\mu}_i$, where $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$. Under χ^2-distance measure with $q_i = 1$, the model-assisted calibrated weights are

$$\mathbf{w}^{\mathrm{MC}} = \mathbf{d}^A + \mathbf{D}^A\mathbf{M}(\mathbf{M}^T\mathbf{D}^A\mathbf{M})^{-1}(\mathbf{T}^{\mathrm{M}} - (\mathbf{d}^A)^T\mathbf{M})^T \tag{8}$$

where $\mathbf{D}^A = \mathrm{diag}(\mathbf{d}^A)$, $\mathbf{T}^{\mathrm{M}} = (N, \Sigma_i^N\hat{\mu}_i)$ and $\mathbf{M} = (\mathbf{1}^A, (\hat{\mu}_i)_{i\in s_A})$. The estimate for the population total based on model-assisted calibrated weights is given by

$$\begin{aligned}\hat{T}_y^{\mathrm{MC}} &= (\mathbf{w}^{\mathrm{MC}})^{\mathrm{T}}\mathbf{y} \\ &= (\mathbf{d}^{\mathrm{A}})^{\mathrm{T}}\mathbf{y} + (\mathbf{T}^{\mathbf{M}} - (\mathbf{d}^{\mathrm{A}})^{\mathrm{T}}\mathbf{M})(\mathbf{M}^{\mathrm{T}}\mathbf{D}^{\mathrm{A}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{D}^{\mathrm{A}}\mathbf{y} \\ &= (\mathbf{d}^{\mathrm{A}})^{\mathrm{T}}\mathbf{y} + \left(\sum_{i}^{N}\hat{\mu}_k - \sum_{i\in s_{\mathrm{A}}} d_i^{\mathrm{A}}\hat{\mu}_i\right)\hat{B}^{\mathrm{MC}} \end{aligned} \tag{9}$$

where $\hat{B}^{\mathrm{MC}}$ is the calibration slope that satisfies the calibration constraints:

$$\hat{B}^{\mathrm{MC}} = \frac{\sum_{i\in s_{\mathrm{A}}} d_i^{\mathrm{A}}(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{\sum_{i\in s_{\mathrm{A}}} d_i^{\mathrm{A}}(\hat{\mu}_i - \hat{\bar{\mu}})^2} \tag{10}$$

where $\hat{\bar{\mu}}$ and $\bar{y}$ are the design-weighted means of the predicted values $\hat{\mu}_i$ and the observed data y_i. (Note that $\hat{B}^{\mathrm{MC}}$ is different from the model parameter estimates $\hat{\mathbf{B}}$.) Wu and Sitter (2001) have shown that $\hat{T}_y^{\mathrm{MC}}$ is asymptotically design unbiased, even when the model is misspecified. As long as the original design weights produce unbiased estimates, $\hat{T}_y^{\mathrm{MC}}$ is approximately unbiased when the sample size is large. Similarly to ECGREG, to account for uncertainties in the benchmark sample, we replace $\mathbf{T}^{\mathrm{M}} = (N, \Sigma_{i\in U}\hat{\mu}_i)$ with estimates from a benchmark sample: $\hat{\mathbf{T}}^{\mathrm{M}} = (\Sigma_{i\in s_{\mathrm{B}}} d_i^{\mathrm{B}}, \Sigma_{i\in s_{\mathrm{B}}} d_i^{\mathrm{B}}\hat{\mu}_i)$, where s_{B} denotes the benchmark sample and d_i^{B} is the probability-based design weights of the benchmark sample:

$$\hat{T}_y^{\mathrm{ECMC}} = (\mathbf{d}^{\mathrm{A}})^{\mathrm{T}}\mathbf{y} + \left(\sum_{i\in s_{\mathrm{B}}} d_i^{\mathrm{B}}\hat{\mu}_i - \sum_{i\in s_{\mathrm{A}}} d_i^{\mathrm{A}}\hat{\mu}_i\right)\hat{B}^{\mathrm{MC}}. \tag{11}$$

3. Estimated control LASSO calibration

Because we are relying so heavily in non-probability samples on models that can approximate the expected value of y_i to compensate for the lack of design weights, a large number of covariates and, consequently, control totals may be required to obtain accurate models. This can greatly increase the probability that the information to estimate totals in the available data may be sparse, resulting in unstable calibrated weights. The problem is made worse in ECGREG, where the benchmark sample is small. Thus we consider the use of the adaptive LASSO for the development of the calibration models, which will allow the inclusion of large numbers of potential predictors while simultaneously penalizing any potential overfitting.

3.1. Assisting model—adaptive LASSO

The adaptive LASSO regression coefficients are obtained by solving a penalized regression equation (Zou, 2006). For linear adaptive LASSO regression,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i\in s_{\mathrm{A}}}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda_n\sum_{j=1}^{p}\alpha_j^{\gamma}|\beta_j|\right\}. \tag{12}$$

Similarly, for the logistic adaptive LASSO,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left(\sum_{i\in s_{\mathrm{A}}}[-y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \log\{1 + \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}] + \lambda_n\sum_{j=1}^{p}\alpha_j^{\gamma}|\beta_j|\right). \tag{13}$$

The role of the weight parameter α_j is to prevent LASSO from selecting covariates with large effect sizes in favour of lowering prediction error when the sample size is small. Thus the weights are inversely proportional to effect sizes of regression parameters: $\alpha_j \propto 1/|\beta_j|$. A

common choice of α_j is $1/|\hat{\beta}_j^{\mathrm{MLE}}|$, where $\hat{\beta}_j^{\mathrm{MLE}}$ is the maximum likelihood estimate of β_j. The power of the weight parameter γ is a constant greater than 0 that interacts with α_j to control LASSO from selecting or excluding parameters. It is important to consider a reasonable range of γ and λ_n during the model selection process through regularization. Given that α_j is inversely proportional to β_j, small values of γ will favour covariates with large effect sizes (which is useful when there are known dominant predictors), whereas a large value of γ allows regularization to treat all covariates equally (which is useful when there is no prior knowledge of predictors). As there is a threshold value of λ_n that sets all regression coefficients to 0, there is no practical value to fitting LASSO with λ_n greater than the threshold. Thus only a range of positive values less than the λ_n-threshold need to be explored. We recommend a cross-validation approach to select λ_n and γ, given a sensible range of values; see the on-line appendix for details. Once λ_n and γ have been selected, we can calculate $\hat{\boldsymbol{\beta}}$ through iterative procedures; see Friedman *et al.* (2010) for details. These algorithms are implemented in `glmnet`. If design weights are available in the analytic data set, weighted versions of equations (12) and (13) can be fitted (McConville *et al.*, 2017); for this application we focus on the setting where the analytic data set is a non-probability sample, and the weights d_i^{A} are constant and can be ignored.

The adaptive LASSO has a model consistency property known as the oracle property, which states that, under the condition that λ_n grows at least at the rate of $\sqrt{n}/(\sqrt{n})^{\gamma}$ but not faster than $\sqrt{n}$, the true model will be discovered, i.e., for a regression model in which the parameters have both non-zero $\boldsymbol{\beta}^{(1)}$ and zero components $\boldsymbol{\beta}^{(2)}$, $\Pr(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}) \to 1$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}) \to N(\mathbf{0}, \mathbf{C})$ where $\mathbf{C} = I^{-1}(\boldsymbol{\beta}^{(1)})$ is the inverse of the Fisher information matrix of $\boldsymbol{\beta}$.

3.2. Estimated control LASSO calibration

The asymptotic properties of $\hat{T}_y^{\mathrm{ECMC}}$, and in particular its development using estimated control totals under LASSO, have not been established in the literature. This section develops the asymptotic expectation and the asymptotic linearized variance estimate of the ECLASSO estimator of a population total. We make the following assumptions.

Assumption 1. The analytical samples, s_{A} with size n_{A}, are drawn from a single-stage, unequal probability of selection sampling design $\mathscr{A}$, with selection probability for unit i denoted by π_i^{A}, and the joint selection probability of units i and j denoted by π_{ij}^{A}. We denote the design weight for unit i by $d_i^{\mathrm{A}} = 1/\pi_i^{\mathrm{A}}$, the vector of design weights by $\mathbf{d}^{\mathrm{A}}$ and the diagonal matrix of design weights by $\mathbf{D}^{\mathrm{A}}$. A set of calibration variables is denoted by $\mathbf{X}^{\mathrm{A}}$. For non-probability samples, $\pi_i^{\mathrm{A}} = n_{\mathrm{A}}/N$ and $\pi_{ij}^{\mathrm{A}} = n_{\mathrm{A}}(n_{\mathrm{A}} - 1)/\{N(N-1)\}$.

Assumption 2. The benchmark samples, s_{B} with size n_{B}, are drawn from a single-stage sampling design $\mathscr{B}$, allowing for unequal probabilities of selection. The selection probability for unit i is denoted by π_i^{B}, and the joint selection probability of units i and j is denoted by π_{ij}^{B}. We denote the design weight for unit i by $d_i^{\mathrm{B}} = 1/\pi_i^{\mathrm{B}}$, the vector of design weights by $\mathbf{d}^{\mathrm{B}}$ and the diagonal matrix of design weights by $\mathbf{D}^{\mathrm{B}}$. A set of calibration variables is denoted by $\mathbf{X}^{\mathrm{B}}$.

Assumption 3. A superpopulation model is assumed, as is described in Section 3.1:

$$E_\xi[y_k|\mathbf{x}_k] = \mu(\mathbf{x}_k, \boldsymbol{\beta}),$$
$$V_\xi(y_k|\mathbf{x}_k) = \nu_k^2 \sigma^2.$$

Assumption 4. The true superpopulation parameters $\boldsymbol{\beta}_v$ are a subset of the full regression model for LASSO:

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

where, without loss of generality, $\boldsymbol{\beta} \equiv \boldsymbol{\beta}^{(1)}$ consists of the p non-zero components of the full model and $\boldsymbol{\beta}^{(2)} \equiv \mathbf{0}_{q\times 1}$.

Assumption 5. The full range of $\mathbf{X}$ in the population has non-zero probability of being observed in both analytical and benchmark samples. (This is needed because predictions are implicitly made for the non-sample part of the population. This assumption would hold trivially if both the analytic and the benchmark samples were probability samples from the desired population. However, when the analytic sample is non-probability, undercoverage is a real danger that should be guarded against by using allocation methods like quota sampling that control the spread of the sample over covariate values.)

The ECLASSO calibration estimate of the total can be obtained from the following steps.

Step 1. Obtain LASSO regression coefficients $\hat{\boldsymbol{\beta}}$ as described in Section 3.1. We use the R package `glmnet` (Friedman *et al.*, 2010) to obtain the LASSO coefficients $\hat{\boldsymbol{\beta}}$, given a pair of (λ_n, γ) selected by cross-validation.
Step 2. Use $\hat{\boldsymbol{\beta}}$ to calculate $\hat{\mu}_i = \mu(\mathbf{x}_i^{\mathrm{A}}, \hat{\boldsymbol{\beta}})$ in the analytic sample, and $\hat{\mu}_i = \mu(\mathbf{x}_i^{\mathrm{B}}, \hat{\boldsymbol{\beta}})$ in the benchmark sample.
Step 3. Define $\hat{\mathbf{T}}^{\mathrm{M}} = (\Sigma_{i\in s_B} d_i^{\mathrm{B}}, \Sigma_{i\in s_B} d_i^{\mathrm{B}} \hat{\mu}_i)$ and $\mathbf{M} = (\mathbf{1}^{\mathrm{A}}, (\hat{\mu}_i)_{i\in s_{\mathrm{A}}})$, under χ^2-distance measure with $q_i = 1$. The model-assisted calibration weights are given by

$$\mathbf{w}^{\mathrm{LASSO}} = \mathbf{d}^{\mathrm{A}} + \mathbf{D}^{\mathrm{A}}\mathbf{M}(\mathbf{M}^{\mathrm{T}}\mathbf{D}^{\mathrm{A}}\mathbf{M})^{-1}(\hat{\mathbf{T}}^{\mathrm{M}} - (\mathbf{d}^{\mathrm{A}})^{\mathrm{T}}\mathbf{M})^{\mathrm{T}}. \tag{14}$$

Step 4. The ECLASSO calibration estimator of total is then given by

$$\begin{aligned} \hat{T}_y^{\mathrm{ECLASSO}} &= (\mathbf{w}^{\mathrm{ECLASSO}})^{\mathrm{T}}\mathbf{y} \\ &= (\mathbf{d}^{\mathrm{A}})^{\mathrm{T}}\mathbf{y} + \left(\sum_{i\in s_{\mathrm{B}}} d_i^{\mathrm{B}} \hat{\mu}_i - \sum_{i\in s_{\mathrm{A}}} d_i^{\mathrm{A}} \hat{\mu}_i \right) \hat{B}^{\mathrm{MC}} \end{aligned} \tag{15}$$

where $\hat{B}^{\mathrm{MC}}$ is the calibration slope computed as in Section 2.3 to satisfy the calibration constraints.

Under conditions given in the on-line appendix—which do not require design consistent estimates of the lasso parameters $\boldsymbol{\beta}$, only that the benchmark probability sample has the correct design weights—$\hat{T}_y^{\mathrm{ECLASSO}}$ is asymptotically design and model unbiased, with the asymptotic design variance given by

$$\begin{aligned} v_{\mathscr{A}}(\hat{T}_y^{\mathrm{ECLASSO}}) = &\sum_{i\in s_{\mathrm{A}}} \left(\frac{y_i - \hat{\mu}_i \hat{B}^{\mathrm{MC}}}{\pi_i^{\mathrm{A}}} \right)^2 (1 - \pi_i^{\mathrm{A}}) + \sum_{i\in s_{\mathrm{A}}} \sum_{j\neq i} \frac{\pi_{ij}^{\mathrm{A}} - \pi_i^{\mathrm{A}}\pi_j^{\mathrm{A}}}{\pi_{ij}^{\mathrm{A}}} \frac{y_i - \hat{\mu}_i \hat{B}^{\mathrm{MC}}}{\pi_i^{\mathrm{A}}} \frac{y_j - \hat{\mu}_j \hat{B}^{\mathrm{MC}}}{\pi_j^{\mathrm{A}}} \\ &+ \sum_{i\in s_{\mathrm{B}}} \left(\frac{\hat{\mu}_i \hat{B}^{\mathrm{MC}}}{\pi_i^{\mathrm{B}}} \right)^2 (1 - \pi_i^{\mathrm{B}}) + \sum_{i\in s_{\mathrm{B}}} \sum_{j\neq i} \frac{\pi_{ij}^{\mathrm{B}} - \pi_i^{\mathrm{B}}\pi_j^{\mathrm{B}}}{\pi_{ij}^{\mathrm{B}}} \frac{\hat{\mu}_i \hat{B}^{\mathrm{MC}}}{\pi_i^{\mathrm{B}}} \frac{\hat{\mu}_j \hat{B}^{\mathrm{MC}}}{\pi_j^{\mathrm{B}}}. \end{aligned} \tag{16}$$

See the on-line appendix for proofs.

Since both linearized variance estimates are based on the asymptotic LASSO calibration estimate of a total, they might not perform well for small sample sizes. Thus we also obtain

naive bootstrap variance estimates, $v_{\text{boot}}^{\text{ECLASSO}}$, as follows: for each simulation sample, draw one finite population bootstrap of the benchmark sample, and one simple random sample with replacement of the analytical sample. For each benchmark and analytical bootstrap sample, calculate $\hat{T}_y^{\text{ECLASSO}}$.

4. Predicting the 2014 US Senate and Governors races

4.1. Data description

The on-line polling data (analytic sample) is a random sample of people who have completed a SurveyMonkey survey during the 4 weeks before the election (`http://www.surveymonkey.com`). On average, 3 million unique surveys were completed per day, with a random 10% of respondents who completed the survey receiving an invitation to complete the on-line poll. Approximately 2–3% of respondents who received the invitation completed the poll (roughly 6000 per day). Although the sample was randomly selected among the survey takers, the response rate was low and, more importantly, the pool of respondents who completed an initial SurveyMonkey survey is non-probability based and may not be representative of the voting population. The data were collected between October 3rd and November 4th, 2014 (the election day). Because conditioning on likely voters improves election prediction (Bolstein, 1991; Gutsche *et al*., 2014), we restricted our analysis to those who indicated that they

(a) had already voted,
(b) were absolutely certain to vote or
(c) were very likely to vote.

Since this paper focuses on binary outcomes, we further narrow the analytical sample to the likely voters who indicated a vote for either a Democratic or Republican candidate, which are the two major US political parties. With the further restrictions in the states to be analysed that are described below, the final analytical sample sizes are 33199 for the collection of Governor races and 28686 for the collection of Senate races.

A probability sample (benchmark sample) of potential voters was obtained by the Pew Research Center (`http://www.pewresearch.org`). Probability samples of telephone and cellphone users were selected during September and October 2014 to measure political opinions, including job approval rating for the President, agreement on recent healthcare reform policies and likelihood to vote for the November 2014 elections. The survey also includes religion and political party identification along with other demographic variables that are also collected in the SurveyMonkey sample. 'Likely voter' weights were constructed by using a 10-point-scale voting interest variable.

Our analysis focuses on states with sufficient benchmark sizes (at least 55 likely voters in a state), again restricted to support for either the Democratic and Republican parties. This yields 11 states (Arizona, California, Florida, Georgia, Illinois, Michigan, New York, Ohio, Pennsylvania, Texas and Wisconsin) for the gubernatorial elections and eight states (Georgia, Illinois, Michigan, Minnesota, New Jersey, North Carolina, Texas and Virginia) for the senatorial elections. The final benchmark sample sizes are 1094 for the collection of Governor races and 656 for collection of Senate races.

Tables 1 and 2 in the on-line appendix display the final sample size, and distributions of the common set of variables between the benchmark and election polling samples. The analytical sample distributions are unweighted, whereas the benchmark sample distributions are weighted by the likely voter weights. The Senate races have one more variable than the Governors' races—support for the House of Representatives candidate. Since both the House of Representatives

and the Senate are part of Congress, this variable is more relevant for Senate elections. The Internet-based analytical sample tends to contain individuals who are younger, more educated, white and less certain of religious beliefs. For many states, there are also much higher proportions of people identified as Republicans in the analytical sample than in the benchmark sample.

4.2. Estimation

The outcome variable y_i is an indicator for voting for a Democratic (*versus* a Republican) candidate. The analytical sample s_A is the Internet-based polling data. Let $s_A(r)$ be the sample of respondents in state r. Our target of inference is the voting spread in state r, $S_{D-R(r)}$, estimated by

$$\begin{aligned}\hat{S}_{D-R(r)} &= \sum_{i\in s_A(r)} w_i y_i \Big/ \sum_{i\in s_A(r)} w_i - \sum_{i\in s_A(r)} w_i(1-y_i) \Big/ \sum_{i\in s_A(r)} w_i \\ &= 2\sum_{i\in s_A(r)} w_i y_i \Big/ \sum_{i\in s_A(r)} w_i - 1\end{aligned}$$

where w_i is the weight for respondent i. Thus positive values are the winning margins of Democratic candidates, and absolute values of negative values are the winning margins of Republican candidates. We compare the weighted estimates based on ECLASSO with unweighted estimates UNWT, as well as estimates based on weights from traditional weighting adjustment methods—calibration to census level state demographic totals, STATEWT, propensity score weighting, PSCORE, and the estimated control regression estimator ECGREG. STATEWT uses standard post-stratification approaches to adjust to known population totals (not registered voter totals) for ages (18–29, 30–39, 40–49, 50–59, 60–74, 75 and older years), gender, race or ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, other) and education (high school or less, some college, college degree, graduate degree). PSCORE develops propensity score weights by using the benchmark sample, which, in addition to age, gender, race and education, includes religion (Protestant, Catholic, other Christian, other, none), 'born again' Evangelical, frequency of attending religious services (more than one a week, once a week, a few times a month, less than a few times a month), approval of Obama, political party favoured and five categories of state type based on their voting behaviour in the 1992–2012 presidential elections: 1, voted Republican candidate all four times, 2, voted Republican candidate three times and Democratic candidate once, 3, voted Republican and Democratic candidate each twice, 4, voted Republican candidate once and Democratic candidate three times, and 5, voted Democratic candidate all four times. In addition to these main effects, interactions between gender and age, gender and race, race and age, party and Obama approval, state type and party, and state type and Obama approval are included. Models for the Senate races also include a measure of support for the (Republican-controlled) House of Representatives. ECGREG calibrates to the estimated benchmark measures (including interactions) by using the standard GREG weights. ECLASSO uses the same estimated benchmark predictors and their interactions for the working models.

4.3. Variance estimates

For estimators that do not rely on a small benchmark sample, method UNWT and STATEWT, we estimate the variance of estimated spread $D-R$ in state r as follows:

$$\mathrm{var}(\hat{S}^{\mathrm{method}}_{D-R(r)}) = \mathrm{var}\left(2\sum_{i\in s_A(r)} w_i^{\mathrm{method}} y_i \Big/ \sum_{i\in s_A(r)} w_i^{\mathrm{method}} - 1\right) = 4\,\mathrm{var}(\hat{\bar{y}}_r^w)$$

where $\mathrm{var}(\hat{\bar{y}}_r^w)$ is the linearized variance estimator of the weighted sample mean in state r.

For estimators that use a small benchmark sample (PSCORE, ECGREG and ECLASSO), we use bootstrap variance estimates to incorporate the uncertainty of the benchmark data. For each bootstrap indexed by b, we draw a weighted bootstrap sample of the benchmark sample, and a simple random sample with replacement of the analytical sample; then we calculate the statistic

$$\hat{S}^{\text{method}}_{D-R(r)}(b) = 2 \sum_{i \in s_A(r)(b)} w_i^{\text{method}} y_i \Big/ \sum_{i \in s_A(a)(b)} w_i^{\text{method}} - 1.$$

We generate 1000 bootstrap samples and use the distribution of $\hat{S}^{\text{method}}_{D-R(r)}(b)$ to estimate the variance of $\hat{S}^{\text{method}}_{D-R(r)}$.

4.4. Results

4.4.1. Direction and error

Tables 1 and 2 list results for 11 Governor election forecasts. UNWT, STATEWT, PSCORE and ECLASSO predicted the correct winning political party for all states in the analysis. ECGREG predicted Arizona and Florida incorrectly.

We define the relative bias as $(\hat{S}^{\text{method}}_{D-R(r)} - S_{D-R(r)}) / S_{D-R(r)}$; if this is positive, the relative bias is towards the Democrats and is denoted with a D; if negative, the relative bias is towards the Republicans, denoted with an R. Without weighting adjustments, the sample has Republican overrepresentation, with 10 out of 11 states biasing towards Republican candidates. STATEWT reduced the bias for most states, whereas PSCORE and ECGREG appear to have overadjusted towards the Democratic direction. ECLASSO reduced the unadjusted absolute sample bias to a maximum of 6% of true values across the 11 states, *versus* 10–25% for the other estimators. On average, ECLASSO also has the smallest relative error across the states (0.5% D *versus* 1.9% R to 7.0% D for the other estimators).

Tables 3 and 4 list results for eight Senate election forecasts. UNWT, STATEWT and ECLASSO predicted the correct winning political party for all states in the analysis. PSCORE predicted North Carolina incorrectly whereas ECGREG predicted Georgia and North Carolina

Table 1. US 2014 midterm election Governor direction

State	*Analytical n*	*Benchmark n*	*True D − R*	*D − R estimates*				
				UNWT	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*
Arizona	974	64	+12% R	+13% R	+10% R	+3% R	+12% D	+8% R
California	2354	166	+19% D	+14% D	+19% D	+20% D	+36% D	+18% D
Florida	2566	134	+1% R	+6% R	+2% R	+2% R	+7% D	+1% R
Georgia	2306	67	+8% R	+14% R	+9% R	+10% R	+2% R	+8% R
Illinois	2955	78	+5% R	+14% R	+8% R	+14% R	+17% R	+10% R
Michigan	6025	75	+4% R	+14% R	+12% R	+12% R	+18% R	+10% R
New York	1962	106	+13% D	+13% D	+18% D	+18% D	+38% D	+17% D
Ohio	2299	87	+31% R	+35% R	+35% R	+31% R	+35% R	+31% R
Pennsylvania	2318	107	+10% D	+11% D	+8% D	+23% D	+33% D	+15% D
Texas	2575	150	+20% R	+26% R	+19% R	+20% R	+20% R	+21% R
Wisconsin	6865	60	+6% R	+6% R	+17% R	+2% R	+1% R	+1% R
Total	33199	1094						

Table 2. US midterm election voting spread estimates (RMSE = $\sqrt{(\text{Bias}^2 + \text{SE}^2)}$)

State	*Relative bias*					*SE (%)*					*RMSE (%)*				
	UNWT	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*	*UNWT*	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*	*UNWT*	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*
Arizona	1.29% R	1.63% D	8.65% D	23.51% D	3.74% D	3.18	5.07	7.04	8.51	4.26	3.43	5.33	11.15	25.01	5.67
California	4.98% R	0.50% D	1.44% D	17.44% D	0.42% R	2.04	3.07	4.72	9.90	3.18	5.38	3.11	4.94	20.05	3.20
Florida	4.69% R	0.98% R	0.50% R	8.08% D	0.02% D	1.97	3.14	6.17	5.55	3.19	5.09	3.29	6.19	9.81	3.19
Georgia	5.84% R	0.69% R	1.77% R	5.51% D	0.38% R	2.06	3.40	5.69	6.16	3.67	6.20	3.47	5.96	8.27	3.69
Illinois	9.62% R	3.86% R	9.37% R	12.89% R	5.11% R	1.82	2.81	4.42	8.93	2.97	9.79	4.77	10.36	15.68	5.91
Michigan	10.00% R	7.87% R	7.69% R	14.31% R	5.71% R	1.28	2.03	3.32	5.43	2.68	10.08	8.12	8.38	15.31	6.31
New York	0.11% R	4.83% D	4.56% D	25.16% D	4.04% D	2.24	3.30	5.12	8.61	3.06	2.24	5.85	6.85	26.60	5.06
Ohio	4.49% R	3.66% R	0.39% R	4.47% R	0.45% R	1.95	3.02	5.41	5.71	2.96	4.90	4.75	5.42	7.25	3.00
Pennsylvania	1.53% D	1.97% R	12.93% D	23.78% D	5.78% D	2.06	3.30	4.39	8.09	3.04	2.57	3.84	13.65	25.12	6.53
Texas	5.32% R	1.72% D	0.05% R	0.36% D	0.29% R	1.91	3.12	5.47	4.79	3.43	5.65	3.56	5.47	4.81	3.44
Wisconsin	0.73% R	10.79% R	3.49% D	4.60% D	4.36% D	1.20	1.84	3.66	5.79	2.94	1.41	10.94	5.06	7.39	5.26
Average	4.14% R	1.92% R	1.03% D	6.98% D	0.51% D	1.97	3.10	5.04	7.04	3.22	5.16	5.19	7.59	15.03	4.66

Table 3. US 2014 midterm election Senate direction

State	*Analytical n*	*Benchmark n*	*True D − R*	*D − R estimates*				
				UNWT	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*
Georgia	2307	67	+8% R	+13% R	+7% R	+4% R	+2% D	+11% R
Illinois	2989	78	+10% D	+1% D	+5% D	+15% D	+13% D	+6% D
Michigan	5851	75	+13% D	+5% D	+3% D	+21% D	+16% D	+8% D
Minnesota	2951	57	+10% D	+6% D	+1% D	+12% D	+6% D	+10% D
New Jersey	841	58	+13% D	+15% D	+19% D	+31% D	+34% D	+16% D
North Carolina	6093	90	+2% R	+5% R	+7% R	+1% D	+15% D	+3% R
Texas	2487	150	+27% R	+35% R	+27% R	+28% R	+27% R	+32% R
Virginia	5167	81	+1% D	+5% D	+6% D	+18% D	+24% D	+8% D
Total	28686	656						

incorrectly. Similarly to the Governor sample, the Senate sample has more Republican votes than the true voting spread, with six out of eight states biasing towards Republican candidates. STATEWT reduced the bias for the majority of states, whereas PSCORE and ECGREG over-adjusted in the Democratic direction. ECLASSO reduced the unadjusted absolute sample bias to a maximum of 8% of true values across the eight states, *versus* 9–27% for the other estimators. On average, ECLASSO also has the smallest relative error across the states (1.0% R *versus* 2.4% R to 9.0% D for the other estimators).

4.4.2. Root-mean-square error

Table 2 gives the standard error SE and root-mean-square error RMSE (the square root of the sum of the squared bias and squared SE) of each estimator in predicting Governor voting spreads. As expected, without any weighting adjustments, UNWT-estimates have the lowest standard error among the estimators. We expect the variance of STATEWT-estimates to be small, as the weights are derived from census level counts rather than from a benchmark sample. However, on average, the bias reduction of STATEWT was not enough to offset the increased variance in the estimates due to weighting, so the average RMSE of STATEWT is about the same as UNWT's. Both PSCORE and ECGREG have overadjusted the sample to produce large biases. The use of a small benchmark sample also increased the variance of the PSCORE- and ECGREG-estimates, as both estimators have larger average RMSE than UNWT's. With the same benchmark sample, working model and variance estimator as for PSCORE and ECGREG, ECLASSO can produce standard errors that are comparable with STATEWT's, and, with smaller average absolute bias, produces the lowest average RMSE across the states, with reductions of 10–69% over the other estimators.

Table 4 gives the standard error SE and root-mean-square error RMSE of each estimator in predicting Senate voting spreads. The results were similar to the gubernatorial results, with ECLASSO having average RMSE-reductions of 15–58% over the other estimators.

4.4.3. Coverage

Fig. 1 displays the plots of 90% confidence intervals computed via a normal distribution approximation based on each Governor's race estimator across 11 states, as well as the true values in the

Table 4. US 2014 midterm election voting spread estimates (RMSE = $\sqrt{(\text{Bias}^2 + \text{SE}^2)}$)

State	*Relative bias*					*SE (%)*					*RMSE (%)*				
	UNWT	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*	*UNWT*	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*	*UNWT*	*STATEWT*	*PSCORE*	*ECGREG*	*ECLASSO*
Georgia	5.63% R	0.31% D	4.12% D	9.75% D	2.83% R	2.06	3.39	5.26	4.89	3.61	5.99	3.41	6.68	10.91	4.59
Illinois	9.08% R	5.51% R	4.46% D	2.43% D	4.25% R	1.83	2.75	4.59	7.19	2.98	9.26	6.16	6.40	7.59	5.20
Michigan	8.15% R	10.05% R	7.61% D	2.60% D	5.19% R	1.31	2.06	4.40	4.46	2.81	8.25	10.26	8.79	5.16	5.90
Minnesota	4.04% R	9.23% R	1.98% D	3.98% R	0.42% R	1.84	2.76	4.35	4.00	3.20	4.44	9.63	4.78	5.64	3.23
New Jersey	2.03% D	5.99% D	17.55% D	20.70% D	3.11% D	3.41	4.79	6.72	9.24	3.79	3.97	7.67	18.79	22.66	4.90
North Carolina	3.00% R	5.28% R	2.46% D	17.11% D	1.10% R	1.28	2.07	5.10	6.59	3.23	3.27	5.67	5.66	18.33	3.41
Texas	7.76% R	0.03% D	0.52% R	0.54% D	4.51% R	1.88	3.16	4.50	4.03	3.25	7.99	3.16	4.53	4.06	5.56
Virginia	4.36% D	4.73% D	17.54% D	23.06% D	7.54% D	1.39	2.13	4.27	5.06	2.90	4.57	5.18	18.05	23.61	8.08
Average	3.91% R	2.38% R	6.90% D	9.03% D	0.96% R	1.87	2.89	4.90	5.68	3.22	5.97	6.39	9.21	12.25	5.11

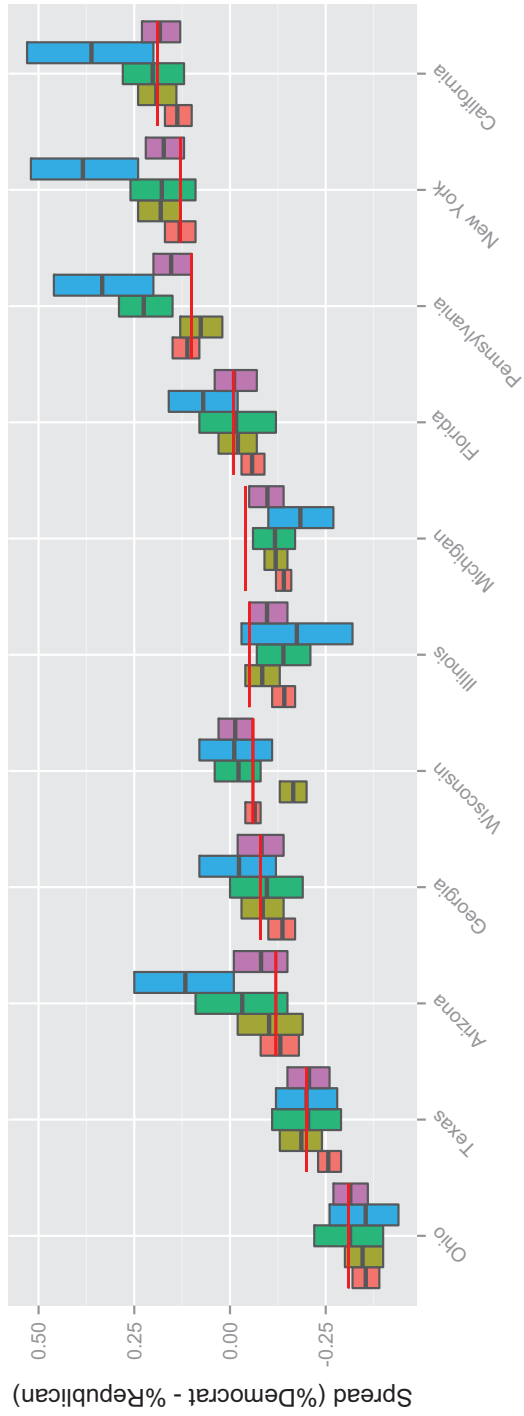

Fig. 1. Estimated voting spread for 2014 US Governors' races, together with 90% confidence intervals: ■, UNWT; ■, STATEWT; ■, PSCORE; ■, ECGREG; ■, ECLASSO; ——, true value

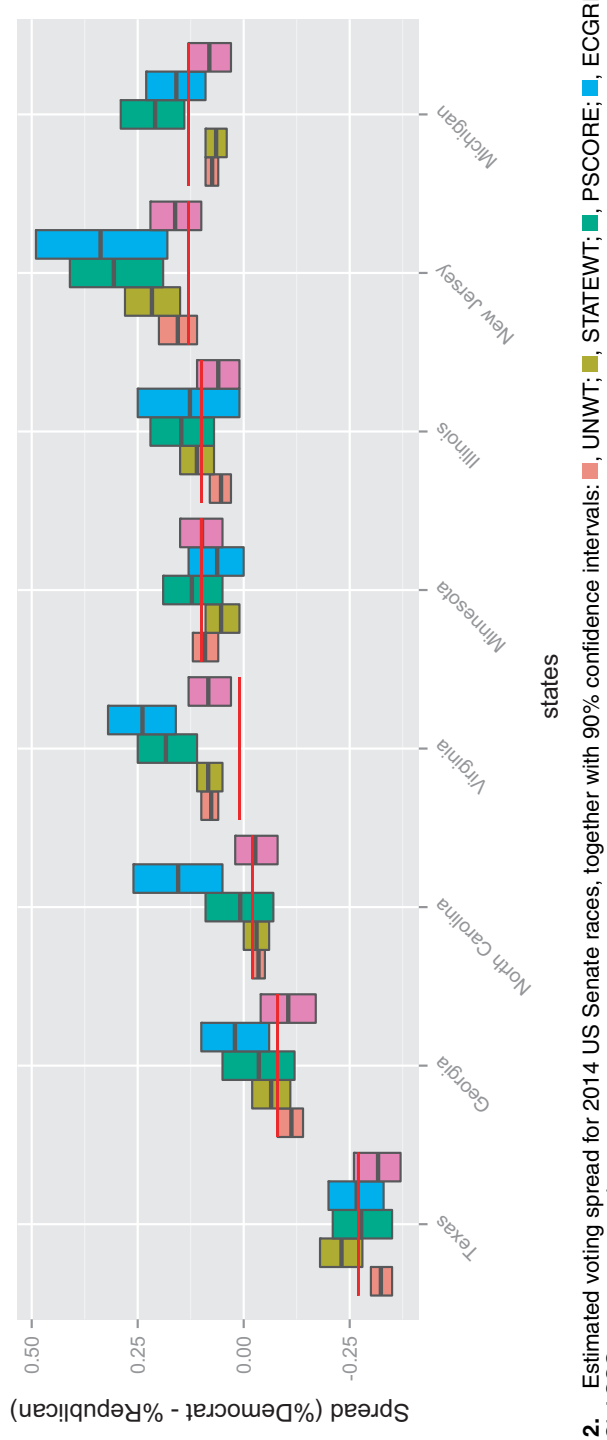

Fig. 2. Estimated voting spread for 2014 US Senate races, together with 90% confidence intervals: ■, UNWT; ■, STATEWT; ■, PSCORE; ■, ECGREG; ■, ECLASSO; ——, true value

full red horizontal lines, for the Governors' elections. The UNWT confidence intervals are too narrow, covering true spreads in only four out of 11 states (36%). ECLASSO and STATEWT confidence intervals both covered nine out of 11 true spreads (82%), which is close to the expected 90% coverage rate. PSCORE covered eight (73%), and ECGREG covered only six (55%). Among the weighted estimators, ECLASSO also has an interval width that is comparable with that of STATEWT, if not narrower.

Fig. 2 displays the plots of 90% confidence intervals based on the Senate race estimator across eight states, as well as the true values in the full red horizontal lines for the Senate elections. The UNWT confidence intervals performed even worse than for the Governor forecasts, covering only one out of eight true spreads (12%). ECLASSO confidence intervals have the highest coverage rate, with six out of eight true spreads within the intervals (75%), which is the closest to the expected 90% coverage rate among the estimators. The confidence intervals of STATEWT covered three (38%), ECGREG covered four (50%) and PSCORE covered five (62%). Aside from estimates for Virginia, where no estimator performed well, the ECLASSO confidence intervals are consistently around the true values.

5. Simulation study

Although our application is unusual in that the target parameters of interest are (eventually) known, we also conduct a simulation study, treating the 2013 National Health Interview Survey (NHIS) as the population of interest. The NHIS 2013 data are particularly suitable for simulating Internet-based non-probability samples, because the survey asks respondents about Internet use (internet_use), as well as whether a respondent has looked up health-related information on the World Wide Web (internet_health). We construct a model predicting internet_use, with internet_health as a predictor. The predicted probabilities, estimated from NHIS data, are related to both Internet usage as well as interest in health-related information on line and are used as selection probabilities to draw our simulation samples. Under such a design, if the outcome of interest is associated with the general health of a respondent, our samples will be subject to selection bias. The outcome of interest y_i is health insurance status (equal to 1 if insured; 0 if not). Restricting data records to adults and removing respondents with missing values on demographics, income and health indicators leave a population size of $N = 31914$. The goal is to predict the total number of individuals in the population without health insurance, $T_y = \Sigma_{i=1}^{N} y_i = 5432$. We use age (agegrp), gender (sex), race or ethnicity (race), education (educ), marital status (marst), employment status (wrk_private), having seen a health professional in the last year (sathc), diagnosis of cancer (cancer), family income (faminc_q), Internet use (internet_use) and obtaining health information over the internet (internet_health) as covariates in the simulation.

The main goal of the simulation is to evaluate $\hat{T}_y^{\mathrm{ECLASSO}}$ under various levels of sample and benchmark sizes. For the analytical sample, we consider $n = 250, 500, 1000$; for the benchmark sample, we consider $n = 250, 1000, 4000, 16000$. In addition to $\hat{T}_y^{\mathrm{ECLASSO}}$, we consider a Horvitz–Thompson estimator of total, assuming that an equal probability sample was selected, HT: $\hat{T}_y^{\mathrm{HT}} = (N/n)\Sigma_{i \in s_A} y_i$, as well as $\hat{T}_y^{\mathrm{GREG}}$, $\hat{T}_y^{\mathrm{ECGREG}}$ and $\hat{T}_y^{\mathrm{PSCORE}}$. To generate non-probability samples, we draw samples from the population with unequal probabilities as described in Section 5.2, but we set the design weights to N/n.

5.1. Working models

Five sets of working models are defined for the estimators. All variables are categorical, and $k[i]$ denotes the category that respondent i belongs to for a given variable:

(a) Demographics1, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{\mathrm{region}} + \beta_{k[i]}^{\mathrm{sex}} + \beta_{k[i]}^{\mathrm{agegrp}} + \beta_{k[i]}^{\mathrm{race}}$;

(b) Demographics2, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{\mathrm{region}} + \beta_{k[i]}^{\mathrm{sex}} + \beta_{k[i]}^{\mathrm{agegrp}} + \beta_{k[i]}^{\mathrm{race}} + \beta_{k[i]}^{\mathrm{educ}}$;

(c) Trimmed, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{\mathrm{sex}} + \beta_{k[i]}^{\mathrm{agegrp}} + \beta_{k[i]}^{\mathrm{race}} + \beta_{k[i]}^{\mathrm{educ}} + \beta_{k[i]}^{\mathrm{faminc\_q}} + \beta_{k[i]}^{\mathrm{employed}}$;

(d) Partial, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{\mathrm{sex}} + \beta_{k[i]}^{\mathrm{agegrp}} + \beta_{k[i]}^{\mathrm{race}} + \beta_{k[i]}^{\mathrm{educ}} + \beta_{k[i]}^{\mathrm{faminc\_q}} + \beta_{k[i]}^{\mathrm{employed}} + \beta_{k[i]}^{\mathrm{sex}} \times \beta_{k[i]}^{\mathrm{age65}} + \beta_{k[i]}^{\mathrm{race}} \times \beta_{k[i]}^{\mathrm{age65}}$;

(e) Full, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{\mathrm{sex}} + \beta_{k[i]}^{\mathrm{agegrp}} + \beta_{k[i]}^{\mathrm{race}} + \beta_{k[i]}^{\mathrm{educ}} + \beta_{k[i]}^{\mathrm{faminc\_q}} + \beta_{k[i]}^{\mathrm{employed}} + \beta_{k[i]}^{\mathrm{sex}} \times \beta_{k[i]}^{\mathrm{age65}} + \beta_{k[i]}^{\mathrm{race}} \times \beta_{k[i]}^{\mathrm{age65}} + \beta_{k[i]}^{\mathrm{race}} \times \beta_{k[i]}^{\mathrm{faminc\_q}}$.

Depending on the estimator, the $\hat{\boldsymbol{\beta}}$ is obtained differently. For GREG and ECGREG, $\hat{\boldsymbol{\beta}}$ is obtained from a linear regression of y_i on $\mathbf{x}_i$. For PSCORE, $\hat{\boldsymbol{\beta}}$ is obtained from a logistic regression of y_i on $\mathbf{x}_i$. And, for ECLASSO, $\hat{\boldsymbol{\beta}}$ is obtained through LASSO regression described in Section 3.1. Table 5 lists the regression estimates from the five working models. Except for sex, all variables are highly significant. The effect of sex is reduced once interaction terms are introduced to the model, indicating that not all interaction terms are necessary. The trimmed and partial working models may perform well. We expect all working models to help to reduce sample bias when the selection weights are ignored.

We denote GREG1 and GREG2 to be the estimators by using Demographics1 and Demographics2 respectively, which are working models that are often used for traditional calibration estimators. We expect GREG1 to perform worse than estimators using other models, because the Demographics1 model has the worst model fitness measure for the population. Demographics2 adds the education variable to Demographics1, improving model fitness substantially.

Models Trimmed, Partial and Full represent three levels of complexity. ECLASSO uses the Full model in all experimental groups. Because the larger models cannot be estimated in a stable manner from the small data sets, ECGREG and PSCORE1 use the Trimmed, Partial and Full models when the minimum of the analytical and benchmark sample size is 250, 500 and 1000 respectively.

The final estimator, PSCORE2, is the propensity score estimator that uses the correct model, i.e. the same working model as the model that generates the samples, described below.

5.2. Sample generation

The selection probabilities simulate a person's propensity to be in a non-probability Internet-based sample:

$$\begin{aligned}\mathrm{logit}(\pi_i^{\mathrm{A}}) = {} & \beta_0 + \beta_{k[i]}^{\mathrm{region}} + \beta_{k[i]}^{\mathrm{sex}} + \beta_{k[i]}^{\mathrm{agegrp}} + \beta_{k[i]}^{\mathrm{race}} + \beta_{k[i]}^{\mathrm{educ}} + \beta_{k[i]}^{\mathrm{faminc\_q}} + \beta_{k[i]}^{\mathrm{marst}} + \beta_{k[i]}^{\mathrm{sathc}} \\ & + \beta_{k[i]}^{\mathrm{wrk\_private}} + \beta_{k[i]}^{\mathrm{internet\_health}}\end{aligned}$$

where π_i^{A} is the probability of Internet use. The model is fitted to the NHIS data to obtain the predicted probabilities $\hat{\pi}_i^{\mathrm{A}}$ for each observation. These predicted probabilities are then used as selection probabilities in a Poisson sampling design. The probabilities are rescaled to generate a sample size that is close to n in expectation: $\hat{\pi}_i^{\mathrm{A}*} = n\hat{\pi}_i^{\mathrm{A}} / \Sigma_{i=1}^{N} \hat{\pi}_i^{\mathrm{A}}$.

5.3. Simulation results

The simulation results are based on 1000 simulation samples. We evaluate the empirical bias, variance and RMSE for each estimator of the total. In addition, we evaluate the linearized variance estimates and bootstrap variance estimates by their 95% nominal coverage, using a normal

Table 5. Logistic regression coefficients for working models fitted on the NHIS population for the PSCORE and ECGREG methods

Variable	*Results for the following dependent variables:*				
	Demographics1	*Demographics2*	*Trimmed*	*Partial*	*Full*
region[2]	0.199†	0.164†			
region[3]	0.519†	0.502†			
region[4]	0.403†	0.404†			
employed[1]			0.258†	0.256†	0.262†
race[2]	0.510†	0.325†	0.216†	0.208†	0.147§
race[3]	1.272†	0.911†	0.820†	0.797†	0.632†
race[4]	0.090	0.171†	0.007	−0.053	−0.331†
age65[1]			−1.954†	−2.326†	−2.360†
sex[2]	−0.262†	−0.223†	0.018	0.015	0.018
agegrp[2]	−0.100‡	−0.049	0.157†	0.158†	0.163†
agegrp[3]	−0.279†	−0.251†	0.087	0.085	0.091§
agegrp[4]	−0.442†	−0.491†	−0.129‡	−0.133‡	−0.125‡
agegrp[5]	−1.352†	−1.447†	−0.261†	−0.266†	−0.256†
agegrp[6]	−2.938†	−3.186†	−0.774†	−0.759†	−0.752†
agegrp[7]	−2.763†	−3.103†	−0.683†	−0.650‡	−0.640‡
faminc_q[1]			−0.213†	−0.211†	−0.253†
faminc_q[2]			−0.972†	−0.971†	−1.178†
faminc_q[3]			−2.109†	−2.109†	−2.253†
educ[1]		−0.414†	−0.266†	−0.262†	−0.263†
educ[2]		−0.833†	−0.588†	−0.585†	−0.592†
educ[3]		−1.187†	−0.674†	−0.672†	−0.677†
educ[4]		−2.053†	−1.191†	−1.184†	−1.186†
sathc[1]			2.057†	2.058†	2.059†
cancer[1]			−0.189‡	−0.178§	−0.180§
sex[2]:age65[1]				0.086	0.080
race[2]:age65[1]				0.195	0.236
race[3]:age65[1]				0.581†	0.649†
race[4]:age65[1]				1.375†	1.455†
race[2]:faminc_q[1]					−0.151
race[3]:faminc_q[1]					0.151
race[4]:faminc_q[1]					0.259
race[2]:faminc_q[2]					0.358†
race[3]:faminc_q[2]					0.353†
race[4]:faminc_q[2]					0.669†
race[2]:faminc_q[3]					0.303
race[3]:faminc_q[3]					0.269
race[4]:faminc_q[3]					0.440§
Constant	−1.719†	−0.869†	−1.100†	−1.088†	−1.012†

†$p < 0.01$.
‡$p < 0.05$.
§$p < 0.1$.

distribution approximation to generate confidence intervals. We ignore the finite population correction factor in variance estimation, as the sampling fraction is no more than about 0.03.

Tables 6 and 7 list the numerical summaries of each estimator under various sample and benchmark sizes. The HT, GREG1 and GREG2 estimators do not use benchmark samples. GREG1 and GREG2 control population totals by basic demographics, with GREG1 omitting the education variable.

Table 6. Simulation summary†

Sample n	*Results for estimator HT*			*Results for estimator GREG1*			*Results for estimator GREG2*		
	Bias	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
250	−383	735	828	−622	722	953	18	837	837
500	−378	520	643	−622	498	797	6	562	562
1000	−355	370	513	−602	348	695	25	399	400

†The target is the number of uninsured in the NHIS sample population: $T = 5432$.

5.3.1. Bias

As expected, assuming simple random sampling without weighting adjustment, HT underestimates the true population total. Without education as a calibration variable, GREG1 actually performed worse than HT. When education is included (GREG2), the bias is small and comparable with that of ECLASSO. This demonstrates that it may often be important to include key control totals that might only be available in benchmark samples.

Among the estimators that utilized benchmark samples, ECLASSO is the only estimator which produced unbiased estimates for all experimental groups. The PSCORE1 and PSCORE2 estimators' bias depends on both sample and benchmark sizes. For PSCORE1 and PSCORE2, the bias improves as the benchmark size increases. However, when analytic sample sizes increase for a fixed benchmark sample size, the bias tends to grow worse for PSCORE1 and especially PSCORE2. One explanation is that the sample bias persists after propensity score weighting. Thus as the sample size grows, the bias accumulates. For ECGREG, the bias remains fairly constant given different benchmark sizes and improves slightly as the analytical sample size increases.

5.3.2. RMSE

When population control variables are strongly related to both the outcome of interest and selection probabilities, we expect the traditional calibration to perform well over estimators that utilize benchmark samples. This is so for GREG2. Comparing with GREG2, ECLASSO still has gains in RMSE when the benchmark size is at least as large as the analytical sample size. For example, when the analytical sample size is 500, ECLASSO starts to have comparable and smaller RMSE relative to GREG1 for benchmark sample sizes 1000 or larger. ECLASSO produced a smaller RMSE than GREG1, even when the benchmark sample is just 250. At sample size 1000, and benchmark sample size 1000 or greater, PSCORE1, ECGREG and ECLASSO use the same working models. ECLASSO outperformed all the other methods given the same working model, suggesting that ECLASSO is most effective in leveraging information from an external benchmark sample.

5.3.3. Variance estimates

Table 8 lists the average length and the 95% nominal coverage for T_y that is obtained by using the asymptotic linearized variance estimates and naive bootstrap estimates of the ECLASSO estimator, along with the average length and the 95% nominal coverage for T_y by using the naive bootstrap estimates for the PSCORE and ECGREG estimators. The linearized variance estimates tend to undercover, with substantial undercoverage when the sample size is small.

Table 7. Simulation summary, bias, standard error SE and root-mean-square error RMSE†

Sample n	*Benchmark n*	*Result for estimator PSCORE1*			*Result for estimator PSCORE2*			*Result for estimator ECGREG*			*Result for estimator ECLASSO*		
		Bias	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
250	250	260	1052	1084	442	1268	1343	344	917	979	20	841	841
250	1000	118	827	835	109	877	884	343	826	894	28	757	758
250	4000	90	782	788	62	817	819	337	799	867	19	724	724
250	16000	93	776	781	59	805	807	339	739	862	19	714	714
500	250	258	756	799	365	868	942	328	683	757	−5	654	661
500	1000	104	576	586	116	602	614	276	582	644	−3	533	533
500	4000	79	530	535	82	549	555	274	551	616	−10	499	499
500	16000	74	520	525	74	535	541	272	546	610	−14	488	488
1000	250	318	622	698	409	698	809	320	536	624	−17	531	532
1000	1000	215	440	490	202	442	486	296	441	531	−9	404	404
1000	4000	193	395	439	180	394	433	299	410	507	−6	369	369
1000	16000	186	377	420	171	378	415	295	396	494	−11	352	352

†The target is the number of uninsured in the NHIS sample population: $T = 5432$.

Table 8. Simulation summary, coverage of the 95% nominal confidence intervals and average interval length for the number of uninsured in the NHIS sample population

Sample n	*Benchmark n*	$v_{\text{boot}}^{\text{PSCORE1}}$		$v_{\text{boot}}^{\text{PSCORE2}}$		$v_{\text{boot}}^{\text{ECGREG}}$		v^{ECLASSO}		$v_{\text{boot}}^{\text{ECLASSO}}$	
		Coverage (%)	*Length*	*Coverage (%)*	*Length*	*Coverage (%)*	*Length*	*Coverage (%)*	*Length*	*Coverage (%)*	*Length*
250	250	99.0	3286	99.1	6473	97.1	1876	88.6	1435	97.4	1925
250	1000	97.6	1786	98.4	1984	96.9	1649	88.9	1274	96.4	1619
250	4000	97.2	1618	97.3	1714	97.2	1589	88.8	1229	96.3	1531
250	16000	96.8	1589	97.0	1668	96.7	1569	89.6	1218	95.9	1509
500	250	98.9	2112	99.0	3120	96.7	1395	92.4	1236	97.0	1435
500	1000	97.1	1232	98.1	1296	90.3	1160	92.3	973	96.0	1127
500	4000	97.1	1090	97.9	1126	91.0	1095	91.5	894	96.3	1033
500	16000	97.0	1057	97.6	1088	91.2	1076	91.2	873	96.2	1008
1000	250	98.7	1590	98.9	2105	95.9	1110	93.0	991	96.1	1151
1000	1000	98.2	959	98.2	934	97.1	879	92.8	724	96.6	834
1000	4000	96.6	781	96.8	785	96.9	790	90.6	641	95.8	732
1000	16000	96.6	745	97.3	750	97.1	766	92.1	618	95.9	704

Table 9. Percentage of times that variables are selected by LASSO across 1000 simulation samples

Variable	*Result (%) for the following sample sizes:*		
	250	*500*	*1000*
employed[1]	40	47	55
sex[2]	45	48	53
race[2]	36	45	58
race[3]	74	93	99
race[4]	25	27	33
age65[1]	73	94	100
agegrp[2]	42	49	59
agegrp[3]	38	39	47
agegrp[4]	33	40	47
agegrp[5]	33	40	52
agegrp[6]	3	4	6
agegrp[7]	1	1	2
faminc_q[1]	43	44	47
faminc_q[2]	64	87	99
faminc_q[3]	98	100	100
educ2[1]	41	44	54
educ2[2]	33	40	54
educ2[3]	52	63	77
educ2[4]	42	61	81
sathc[1]	99	100	100
cancer[1]	19	23	28
sex[2]:age65[1]	4	7	8
race[2]:age65[1]	1	1	1
race[3]:age65[1]	2	2	3
race[4]:age65[1]	1	1	2
race[2]:faminc_q[1]	17	17	23
race[3]:faminc_q[1]	25	29	32
race[4]:faminc_q[1]	12	14	17
race[2]:faminc_q[2]	15	16	18
race[3]:faminc_q[2]	17	16	23
race[4]:faminc_q[2]	10	11	14
race[2]:faminc_q[3]	7	8	9
race[3]:faminc_q[3]	11	11	12
race[4]:faminc_q[3]	5	7	8

(Coverage is only slightly affected by the benchmark sample size.) The bootstrap variance estimate $v_{\text{boot}}^{\text{ECLASSO}}$ significantly overcovers when the benchmark sample is small. As both the analytical and the benchmark sample sizes increase, $v_{\text{boot}}^{\text{ECLASSO}}$ improves. The bootstrap overcoverage is worse for PSCORE1 and PSCORE2, with very wide interval lengths. As the benchmark sample size increases, the empirical coverage of the PSCORE1 and PSCORE2 bootstrap variance estimates grow closer to 95%, and the average interval length shrinks to be similar to other estimators. This suggests that the propensity score weighting adjustment method can be very sensitive to the benchmark sample sizes. ECGREG bootstrap variance estimates seem to be sensitive to the working models. For sample size $n = 500$ and benchmark sample size 500 or greater, ECGREG uses the Partial working model, which gives lower than desired coverage: around 90–91%. Given that interval widths are not small, this can be a combination of bias and model complexity—ECGREG's variances based on the Partial working model are

not sufficiently large to compensate for the bias at sample size 500. With the Full model that has more calibration cells (when the sample size is 1000 and the benchmark sample is 1000 or more), the ECGREG nominal coverages rates increase to 96–97%. Among the estimators that use benchmark samples, ECLASSO is the least sensitive to both sample and benchmark sizes, with coverages in the 96–97% range, and narrower average interval lengths than all other estimators with nominal or above coverage.

5.3.4. Adaptive LASSO model results

To gain more insight into why ECLASSO has improved performance, Table 9 lists the percentage of times that each variable is selected by LASSO across the simulation samples. The higher the percentage, the more important a variable is to predict whether a person has health insurance coverage. As the sample size increases, the proportion of times that each variable is selected by LASSO is fairly consistent for the majority of the variables, except for race[3], age65[1] and faminc_q[2], and all categories of the educ variable where the percentage increases significantly as the sample size increases. These variable categories are likely to be strong predictors of health insurance coverage that are also related to sample selection, which may explain why GREG1 performed poorly without controlling for the education variable. Age groups 6 and 7 are seldom selected by LASSO in all sample sizes, allowing ECLASSO to gain efficiency by setting these age categories to 0. Similarly, some interaction terms such as race and sex and race and age are almost always dropped, allowing ECLASSO further gains in efficiency over ECGREG under the Full model.

6. Discussion

This paper develops the framework for ECLASSO calibration and applies it to the estimation of 2014 US Governor and Senate races by using a non-probability poll of SurveyMonkey users, and to a simulation using 'Internet user' samples generated from a 'population' of the 2013 NHIS. In the application to the 2014 elections, ECLASSO was the most successful in reducing the bias in predicting voting spreads. For both Governor and Senate elections, ECLASSO reduced the overall bias from roughly 4% to under 1%. Although we expected larger variances for PSCORE, ECGREG and ECLASSO relative to the variances of STATEWT due to the small benchmark sample size, this was not so for ECLASSO, whose standard errors were comparable with STATEWT's in both races. The election data analysis shows that a benchmark sample size of 1000 is sufficient for ECLASSO to generate estimates with similar standard errors to those of estimates based on census level benchmarks. In terms of root-mean-square error and coverage, ECLASSO consistently outperforms the other estimators in both Governor and Senate election forecasts. The working models for PSCORE, ECGREG and ECLASSO are the same, indicating that ECLASSO leverages the most useful information from the benchmark.

In the simulations that were considered, the ECLASSO estimator uniformly outperforms traditional weighting adjustment methods that utilize the same benchmark data. ECLASSO could achieve the same performance as a calibration estimator controlled to a strong population level variable, even with small benchmark samples. Although the simulation models are, by definition, not inclusive of all possible applications, we expect that the key findings will be applicable across a broad range of settings: namely, that ECLASSO will allow more efficient use of high dimensional predictors, including interaction terms, that are unstable or even impossible to fit by using standard GREG estimators, that even modest benchmark sample sizes when using ECLASSO can yield substantial reductions in RMSE, especially relative to propensity

score estimators or misspecified calibration models, and that ECLASSO linearized variance estimates tend to undercover when benchmark samples are small, whereas bootstrap estimators are uniformly (if modestly) conservative.

There are many potential extensions for this work. Although ECLASSO can be extended to a multinomial setting, we stayed within a binary outcome framework and removed with non-major party supporters from the analytical sample. Another limitation is the use of a national level model to make state level forecasts. Given a small benchmark sample, the national level model enables more stable estimates by calibrating to pooled benchmark information, but alternatives that consider more complex multilevel models to smooth state level benchmark measures might be of value. Similarly, although we illustrated that the ECLASSO estimator made the most effective use of benchmark data at several different benchmark sample sizes, a topic for additional research would be to determine how large a benchmark sample should be relative to the analytic sample for ECLASSO to reduce bias most effectively without inflating mean-square errors. Finally, we have focused on the single-stage survey setting; extensions to clustered designs for model-based calibration can be developed as well (Kennel, 2013).

Although probability-based samples have always been less common outside official statistics compared with non-probability samples, their increasing expense and the proliferation of data collection from administrative sources, social media and other non-traditional sources means that methods such as those developed here will play increasingly important roles in health and social science research. Indeed, the development of methods to leverage information from probability surveys suggests a strategy of investment in a small number of very high quality probability surveys targeted towards specific research areas (e.g. behavioural health and voting behaviour) to provide calibration measures for a large set of non-probability surveys. We hope that the application discussed here will encourage such strategies.

Acknowledgements

The authors thank the Joint Editor, Associate Editor and two referees whose suggestions greatly improved the manuscript.

References

Abramowitz, A. (2008) Forecasting the 2008 presidential election with the time-for-change model. *Polit. Sci. Polit.*, **41**, 691–695.

Bolstein, R. (1991) Predicting the likelihood to vote in pre-election polls. *Statistician*, **40**, 277–283.

Dever, J. and Valliant, R. (2010) A comparison of variance estimators for poststratification to estimated control totals. *Surv. Methodol.*, **36** 45–56.

Deville, J.-C. and Sarndal, C.-E. (1992) Calibration estimators in survey sampling. *J. Am. Statist. Ass.*, **87**, 376–382.

Dutwin, D. and Lavrakas, P. (2016) Trends in telephone outcomes. *Surv. Pract.*, **9**, no. 2, 1–9.

Elliott, M. and Valliant, R. (2017) Inference for non-probability samples. *Statist. Sci.*, **32**, 249–264.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Gutsche, T., Kapteyn, A., Meijer, E. and Weerman, B. (2014) The Rand continuous 2012 presidential election poll. *Publ. Opin. Q.*, **78**, 233–254.

Kennel, T. (2013) Topics in model-assisted point and variance estimation in clustered samples. *PhD Thesis*. University of Maryland, College Park.

Kohut, A., Keeter, S., Doherty, C., Dimock, M. and Christian, L. (2012) Assessing the representativeness of public opinion surveys. Pew Research Center, Washington DC. (Available from `http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/`.)

Krosnick, J. A. (1988) The role of attitude importance in social evaluation: a study of policy preferences, presidential candidate evaluations, and voting behavior. *J. Personlty Socl Psychol.*, **55**, 196–210.

McConville, K., Bredit, F., Lee, T. and Moisen, G. (2017) Model-assisted survey regression estimation with the lasso. *J. Surv. Statist. Methodol.*, **5**, 131–158.

Mercer, A., Deane, C. and McGeeney, K. (2016) Why 2016 election polls missed their mark. Pew Research Center, Washington DC. (Available from `http://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/`.)

Schonlau, M., Zapert, K., Simon, L., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R. and Berry, S. (2004) A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Socl Sci. Comput. Rev.*, **22**, 128–138.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. and Smith, P. (2016) Report of the Inquiry into the 2015 British general election opinion polls. University of Southampton, Southampton. (Available from `http://eprints.ncrm.ac.uk/3789/`.)

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I. (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In *Proc. Int. Conf. Web and Social Media*, vol. 10, pp. 178–185.

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

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015) Forecasting elections with non-representative polls. *Int. J. Forecast.*, **31**, 980–991.

Wu, C. and Sitter, R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Statist. Ass.*, **96**, 185–193.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix'.