Calibrating Non-Probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling

Jack Kuang Tsung Chen[1], Richard L. Valliant[2] Michael R. Elliott[2,3],

[1]SurveyMonkey, Palo Alto, CA 94301, U.S.A. *email:* jjkkttcc@gmail.com

[2]Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A. *email:* valliant@umich.edu, mrelliot@umich.edu

[3]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

SUMMARY: Declining response rates and increasing costs have lead 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 U.S. 2014 midterms using large non-probability surveys obtained from SurveyMonkey users, calibrated to estimated control totals using model-assisted calibration combined with adaptive LASSO regression, or 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 proposed methodology 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: Probability survey; Propensity weighting; General regression estimator; Model-assisted calibration; Election polls

1. Introduction

One of the most prominent applications of survey research is election polling. The time-frame to collect critical voting intention is short, typically spanning just the last few weeks prior to 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 due to 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 has 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.

Recent research has shown the potential use of non-probability samples to predict election outcomes. Wang et al. (2015) performed multi-level regression and post-stratification on Xbox users to accurately predict the U.S. 2012 presidential election results. Tumasjan et al. (2010) found success in analyzing 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, compared to less than 50% male and 50% age 18-44 in the 2008 presidential election exit polls. Yet by making post-survey adjustments to match Xbox sample characteristics to 2008 exit

poll characteristics, they were able to 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 prior to 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 toward 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 using probability samples. Elliott and Valliant (2017) review work in this area, dividing methods into "quasi-likelihood" approaches (Schonlau et al., 2004) versus "superpopulation" modeling approaches (Valliant et al., 2000). The quasi-likelihood approaches includes propensity-score weighting, which combines probability and non-probability samples to generate pseudo-selection-weights for non-probability sample respondents. Superpopulation modeling 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 modeling approaches by utilizing a probability-based benchmark sample similar to the probability-based reference sample 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, while 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 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 adaptive LASSO has shown to have model-consistency properties under mild conditions (i.e., able to 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 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 manuscript is as follows. Section 2 provides background and notation for traditional post-survey weighting schemes 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, $\hat{T}_y^{ECLASSO}$. Section 4 applies ECLASSO to predict the voting spread (proportion of Demo-

cratic votes minus the proportion of Republican votes) for 11 gubernatorial elections and 8 Senate elections in the U.S. 2014 midterm election. Section 5 describes the simulation used to evaluate $\hat{T}_y^{ECLASSO}$ and the asymptotic linearized variance estimates. We summarize our findings in Section 6.

2. Post-survey Weighting Schemes for Non-probability Samples

2.1 Propensity-score weighting

Suppose 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 0 otherwise, with the propensity to be in the non-probability sample given by $p_i = Pr\left(Z_i = 1 \middle| \mathbf{X}\right)$. The propensity-score weights are simply the inverse of propensity-scores, $w_i^{PSCORE} = 1/p_i$. For an outcome of interest Y, the weighted estimates 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 has to be estimated, typically via logistic regression. Estimators of totals based on propensity-score weights are given by

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

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 dataset containing the targeted data for analysis (Y). We consider the general setting where this could be either a probability

sample with known design weights $\underset{n_A \times 1}{\mathbf{d}^A}$, 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, equivalent to assuming a simple random sample design. Defining the diagonal matrix of design or pseudo-design weights as $\mathbf{D}^A$, the calibrated weights $\underset{n_A \times 1}{\mathbf{w}}$ minimize an expected distance measure with respect to the design of A, $\mathscr{A}$ (Deville and Sarndal, 1992):

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

under the constraint $\sum_{i \in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}}$ where $\mathbf{T}^X$ 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 chi-square distance measure $g(w_i, d_i^A) = (w_i - d_i^A)^2/d_i^A$ with $q_i = 1$ yields the the generalized regression estimator (GREG):

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

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

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

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, $\mathbf{d}^A = \frac{N}{n}\mathbf{1}$ and $\mathbf{D}^A = \frac{N}{n}\mathbf{I}$.) The calibrated weights 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 $\hat{\mathbf{T}}^{\mathbf{X}}$:

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

The resulting estimator of population total is:

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

The estimate-control calibration estimator (ECGREG) has the same general form as GREG; thus we use the notation $\mathbf{w}^{ECGREG}$ and $\hat{T}_y^{ECGREG}$ 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:

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

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 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 $\sum_i^N(y_i - \mu(\mathbf{x}_i, \mathbf{B})) = 0$, and $\hat{\mathbf{B}}$ be the quasilikelihood estimator of $\mathbf{B}$ given by $\sum_{i\in s_A} d_i(y_i - \mu(\mathbf{x}_i, \hat{\mathbf{B}})) = 0$. The model-assisted calibrated weights $\mathbf{w}$ minimize a distance measure $E_{\mathscr{A}}\left[\sum_{i\in s_A} g(w_i, d_i^A)/q_i\right]$ under the constraints $\sum_{i\in s_A} w_i = N$, $\sum_{i\in s_A} w_i\hat{\mu}_i = \sum_i^N \hat{\mu}_i$, where $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$. Under chi-square distance measure with $q_i = 1$, the model-assisted calibrated weights are:

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

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

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

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

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

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 $\hat{B}^{MC}$ is different from the model parameter estimates $\hat{\mathbf{B}}$.) Wu and Sitter (2001) have shown that $\hat{T}_y^{MC}$ is asymptotically design unbiased, even when the model is misspecified. As long as the original design weights produce unbiased estimates, $\hat{T}_y^{MC}$ is approximately unbiased when the sample size is large. Similar to ECGREG, to account for uncertainties in the benchmark sample, we replace $\mathbf{T}^M = (N, \sum_{i \in U} \hat{\mu}_i)$ by estimates from a benchmark sample: $\hat{\mathbf{T}}^M = (\sum_{i \in s_B} d_i^B, \sum_{i \in s_B} d_i^B \hat{\mu}_i)$, where s_B denotes the benchmark sample and d_i^B is the probability-based design weights of the benchmark sample:

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

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 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:

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

Similarly for logistic adaptive LASSO:

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

The role of the weight parameter, α_j, is to prevent LASSO from selecting covariates with large effect sizes in favor of lowering prediction error when the sample size is small. Thus the weights are inversely proportional to effect sizes of regression parameters: $\alpha_j \propto 1/\left|\beta_j\right|$. A common choice of α_j is $1/\left|\hat{\beta}_j^{MLE}\right|$, where $\hat{\beta}_j^{MLE}$ 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 model selection process through regularization. Given that α_j is inversely proportional to β_j, small values of γ will favor covariates with large effect sizes (useful when there are known dominant predictors), while a large value of γ allows regularization to treat all covariates equally (useful when there is no prior knowledge of predictors). As there is a threshold value of λ_n that sets all regression coefficients to zero, 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 are 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 dataset, weighted versions of (12) and (13) can be fit (McConville et al., 2017); for this application we focus on the setting where the analytic dataset is a non-probability sample, and the weights d_i^A are constant and can be ignored.

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: that is, for a regression model in which the parameters have both non-zero $\boldsymbol{\beta}^{(1)}$ and zero components $\boldsymbol{\beta}^{(2)}$, $Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \rightarrow 1$ and $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \rightarrow N\left(\mathbf{0}, \mathbf{C}\right)$ 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 (ECLASSO)

The asymptotic properties of $\hat{T}_y^{ECMC}$, 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:

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 $d_i^A = 1/\pi_i^A$, the vector of design weights by $\mathbf{d}^A$, and the diagonal matrix of design weights by $\mathbf{D}^A$. A set of calibration variables is denoted by

$\mathbf{X}^A$. For non-probability samples, $\pi_i^A = \frac{n_A}{N}$ and $\pi_{ij}^A = \frac{n_A(n_A-1)}{N(N-1)}$.

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 $d_i^B = 1/\pi_i^B$, the vector of design weights by $\mathbf{d}^B$, and the diagonal matrix of design weights by $\mathbf{D}^B$. A set of calibration variables is denoted by $\mathbf{X}^B$.

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

$$E_\xi(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta})$$

$$V_\xi(y_k|\mathbf{x}_k) = \nu_k^2\sigma^2.$$

4. The true superpopulation parameters β_v are a subset of the full regression model for LASSO: $\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)}, \end{pmatrix}$, where, without loss of generality, $\boldsymbol{\beta} \equiv \boldsymbol{\beta}^{(1)}$ consists of the p non-zero components of the full model and $\boldsymbol{\beta}^{(2)} \equiv \mathbf{0}_{q\times 1}$.

5. The full-range of $\mathbf{X}$ in the population has non-zero probability of being observed in both analytical and benchmark samples. (Note that this is needed because predictions are implicitly made for the nonsample part of the population. This assumption would hold trivially if both the analytic and 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 total can be obtained following the steps:

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.

2. Use $\hat{\boldsymbol{\beta}}$ to calculate $\hat{\mu}_i = \mu(\mathbf{x}_i^A, \hat{\boldsymbol{\beta}})$ in the analytic sample, and $\hat{\mu}_i = \mu(\mathbf{x}_i^B, \hat{\boldsymbol{\beta}})$ in the benchmark sample.

3. Define $\hat{\mathbf{T}}^M = \left(\sum_{i\in s_B} d_i^B, \sum_{i\in s_B} d_i^B \hat{\mu}_i\right)$ and $\mathbf{M} = \left[\mathbf{1}^A, (\hat{\mu}_i)_{i\in s_A}\right]$, under chi-square distance measure with $q_i = 1$. The model-assisted calibration weights are given by

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

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

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

where $\hat{B}^{MC}$ 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 have the correct design weights – $\hat{T}_y^{ECLASSO}$ is asymptotically design and model-unbiased, with the

asymptotic design variance is given by

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

See the on-line appendix for proofs.

Since both linearized variance estimates are based on asymptotic LASSO calibration estimate of a total, they might not perform well for small sample sizes. Thus we also obtain naive bootstrap variance estimates, $v_{boot}^{ECLASSO}$, 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 $\hat{T}_y^{ECLASSO}$.

4. Predicting the 2014 US Senate and Governors Races

4.1 Data description

The online polling data (analytic sample) is a random sample of people who have completed a SurveyMonkey survey during the four weeks prior to 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 online poll. Approximately 2-3% of respondents receiving the invitation completed the poll (roughly 6,000 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 3^{rd} and November 4^{th}, 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 they: (1) already voted, (2) were absolutely certain to vote, or (3) were very likely to vote. Since this manuscript 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, the two major US political parties. With the further restrictions in the states to be analyzed described below, the final analytical sample sizes are 33,199 for the collection of governor races and 28,686 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 of 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 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 (AZ, CA, FL, GA, IL, MI, NY, OH, PA, TX, WI) for the gubernatorial elections and 8 states (GA, IL, MI, MN, NJ, NC, TX, VA) for the Senatorial elections. The final benchmark sample sizes are 1,094 for the collection of governor races and 656 for collection of Senate races.

Tables 1 and 2 in the online 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, while 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 House of Representatives and Senate are part of Congress, the variable is more relevant for Senate elections. The internet-based analytical sample tends to be 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:

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

where w_i is the weight for respondent i. Thus positive values are the winning margin 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 Estimated-Control Regression Estimator (ECGREG). STATEWT uses standard poststratification approaches to adjust to known population totals (not registered voter totals) for age (18-29, 30-39, 40-49, 50-59, 60-74, 75+), gender, race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, other), educa-

tion (high school or less, some college, college degree, graduate degree). PSCORE develops propensity score weights 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 favored, and five categories of state type based on their voting behavior in the 1992-2012 Presidential elections: (1) voted Republican candidate all 4 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 4 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) 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:

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

where $var\left(\hat{\bar{y}}_r^w\right)$ is the linearized variance estimator of 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 calculate the statistic:

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

We generate 1,000 bootstrap samples, and use the distribution of $\hat{S}^{method}_{D-R(r)}(b)$ to estimate the variance of $\hat{S}^{method}_{D-R(r)}$.

4.4 Results

4.4.1 Direction and error. Table 1 lists 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 relative bias as $\frac{\hat{S}^{method}_{D-R(r)} - S_{D-R(r)}}{S_{D-R(r)}}$; if this is positive, the relative bias is toward the Democrats, and is denoted with a D; if negative, the relative bias is toward the Republicans, denoted with a R. Without weighting adjustments, the sample has Republican overrepresentation, with 10 out of 11 states biasing toward Republican candidates. STATEWT reduced the bias for most states, while PSCORE and ECGREG appear to have over-adjusted toward Democratic direction. ECLASSO reduced 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).

Table 2 lists results for 8 Senate election forecasts. UNWT, STATEWT, and ECLASSO predicted the correct winning political party for all states in the analysis. PSCORE predicted North Carolina incorrectly while ECGREG predicted Georgia and North Carolina incorrectly. Similar to the governor sample, the Senate sample has more Republican votes

than the true voting spread, with 6 out of 8 states biasing toward Republican candidates. STATEWT reduced the bias for the majority of states, while PSCORE, and ECGREG over-adjusted in the Democratic direction. ECLASSO reduced unadjusted absolute sample bias to a maximum of 8% of true values across the 8 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 1 gives the standard error (SE) and root-mean-square error (RMSE) (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 anticipate 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 over-adjusted the sample to produce large biases. The use of small benchmark sample also increased the variance of 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 PSCORE and ECGREG, ECLASSO is able to produce standard errors that are comparable to STATEWT's, and, with smaller average absolute bias, produces the lowest average RMSE across the states, with reductions of 10% to 69% over the other estimators.

Table 2 gives the standard error (SE) and root-mean-square error (RMSE) of each estimator in predicting Senate voting spreads. Results were similar to the gubernatorial results, with ECLASSO having average RMSE reductions of 15% to 58% over the other estimators.

4.4.3 Coverage. Figure 1 displays the plots of 90% confidence intervals computed via normal approximation based on each governors race estimator across 11 states, as well as the true values in solid red horizontal lines, for the governors' elections. The UNWT confidence intervals are too narrow, covering true spreads in only 4 out of 11 states (36%). ECLASSO and STATEWT confidence intervals both covered 9 out of 11 true spreads (82%), close to the expected 90% coverage rate. PSCORE covered 8 (73%), and ECGREG covered only 6 (55%). Among weighted estimators, ECLASSO also has comparable interval width as STATEWT's, if not narrower.

Figure 2 displays the plots of 90% confidence intervals based on Senate rate estimator across 8 states, as well as the true values in solid red horizontal lines for the Senate elections. The UNWT confidence intervals performed even worse than governor forecasts, covering only 1 out of 8 true spreads (12%). ECLASSO confidence intervals have the highest coverage rate, with 6 out of 8 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 3 (38%), ECGREG covered 4 (50%), while PSCORE covered 5 (62%). Aside from estimates for Virginia, where no estimator performed well, 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 as the population of interest. NHIS 2013 data is 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 online, 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 (=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 the population size of $N = 31,914$. The goal is to predict the total number of individuals in the population without health insurance, $T_y = \sum_{i=1}^{N} y_i = 5,432$. We use age (*agegrp*), gender (*sex*), race/ethnicity (*race*), education (*educ*), marital (*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*), and obtaining health information over the internet (*internet_health*) as covariates in the simulation.

The main goal of the simulation is to evaluate $\hat{T}_y^{ECLASSO}$ under different 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 $\hat{T}_y^{ECLASSO}$, we consider a Horvitz-Thompson estimator of total, assuming that an equal probability sample was selected, HT: $\hat{T}_y^{HT} = (N/n)\sum_{i \in s_A} y_i$, as well as $\hat{T}_y^{GREG}$, $\hat{T}_y^{ECGREG}$, and $\hat{T}_y^{PSCORE}$. To generate non-probability samples, we draw samples from the population with unequal probabilities as described in Section 5.2, but 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 respondent i belongs to for a given variable.

- Demographics1: $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{region} + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race}$
- Demographics2: $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{region} + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ}$
- Trimmed: $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ} + \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{employed}$
- Partial: $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ} + \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{employed} + \beta_{k[i]}^{sex} \times \beta_{k[i]}^{age65} + \beta_{k[i]}^{race} \times \beta_{k[i]}^{age65}$
- Full: $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ} + \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{employed} + \beta_{k[i]}^{sex} \times \beta_{k[i]}^{age65} + \beta_{k[i]}^{race} \times \beta_{k[i]}^{age65} + \beta_{k[i]}^{race} \times \beta_{k[i]}^{faminc\_q}$

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 3 lists the regression estimates from the 5 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 reduce sample bias when the selection weights are ignored.

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

Models Trim, 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 datasets, ECGREG and PSCORE1 use the Trimmed, Partial, and Full models when the minimum of 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 one 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:

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

where π_i^A is the probability of internet use. The model is fit to the NHIS data to obtain the predicted probabilities $\hat{\pi}_i^A$ 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 close to n in expectation: $\hat{\pi}_i^{A*} = n\hat{\pi}_i^A / \sum_{i=1}^N \hat{\pi}_i^A$.

5.3 Simulation results

The simulation results are based on 1,000 simulation samples. We evaluate empirical bias, variance, and RMSE for each estimator of total. In addition, we evaluate the linearized variance estimates and bootstrap variance estimates by their 95% nominal coverage, using a normal 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.

Table 4 lists the numerical summaries of each estimator under different sample and benchmark sizes. HT, GREG1, and GREG2 estimators do not use benchmark samples.

GREG1 and GREG2 control to population totals by basic demographics, with GREG1 omitting the education variable.

5.3.1 Bias. As expected, assuming SRS 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), bias is small and comparable to 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. PSCORE1 and PSCORE2 estimators' bias depends on both sample and benchmark sizes. For PSCORE1 and PSCORE2, bias improves as benchmark size increases. However, when analytic sample sizes increase for a fixed benchmark sample size, bias tends to get worse for PSCORE1 and especially PSCORE2. One explanation is that the sample bias persists after propensity-score weighting. Thus as sample size grows, the bias accumulates. For ECGREG, the bias remains fairly constant given different benchmark sizes, and improves slightly as 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 the case for GREG2. Comparing to GREG2, ECLASSO still has gains in RMSE when benchmark size is at least as large as the analytical sample size. For example, when 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 smaller RMSE than GREG1, even when the benchmark sample is just 250. At sample size 1,000, and benchmark sample size $\geq$ 1,000, PSCORE1, ECGREG,

and ECLASSO use the same working models. ECLASSO out-performed all 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 5 lists the average length and the 95% nominal coverage for T_y obtained 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 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. (Coverage is only slightly affected by the benchmark sample size.) The bootstrap variance estimate, $v_{boot}^{ECLASSO}$, significantly over-covers when the benchmark sample is small. As both analytical and benchmark sample size increase, $v_{boot}^{ECLASSO}$ improves. The bootstrap overcoverage is worse for PSCORE1 and PSCORE2, with very wide interval lengths. As benchmark sample size increases, empirical coverage of PSCORE1 and PSCORE2 bootstrap variance estimates get closer to 95%, and the average interval length shrinks to be similar to other estimators. This suggests that 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, 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 large enough to compensate for the bias at sample size 500. With the Full model that has more calibration cells (when sample size $1,000$ and benchmark sample $\geq 1,000$), 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% to 97% range, and narrower average interval lengths that all other estimators with nominal or above coverage.

5.3.4 Adaptive lasso model results. To gain more insight into why the ECLASSO has improved performance, Table 6 lists the percentage of times 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 sample size increases, the proportion of times each variable selected by LASSO is fairly consistent for the majority of the variables, except for race[3], age65[1], faminc_q[2], and all categories of educ variable where the percentage increases significantly as sample size increases. These variable categories are likely strong predictors of health insurance coverage that are also related to sample selection, which may explain why GREG1 performed poorly without controlling to 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 manuscript develops the framework for ECLASSO calibration, and applies it to the estimation of 2014 US governor and Senate races using a non-probability poll of SurveyMonkey users, and to a simulation using "internet user" samples generated from a "population" of the 2013 National Health Interview Survey. 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 anticipated larger variances for PSCORE, ECGREG, and ECLASSO

relative to the variances of STATEWT due to the small benchmark sample size, this was not the case for ECLASSO, whose standard errors were comparable to STATEWT's in both races. The election data analysis shows that benchmark sample size of 1,000 is sufficient for ECLASSO to generate estimates with similar standard errors as estimates based on Census-level benchmarks. In terms of root-mean-square-error and coverage, ECLASSO consistently outperforms 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 considered, the ECLASSO estimator uniformly outperforms traditional weighting adjustment methods that utilize the same benchmark data. ECLASSO was able to 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 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, while 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 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 allows for 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, while 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 determining how large a benchmark sample should be relative to the analytic sample in order for ECLASSO to most effectively reduce bias 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).

While probability-based samples have always been less common outside of official statistics compared to 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, 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 toward specific research areas (e.g, behavioral health, voting behavior, etc.) to provide calibration measures for a large set of non-probability surveys. We hope the application discussed here will encourage such strategies.

References

Abramowitz, A. (2008). Forecasting the 2008 presidential election with the time-for-change model. *Political Science & Politics 41*, 691–695.

Bolstein, R. (1991). Predicting the likelihood to vote in pre-election polls. *The Statistician*,

277–283.

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

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

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

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

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

Gutsche, T., A. Kapteyn, E. Meijer, and B. Weerman (2014). The rand continuous 2012 presidential election poll. *Public Opinion Quarterly 78*, 233–254.

Kennel, T. (2013). *Topics in Model-Assisted Point and Variance Estimation in Clustered Samples*. Ph. D. thesis, University of Maryland, College Park.

Kohut, A., S. Keeter, C. Doherty, M. Dimock, and L. Christian (2012). Assessing the representativeness of public opinion surveys. `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. *Journal of Personality and Social Psychology 55*, 196.

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

Mercer, A., C. Deane, and K. McGeeney (2016). Why 2016 election polls missed their mark. http://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/.

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

Sturgis, P., N. Baker, M. Callegaro, S. Fisher, J. Green, W. Jennings, J. Kuha, B. Lauderdale, and P. Smith (2016). Report of the Inquiry into the 2015 British general election opinion polls. http://eprints.ncrm.ac.uk/3789/. [accessed 15-June-2018].

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

Tumasjan, A., T. Sprenger, P. Sandner, and I. Welpe (2010). Predicting elections with twitter: what 140 characters reveal about political sentiment. *ICWSM 10*, 178–185.

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

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

Wu, C. and R. Sitter (2001). A model-calibration approach to using complete auxiliary

information from survey data. *Journal of the American Statistical Association 96*, 185–193.

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

Figure 1: Estimated voting spread for 2014 US governor's races, together with 90% CIs

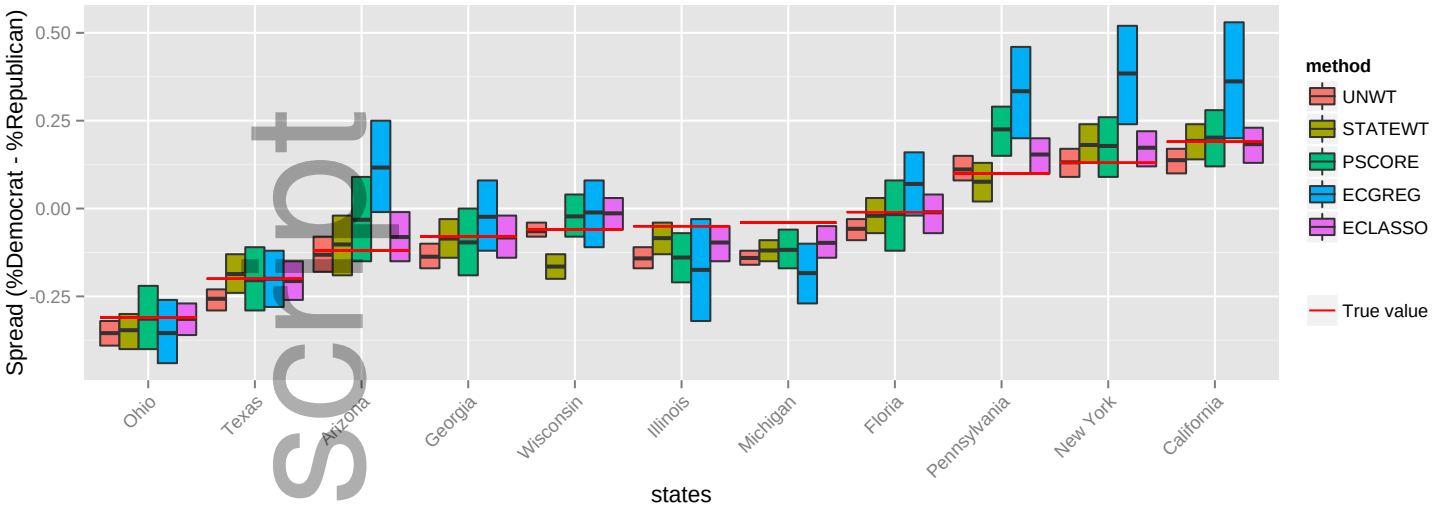

Figure 2: Estimated voting spread for 2014 US Senate races, together with 90% CIs

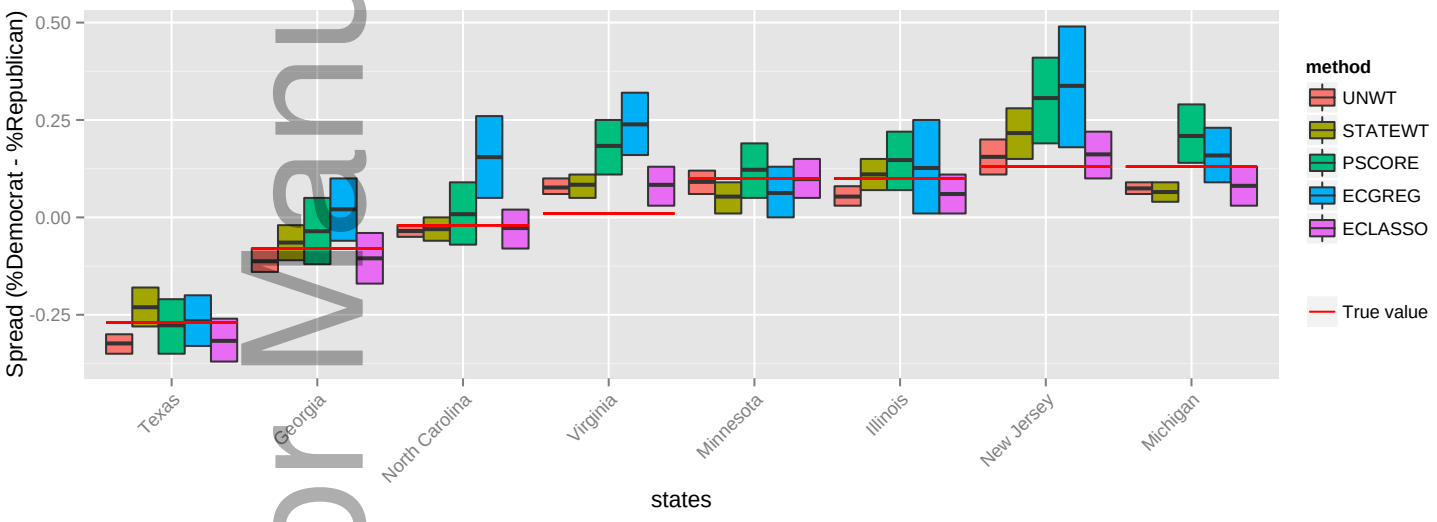

Table 1: U.S. 2014 midterm election governor direction and voting spread estimates (RMSE=$\sqrt{\text{Bias}^2 + \text{SE}^2}$)

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	2,354	166	+19%D	+14%D	+19%D	+20%D	+36%D	+18%D
Florida	2,566	134	+1%R	+6%R	+2%R	+2%R	+7%D	+1%R
Georgia	2,306	67	+8%R	+14%R	+9%R	+10%R	+2%R	+8%R
Illinois	2,955	78	+5%R	+14%R	+8%R	+14%R	+17%R	+10%R
Michigan	6,025	75	+4%R	+14%R	+12%R	+12%R	+18%R	+10%R
New York	1,962	106	+13%D	+13%D	+18%D	+18%D	+38%D	+17%D
Ohio	2,299	87	+31%R	+35%R	+35%R	+31%R	+35%R	+31%R
Pennsylvania	2,318	107	+10%D	+11%D	+8%D	+23%D	+33%D	+15%D
Texas	2,575	150	+20%R	+26%R	+19%R	+20%R	+20%R	+21%R
Wisconsin	6,865	60	+6%R	+6%R	+17%R	+2%R	+1%R	+1%R
Total	33,199	1,094						

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 2: U.S. 2014 midterm election Senate direction and voting spread estimates (RMSE=$\sqrt{\text{Bias}^2 + \text{SE}^2}$)

				D-R estimates				
State	**analytical n**	**benchmark n**	**True D-R**	**UNWT**	**STATEWT**	**PSCORE**	**ECGREG**	**ECLASSO**
Georgia	2,307	67	+8%R	+13%R	+7%R	+4%R	+2%D	+11%R
Illinois	2,989	78	+10%D	+1%D	+5%D	+15%D	+13%D	+6%D
Michigan	5,851	75	+13%D	+5%D	+3%D	+21%D	+16%D	+8%D
Minnesota	2,951	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	6,093	90	+2%R	+5%R	+7%R	+1%D	+15%D	+3%R
Texas	2,487	150	+27%R	+35%R	+27%R	+28%R	+27%R	+32%R
Virginia	5,167	81	+1%D	+5%D	+6%D	+18%D	+24%D	+8%D
Total	28,686	656						

	Relative Bias					SE					RMSE				
State	**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%

Table 3: Logistic regression coefficients for working models fit on the NHIS population for PSCORE and ECGREG methods

	Dependent variable:				
	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***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Simulation summary, target is number of uninsured in the NHIS sample "population": $T = 5,432$

	HT			GREG1			GREG2		
sample n	**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
1,000	-355	370	513	-602	348	695	25	399	400

		PSCORE1			PSCORE2			ECGREG			ECLASSO		
sample n	**benchmark n**	**bias**	**SE**	**rmse**	**bias**	**SE**	**rmse**	**bias**	**SE**	**rmse**	**bias**	**SE**	**rmse**
250	250	260	1052	1,084	442	1268	1,343	344	917	979	20	841	841
250	1,000	118	827	835	109	877	884	343	826	894	28	757	758
250	4,000	90	782	788	62	817	819	337	799	867	19	724	724
250	16,000	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	1,000	104	576	586	116	602	614	276	582	644	-3	533	533
500	4,000	79	530	535	82	549	555	274	551	616	-10	499	499
500	16,000	74	520	525	74	535	541	272	546	610	-14	488	488
1,000	250	318	622	698	409	698	809	320	536	624	-17	531	532
1,000	1,000	215	440	490	202	442	486	296	441	531	-9	404	404
1,000	4,000	193	395	439	180	394	433	299	410	507	-6	369	369
1,000	16,000	186	377	420	171	378	415	295	396	494	-11	352	352

Table 5: 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**	$v_{boot}^{PSCORE1}$		$v_{boot}^{PSCORE2}$		v_{boot}^{ECGREG}		$v^{ECLASSO}$		$v_{boot}^{ECLASSO}$	
		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	1,000	97.6%	1786	98.4%	1984	96.9%	1649	88.9%	1274	96.4%	1619
250	4,000	97.2%	1618	97.3%	1714	97.2%	1589	88.8%	1229	96.3%	1531
250	16,000	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	1,000	97.1%	1232	98.1%	1296	90.3%	1160	92.3%	973	96.0%	1127
500	4,000	97.1%	1090	97.9%	1126	91.0%	1095	91.5%	894	96.3%	1033
500	16,000	97.0%	1057	97.6%	1088	91.2%	1076	91.2%	873	96.2%	1008
1,000	250	98.7%	1590	98.9%	2105	95.9%	1110	93.0%	991	96.1%	1151
1,000	1,000	98.2%	959	98.2%	934	97.1%	879	92.8%	724	96.6%	834
1,000	4,000	96.6%	781	96.8%	785	96.9%	790	90.6%	641	95.8%	732
1,000	16,000	96.6%	745	97.3%	750	97.1%	766	92.1%	618	95.9%	704

Table 6: Percentage of times variables are selected by LASSO across 1,000 simulation samples

	Sample sizes		
Variables	**250**	**500**	**1,000**
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%