Using LASSO to Calibrate Non-probability Samples using Probability Samples

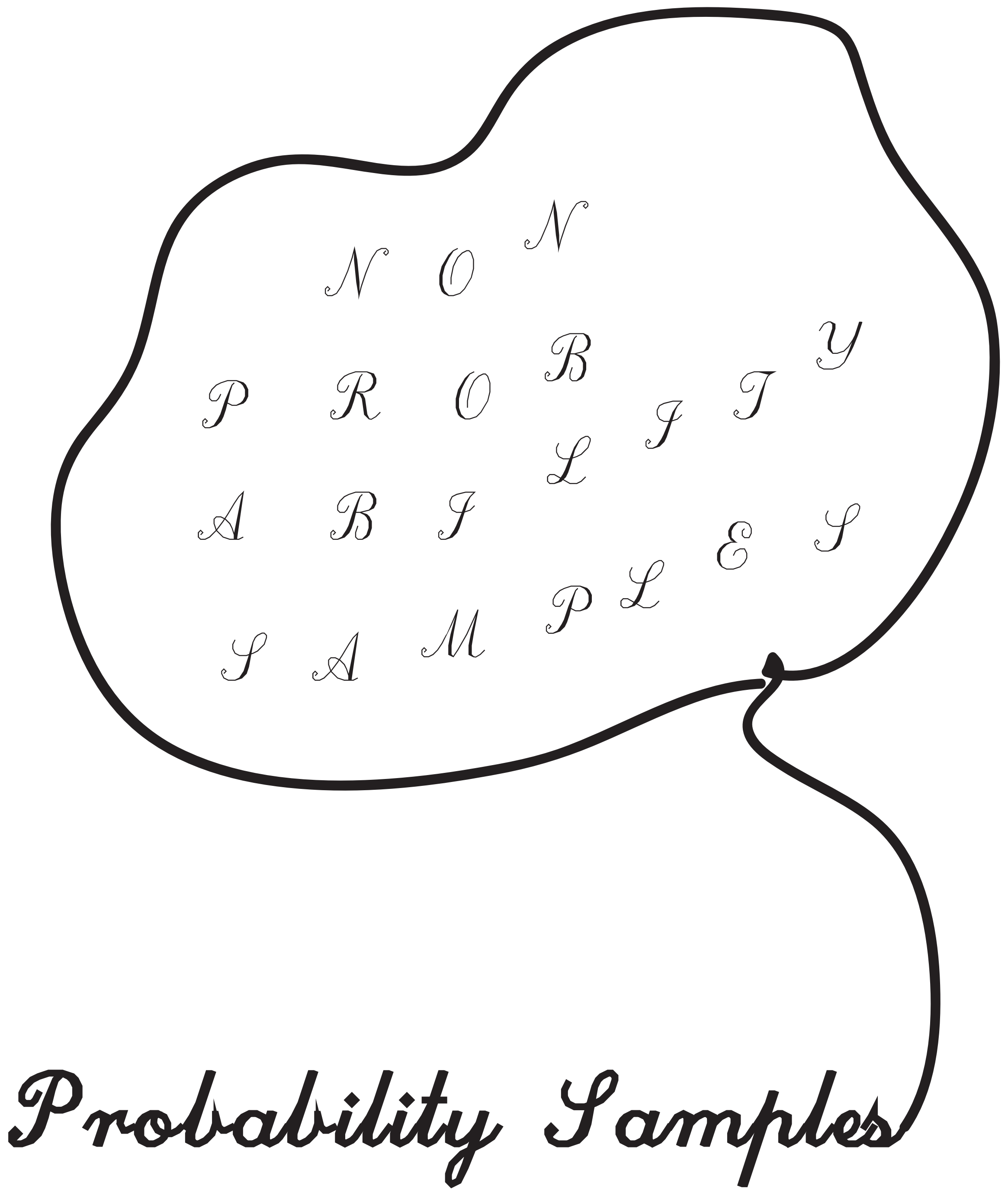
by

Kuang Tsung Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2016

Doctoral Committee:

Professor Michael R. Elliott, Chair
Professor Fred M. Feinberg
Assistant Research Scientist Sunghee Lee
Research Professor Richard L. Valliant


N O N
P R O B A B I L I T Y
S A M P L E S
Probability Samples

For my father.

ACKNOWLEDGEMENTS

I first would like to express my deepest gratitude to my adviser, Dr. Michael Elliott, for his excellent guidance, patience, and confidence in my ability to complete my research. The trajectory of my life had taken an unexpected yet fulfilling turn more than ten years ago, when Mike offered me a chance to enroll in the post-graduate Biostatistics program at the University of Michigan. While studying in this program, I first met Lucy Huang, a fellow classmate who would later become my wife. And while pursuing my degree in Survey Methodology, our little family has grown from two to four with the births of Aiden and Evelyn, our dearest son and daughter. Mike, you are the "Moon-Grandpa" in Chinese mythology (like Cupid in classical mythology), who has drawn the fateful connection between my wife and I.

I am grateful to Dr. Rick Valliant, whose casual mention of a research approach has given me the direction for my dissertation. The potential practical application of my research would not have been fully realized in the dissertation, if Dr. Sunghee Lee had not introduced me to an election-polling data. And thanks to Dr. Fred Feinberg, although being a Bayesian at heart, has graciously agreed to read through these frequentist-based materials.

I would not have been able to complete my research without the generous support from the Health and Retirement Study (HRS) at the University of Michigan. I would like to thank its director, David Weir, for giving me an opportunity to work in the HRS while pursuing my degree. And to Mary Beth Ofstedal, the most amazing supervisor a person can hope for, a simple "thank you" just cannot express this profound

grateful feelings in my heart for you. Your caring friendship, warm understanding, and steadfast support have helped me and my family through some difficult personal as well as professional challenges. You truly are one of the guardian spirits in my life.

I am forever indebted to the support I have received from my family throughout the years. To my mother, Shuan-Hwa Chen, who has always been my soul's spiritual and emotional anchor. To my sister, Victoria Chen, who has taught me the importance of distinguishing between skepticism and cynicism at work and in life. To my mother-in-law, Helen Hsieh, whose boundless energy and love have inspired me to be a better husband and father. To my father-in-law, Ken Huang, for trusting me with your amazing daughter. To aunt Amy Tung, the world's best baby-sitter, for making it possible for me to focus all my time to complete the final stretch of my dissertation. And last but not least, I would like to thank my wife, Lucy Huang, who is by far the most exuberant and optimistic person I have ever met. Your constant encouragement and unconditional support have given me great strength to continue on our life's journey. If a man's fortune is measured by the happiness that surrounds him, you have made me the luckiest person in the world.

Finally to my father in heaven, I did it!

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Using LASSO to Calibrate Non-probability Samples using Probability Samples

by

Kuang Tsung Chen

Chair: Professor Michael R. Elliott

Amidst declining response rates and rapidly increasing costs of probability-based sampling, the resurgence of more cost-effective non-probability sampling has prompted survey researchers to explore different adjustment methods for non-probability samples. The current approach attempts to create one single set of survey weights to correct all imbalances within a non-probability sample. One scheme is to generate estimated selection weights by combining the non-probability sample with a large probability-sampling-based dataset with all variables related to propensity of a respondent being in the non-probability sample. In practice, obtaining an appropriate probability sample is costly, and usually there is no way to determine the correct probability of selection for the non-probability sample, or even if all variables are available in the non-probability data to do so. An alternative approach is to adjust the non-probability sample so that the weighted sample totals of a set of variables, known as calibration variables, equal to their Census benchmark totals. Although the method does not require specialized probability-sampling-based data, the resulting calibrated weights can only correct the imbalance with respect to the limited number of Census benchmark variables, which is insufficient for adjusting all errors of a non-probability sample. To date, no method has shown to be effective in helping researchers make unbiased inference from non-probability samples.

This dissertation addresses the growing demand for making proper inference from non-probability samples. Instead of generating a single set of weights to fix all errors in a non-probability sample, we focus on constructing weights to enable unbiased inference for a specific outcome of interest. We introduce the Least Angle Shrinkage and Selection Operator, LASSO, to the framework of model-assisted calibration. The proposed method, LASSO calibration, determines the set of variables with the strongest relation to the outcome variable, then calibrates to expected outcome in a probability benchmark sample. The estimator of population total based on LASSO calibrated weights can be unbiased, regardless of how samples are generated. The theoretical framework is developed and evaluated through simulations. An application of LASSO calibration to a large-scale internet-based non-probability sample shows the proposed method can make more accurate and precise inference than existing methods.

CHAPTER I

Introduction

1.1 Objective

Probability-based sampling has dominated survey research for the greater part of the past century (*Stephan*, 1948; *Frankel and Frankel*, 1987). Given complete measures on sampled units with known selection probabilities, randomization theory removes selection bias by generating representative samples of the target population. On the other hand, non-probability samples that are generated without selection probabilities are automatically at risk for selection bias as samples can easily differ from the target population on key statistics (*Groves*, 2006). Well-documented failures in 1936 and 1948 presidential election polls highlight the potential downfalls in making population inference from non-probability samples (*Mosteller*, 1949). Although probability-sampling-based framework provides survey practitioners analytical tools to assess and correct sampling errors, declining response rates among all traditional data collection methods — mail, telephone, and face-to-face — are raising concerns over the potentially high nonresponse bias of probability samples (*Curtin et al.*, 2005; *Holbrook et al.*, 2007; *Groves*, 2011; *Kohut et al.*, 2012; *Brick and Williams*, 2013). Faced with increasing cost to conduct probability-based surveys, many researchers are turning to cheaper and more convenient non-probability sampling methods to achieve desired sample size. It is estimated that nearly half of all U.S. survey research spending will be allocated to online data collection, a platform without a universal sampling frame to conduct probability-based sampling (*Terhanian and Bremer*, 2012). With non-probability sampling quickly on the rise, the demand for practical and effective post-survey adjustment methods of non-probability samples has also increased dramatically.

Current approach to adjusting non-probability samples is met with limited success, mainly because researchers attempt to generate one set of sample weights that can work for all variables in the non-probability sample. While the one-size-fit-all is a desirable property of sample weights in public-release surveys, the highly skewed nature of non-probability samples makes it extremely challenging to construct a single set of weights capable of correcting all sample imbalance. This dissertation focuses on making proper inference from a non-probability sample for a specific outcome of interest. We construct weights under the framework of model-assisted calibration (*Wu and Sitter*, 2001). The framework uses a model to estimate expected total of the outcome in the population, then calibrates the non-probability sample with respect to the outcome variable. Unlike traditional weighting schemes, model-assisted calibration aims to reduce root-mean-square-error for the weighted estimate of a specific outcome rather than creating a general set of weights that applies to all variables. We employ the Least Angle Shrinkage and Selection Operator (*Tibshirani*, 1996), LASSO, as the assisting model in model-assisted calibration. LASSO performs estimation and variable selection simultaneously, which can determine the set of auxiliary information that is most strongly related to the outcome variable to improve calibration weighted estimates for the outcome of interest. The dissertation has two main objectives:

(1) Establish the theoretical framework for LASSO calibration in constructing weights to make inference on a population total, given auxiliary benchmark information on the population.

(2) Extend LASSO calibration for constructing weights for non-probability samples

given a small auxiliary benchmark sample.

We develop and evaluate an estimator based on the LASSO calibrated weights, as well as variance estimation methods for the estimator. We make the following assumptions:

(1) There exists a model, ξ, such that the expected value of the outcome variable, $\mathbf{y}$, can be obtained given a set of covariates $\mathbf{X}$ and a vector of parameters $\boldsymbol{\beta}$: $E_\xi\left[y_i \middle| \mathbf{x}_i, \boldsymbol{\beta}\right] = \mu(\mathbf{x}_i, \boldsymbol{\beta})$. Furthermore, the variance of y_i is a function of $\mathbf{x}_i$ or $\mu(\mathbf{x}_i, \boldsymbol{\beta})$: $V_\xi\left(y_i \middle| \mathbf{x}_i\right) = \nu_i \sigma^2$, $\nu_i = f\left(\mu(\mathbf{x}_i, \boldsymbol{\beta})\right)$ or $\nu_i = f(\mathbf{x}_i)$.

(2) The full-range of $\mathbf{X}$ in the population has non-zero probability of being observed in the non-probability sample.

Assumption (1) relates the outcome variable to the data through a superpopulation model ξ. Together with assumption (1), assumption (2) ensures that the model parameters can be estimated correctly in the non-probability sample because the relationship between $\mathbf{y}$ and $\mathbf{X}$ can be fully captured in the sample. If, for example, the outcome of interest is associated with age 65+, a sample without any respondent age 65+ would violate assumption (2). These two assumptions are the key to successful model-assisted calibration: From the non-probability sample, we can train a model that will accurately predict the expected values of the outcome variable. Then we apply the trained model to a probability-based benchmark sample to recover the distribution of $E_\xi\left[y_k \middle| \mathbf{x}_k, \hat{\boldsymbol{\beta}}\right]$ in the population. Calibrating non-probability sample y_i against population total of $E_\xi\left[y_k \middle| \mathbf{x}_k, \hat{\boldsymbol{\beta}}\right]$ will result in calibrated weights to give unbiased estimate of the outcome. Note that we make no assumption about how non-probability sample respondents participate.

Section 1.2 gives an overview of non-probability samples. Section 1.3 reviews existing post-survey adjustment methods for non-probability samples and their limitations. Sections 1.4 and 1.5 provide backgrounds of model-assisted calibration and

LASSO regression. Section 1.6 outlines the content for the remaining chapters in the dissertation.

1.2 Non-probability samples

The American Association for Public Opinion Research (AAPOR) categorized non-probability sampling into three broad categories of non-probability sampling: (1) sample matching, (2) network sampling, and (3) convenience sampling (*Couper et al.*, 2013). Sample matching is a technique in which respondents are recruited to match characteristics of a target population. A well-known sample matching method is quota sampling, which can produce proper inference if the outcome of interest is associated with quota categories. In 1936 election polls, the Literary Digest collected 2.3 million returned surveys from mostly middle-to-upper income respondents, and predicted the wrong winner with a 17% of error. At the same time, based on a quota sample of 3,000 respondents filling various income by gender quota categories, George Gallup of the American Institute of Public Opinion accurately predicted the winner with only a 5% of error (*Squire*, 1988). Quota sampling then was at the forefront of data collection methods, but did not enjoy the same success in the 1948 election polls. Since then, survey research has shifted to full probability-based sampling.

While sample matching can be viewed as a top-down approach, where desired characteristics of a sample are determined a priori to data collection, network sampling takes the bottom-up approach, where a sample starts with an initial set of respondents and gradually builds up through the respondents' social network given certain recruitment protocols (*Coleman*, 1958). An early example of networking sampling is multiplicity sampling (*Sirken*, 1970), used to enumerate households through an initial set of household rosters and expanded to households with related members on the rosters. Network sampling also has the potential to collect more data on rare populations than other sampling methods (e.g., an initial H.I.V. positive respondent can lead to a group of H.I.V. positive patients). It is a popular sampling technique in qualitative sociological research, such as the studies of drug addicts and marijuana smokers (*Lindesmith*, 1947; *Biernacki and Waldorf*, 1981). Given proper conditions (e.g., random selection of friends in referral procedures, known number of persons connected to a sampled person in their social network), we can draw population inference from network samples based on the theory of respondent-driven sampling (*Heckathorn*, 1997).

The last type of non-probability sampling, convenience sampling, is probably the most prevalent type of non-probability sampling method in practice. By definition, convenience sampling is a method where "the ease with which potential participants can be located or recruited is the primary consideration" (*Couper et al.*, 2013). While survey researchers can determine the types of respondents in quota and, to a lesser extent, network samples, there is little control over the characteristics of respondents in convenience samples. We provide more details on the common types of convenience samples and their potential errors in the following section.

1.2.1 Convenience Sampling

There are four common types of convenience samples in practice.

1. **Snowball samples**. Snowball sampling starts with a set of participants, then asks them to suggest other people who might be willing to join the study. The sample size grows quickly with each iteration of referrals, like a snowball rolling down the hill. Snowball sampling, while it is sometimes considered as a special type of network sampling, is formally categorized as convenience sampling in this work. The main difference is that, in network sampling, there is a selection and referral policy in place to establish the network, while snowball sampling recruits anyone out of convenience. It is common to observe a snowball sample of respondents with similar characteristics because they share the same

interests, activities, and remain in close contacts. Due to the selective nature of snowball sampling, it is difficult to generalize the results to the population outside of its network. Early sampling literature categorizes snowball sampling as network sampling, or chain referral sampling (*Goodman*, 1961; *Frank and Snijders*, 1994). We make the subtle point that if the network referral rule is less rigorous and more out of convenience, then the resulting sample is a snowball sample under convenience sampling.

2. **Mall intercepts**. As the name suggests, mall intercepts are collected at shopping centers to gather responses in a short amount of time. At one point, mall-intercept was the second most used data collection method, trailing only telephone surveys (*Nowell and Stanley*, 1991). In addition to missing the coverage on the population that does not have access to or go to shopping centers, selection bias can easily occur in mall intercepts when recruiters prefer specific types of shopping malls over others. For example, when recruiters only visit shopping malls at more affluent neighborhoods, measurements on shoppers' average income and spending can be much higher than those at shopping centers in a low-income neighborhood. Although certain level of probability sampling can be implemented (e.g., selecting shopping centers proportional to size, approaching every n^{th} person through the door), the typical goal of mall intercepts is to collect as many responses as possible with less emphasis (if any) on sample representativeness.

3. **River sample**. River sampling places survey invitations at designated websites to "intercept" web visitors who are willing to join the study. The method is akin to mall intercept, except the recruitment is carried out online. People who do not have internet access are not covered by river sampling. Since website contents vary greatly to attract readership of different demographics, the choice

of websites to recruit participants can greatly influence sample representation, even more so than the choice of shopping malls in mall intercept. Even if a website generates traffic for all types of web users, there can still be significant selection bias because internet users tend to be younger, more educated, and with higher income (*pewinternet.org*, 2015).

4. **Volunteer panel**. Respondents of volunteer panels actively seek and sign up to participate in a survey or study. Taking the stochastic view of nonresponse (*Groves*, 2006), we can treat each person as having a propensity to be a volunteer. From this perspective, the coverage error of volunteer panel is a function of the survey instrument. People without internet access, for instance, cannot volunteer for web surveys. A unique source of error in volunteer panel is self-selection bias. Respondents' purposeful intent to participate can result in highly skewed measures relative to the general population. For example, people with difficulty sleeping may be much more inclined to participate in a sleep behavior study. As a result, if sleep duration is an outcome of interest, the sample average hours of sleep can be much shorter than the population's. While in clinical settings, researchers can recruit from a pool of participants with diverse sleeping patterns, there is no such control in a volunteer sample. A special type of volunteer panel is web volunteer panel. Respondents of web volunteer panel join an online survey company and respond to questionnaires periodically sent out by the survey agency. Researchers in the fields of medicine, sociology, and psychology have already begun to conduct research with web opt-in panels (see, for examples, *Declercq, Sakala, Corry, and Applebaum*, 2007; *Butt, Peipert, Webster, Chen, and Cella*, 2013; *Popova and Ling*, 2014).

While there are many types of non-probability samples, convenience samples are the most prevalent for the obvious reason - they can be obtained the quickest with little cost. Unfortunately, survey researchers have the least control over the composition of convenience samples, making post-survey adjustments of convenience samples particularly challenging. In the following section, we review the common weighting adjustment methods for non-probability samples in practice.

1.3 Weighting adjustments for non-probability samples

The most recent development for adjusting non-probability samples is propensity-score weighting. Propensity-score adjustment combines the non-probability sample with a probability sample and constructs a propensity model predicting the probability of a respondent being in the non-probability sample (*Lee*, 2004; *Hill and Shaw*, 2013). Within the same propensity class, the non-probability sample respondents are matched with probability-selected respondents on all known characteristics that are used in the model. Inverse of the propensity scores serve as pseudo-selection weights for the non-probability sample. For propensity score weights to effectively remove sample bias of an outcome measure, the model covariates must be correlated to both the participation propensity in the non-probability sample and the outcome variable.

Large online survey agencies that supply non-probability samples, including Harris Interactive and Toluna, conduct expensive probability-based parallel surveys as reference samples to construct propensity models. Included in their propensity models is a set of "webographic" variables, lifestyle or attitudinal measures that should theoretically correlate with both survey measures and respondent participation propensity (*Taylor*, 2000). These specialized reference surveys have small sample sizes due to high data collection cost, which can result in highly variable propensity-score weights that can inflate variances of the weighted analysis. Furthermore, there are mixed findings on the effectiveness of propensity score adjustments when used as the only weighting adjustment method. *Schonlau et al.* (2004) found that propensity score was more effective at removing bias for categorical variables with two or more categories, but not effective in other types of measures regardless of whether they were factual or personal. With a longitudinal dataset that allowed for more in-depth analyses of webographic variables, *Schonlau et al.* (2009) concluded that propensity-score weighting with webographic variables was effective in reducing bias in health measures, but not successful for wealth-related variables. The many possibilities of a respondent to participate in non-probability samples present challenges for constructing propensity models. Thus it is common that propensity weights are only successful for a subset of the variables.

To improve propensity-score weighting, *Boboth et al.* (2007) recommended that sample calibration, i.e., using weights to "calibrate" samples such that the weighted sample matches the population on key demographic statistics and internet-use, should always accompany propensity-score adjusted weights. A simulation study by *Bethlehem* (2010) demonstrated that post-stratification, a special case of calibration, was effective in removing selection bias, provided that: (1) the benchmark data is sufficiently large, and (2) the outcome variable is similar for probability and non-probability respondents within the post-stratification cells. The latter case matches the requirements for adjustments of missing at random (MAR) in nonresponse adjustment of probability samples. Through another simulation study, *Valliant and Dever* (2011) suggested that such calibration alone can potentially achieve better bias reduction than propensity-score weighting. We are reminded by *Valliant and Dever* (2011), however, that if the outcome variable is highly correlated with propensity to be a non-probability sample respondent, no amount of weighting adjustment can completely remove sample bias. This situation falls under not missing at random (NMAR), or non-ignorable nonresponse in the probability framework. For NMAR, only models that strongly predict missing responses under the correct missigness mechanism can improve the analysis.

Both weighting adjustment methods in practice – propensity-score weighting, and calibration to benchmark totals, aim to construct one set of weights such that the weighted distributions of key statistics in the sample match those in a probability sample or in the population. Many non-probability samples are convenience samples with skewed distributions on many statistics. Thus it is unlikely that one single set of weights can achieve the goal. One major limitation in propensity-score weighting is the quality and size of the probability-based reference sample. It is costly to obtain a large reference sample, and small reference samples can result in unstable propensity weights. Furthermore, there is no systematic way to determine which variables to be included in the propensity models. In calibration weighting, we are limited by the available information on the benchmark data, which can consist of just basic demographics from large-scale government surveys. In that case, variables outside of demographics likely remain skewed. Instead of addressing overall error of the non-probability sample, we focus on improving the inferential property of a specific outcome variable, i.e. bias and variance of a weighted estimate. We choose an alternative approach with calibration, model-assisted calibration, that allows for post-survey adjustment targeting specific outcome of interest. The details of model-assisted calibration are given in the following section.

1.4 Calibration and model-assisted calibration

1.4.1 Traditional calibration

For an analytical sample s_A (the sample which requires weight calibration) of size n_A drawn from sample design $\mathcal{A}$, *Särndal and Deville* (1992) defined the term "calibrated weights", $\underset{n_A \times 1}{\mathbf{w}}$, as the adjusted weights that are as close as possible, on average, to the original design weights, $\underset{n_A \times 1}{\mathbf{d}}$, with respect to a distance measure $g(w_i, d_i)/q_i$, under the constraints that $\mathbf{w}^T\mathbf{X} = \sum_{i \in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}}$, where q_i is a constant independent of design weight d_i, $\mathbf{T}^{\mathbf{X}}$ is a row vector of known population totals of auxiliary variables in $\mathbf{X}$. Formally, $\mathbf{w}$ is the solution that minimizes:

$$E_{\mathcal{A}}\left[\sum_{i \in s_A} g(w_i, d_i)/q_i\right] \tag{1.4.1.1}$$

under the constraint:

$$\sum_{i \in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}} \tag{1.4.1.2}$$

We require that $g(w_i, d_i)$ be differentiable with respect to w_i, strictly convex on an interval containing d_i (this ensures that the local minimum of the distance function equals the solution when first derivative is zero), and $g(d_i, d_i) = 0$. The expectation in equation (1.4.1.1) is taken over sample design $\mathcal{A}$. The most common distance measure used in practice is the chi-square distance function with $q_i = 1$: $g(w_i, d_i) = (w_i - d_i)^2/d_i$. With chi-square distance, let $\mathbf{D}$ be the diagonal matrix of design weights, the calibrated weights are:

$$\mathbf{w} = \mathbf{d} + \mathbf{D}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)^T \tag{1.4.1.3}$$

The estimate of population total based on calibrated weights:

$$\begin{aligned}\hat{T} &= \mathbf{w}^T\mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}\mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}} \end{aligned} \tag{1.4.1.4}$$

where $\hat{\boldsymbol{\beta}}$ is the weighted least square estimate of the linear regression: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, given weights $\mathbf{D}$. Thus there is an implicitly assumed linear relationship between $\mathbf{y}$ and $\mathbf{X}$ for traditional calibration. The calibrated weights defined in equation (1.4.1.3) do not rely on any outcome variable. Thus the same set of weights can be applied to all variables in the survey. The calibrated weights correspond to generalized regression estimator (GREG) weights, and the weighted total expressed in equation (1.4.1.4) corresponds to the GREG estimate of total. The linear model, $E_\xi\left[y_i \middle| \mathbf{x}_i, \boldsymbol{\beta}\right] = \mathbf{x}_i^T \boldsymbol{\beta}$, is referred to as the working model for GREG.

Calibrated weights have four attractive properties: (1) They ensure that for a set of variables in the sample, the sample weighted totals match known population quantities. (2) The weights correct under-coverage of the sub-groups defined by cells in $\mathbf{X}$. (3) If there are unit non-response in the data, and the missing mechanism is missing at random (MAR) given $\mathbf{X}$, i.e. respondents and non-respondents with the same values of $\mathbf{X}$ have the same means, then calibrated weights can correct non-response bias (*Kott*, 2006). (4) If a survey outcome variable $\mathbf{y}$ has a strong linear relationship with $\mathbf{X}$, then the design-based variance of weighted estimates of $\mathbf{y}$, such as $var_{\mathcal{A}}\left(\sum_{i \in s_A} w_i y_i\right)$, is smaller than the design variance with initial design weights, $var_{\mathcal{A}}\left(\sum_{i \in s_A} d_i y_i\right)$.

There are also three side-effects from calibration: (1) Weighted estimates based on calibrated weights are no longer unbiased, but they are approximately design unbiased given large population and sample sizes and that the initial design weights are probability-based. (2) If the relationship between $\mathbf{y}$ and $\mathbf{X}$ is non-linear, variance of weighted estimates of $\mathbf{y}$ can be larger than the variance of corresponding pure-design based estimate. (3) The chi-square distance function can lead to negative weights, which do not make sense in many settings.

Särndal and Deville (1992) explored a set of distance measures and derived their calibrated weights, for which some are strictly positive. This dissertation focuses on weighted estimates rather than properties of the weights. Thus we focus on calibration with chi-square distance with $q_i = 1$. Throughout this dissertation, we will compare our proposed method with estimates from traditional calibration, GREG, since it is widely used in practice. The next section details model-assisted calibration, which forms the basis of our proposed method.

1.4.2 Model-assisted calibration

In model-assisted calibration (*Wu and Sitter*, 2001), we assume a relationship between an outcome $\mathbf{y}$ with $\mathbf{X}$ through first two moments:

$$E_\xi(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta}), V_\xi(y_k|\mathbf{x}_k) = \nu_k^2\sigma^2 \tag{1.4.2.1}$$

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 $\boldsymbol{\beta}$, 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 (or census) estimate of $\boldsymbol{\beta}$ (i.e., the quasilikelihood estimator of $\boldsymbol{\beta}$ based on the entire finite population), and $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$, where $\hat{B}$ is an estimate of $\mathbf{B}$ based on a sample of the finite population, the model-assisted calibration weights $\mathbf{w}$ minimize a distance measure:

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

under the constraints:

$$\begin{aligned}\sum_{i\in s_A} w_i &= N \\ \sum_{i\in s_A} w_i\hat{\mu}_i &= \sum_{k\in U}\hat{\mu}_k\end{aligned} \tag{1.4.2.2}$$

The main conceptual difference between traditional calibration and model-assisted calibration is that in model-assisted calibration, the constraints are based on two quantities: (1) population total, and (2) population total of predicted values $\hat{\mu}_k$. In traditional calibration, the constraint is a vector of population totals of $\mathbf{X}$ (see equation (1.4.1.2)). Define $\mathbf{T}^M = \left[N, \sum_{k\in U}\hat{\mu}_k\right]$ and $\mathbf{M} = [\mathbf{d}, (\hat{\mu}_i)_{i\in s_A}]$, under chi-square distance measure with $q_i = 1$, the model-assisted calibration weights are:

$$\mathbf{w} = \mathbf{d} + \mathbf{DM}\left(\mathbf{M}^T\mathbf{DM}\right)^{-1}\left(\mathbf{T}^M - \mathbf{d}^T\mathbf{M}\right)^T \tag{1.4.2.3}$$

The estimate for population total based on model-assisted calibrated weights are:

$$\begin{aligned}\hat{T} &= (\mathbf{w})^T\mathbf{y}\\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}\\ &= \mathbf{d}^T\mathbf{y} + \left(\sum_{k\in U}\hat{\mu}_k - \sum_{i\in s_A} d_i\hat{\mu}_i\right)B^{MC} \end{aligned}\tag{1.4.2.4}$$

where B^{MC} is the calibration slope to satisfy the calibration constraints (different from the model parameter estimates $\hat{\mathbf{B}}$):

$$\begin{aligned}B^{MC} &= \frac{\sum_{i\in s_A} d_i(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} di(\hat{\mu}_i - \hat{\bar{\mu}})^2}\\ \hat{\bar{\mu}} &= \sum_{i\in s_A} d_i\hat{\mu}_i / \sum_{i\in s_A} d_i\\ \bar{y} &= \sum_{i\in s_A} d_i y_i / \sum_{i\in s_A} d_i\end{aligned}$$

It is important to note that when the model in equation (1.4.2.1) is linear, i.e., $E_\xi(y_k|\mathbf{x}_k) = \mathbf{x}_k^T\boldsymbol{\beta}$, then we do not need individual auxiliary $\mathbf{x}_k$ values from the population. Instead, we can apply $\hat{\mathbf{B}}$ to the sum of $\mathbf{x}_k$ in the population to calculate the constraint in equation (1.4.2.2). Thus when only population totals are available, model-assisted calibration is still possible under a linear model.

When the relationship between $\mathbf{y}$ and $\mathbf{X}$ is closely captured by the superpopulation model, the resulting calibrated weights are very efficient for estimating a population quantity of the outcome variable $\mathbf{y}$ (i.e., small mean squared error). Estimating $\hat{\mathbf{B}}$ requires $\mathbf{y}$, thus model-assisted calibration weights depend on an outcome variable. The weights are constructed specifically to lower the root-mean-square error of weighted estimates of $\mathbf{y}$, and rely on how well $\hat{\mu}_i$ approximates the true μ_i. This dissertation employs a modern statistical model commonly used in predictive modeling as the assisting model to capture the relationship between $\mathbf{y}$ and $\mathbf{X}$ through μ. The resulting calibrated weights can improve root-mean-square-error of weighted estimates of $\mathbf{y}$ over traditional calibration estimates. We describe the assisting model in the next section.

1.5 LASSO

Introduced by *Tibshirani* (1996), LASSO is the acronym for "least absolute shrinkage and selection operator." LASSO falls under the general framework of regularized regression, where the solution path to regression coefficients is subject to additional constraints. The early use of mathematical regularization can be found in numerical analysis when solving for a system of equations with more unknowns than the number of equations (*Tikhonov*, 1943). There is also a vast literature in image processing and sound wavelets decoding which apply regularization techniques to filter and de-noise signals (*Donoho and Johnstone*, 1994a,b; *Abramovich and Benjamini*, 1996; *Fuchs*, 1998; *Candès et al.*, 2006; *Foucart and Rauhut*, 2013). In statistics, the main objective of regularization is to prevent model over-fitting, so that the same operator (regressors) can produce reliable estimates from different samples of the same population (*Bickel and Li*, 2008). In the past decade, the amount of research on regularized regression has grown exponentially, driven by the availability of high-dimensional data in fields such as genetics, medicine, and marketing (see, for examples: *Butler and Denham*, 2000; *Li, Sung, and Liu*, 2007; *Goldstein and Osher*, 2009; *Witten and Tibshirani*, 2009; *Wang and Zhu*, 2010). LASSO regression, in particular, has been widely used as a model selection technique in analyses involving hundreds or thousands of regressors (*Wu et al.*, 2009; *Jagannathan and Ma*, 2003), as well as a predictive model for forecasting (*Kamarianakis et al.*, 2012; *Kato and Uemura*, 2012). A wide range of applications and studies have demonstrated that LASSO regression is effective in preventing model over-fitting because it automatically selects more accurate and parsimonious models.

The capability of a model trained under a sample to make reliable estimates on a different dataset is a key feature that we need for model-assisted calibration. In the constraint equation of model-assisted calibration, equation (1.4.2.2), the sample predicted values are calibrated against predicted values in the population. If the model is prone to over-fitting in the sample, predicted values in the population would be inaccurate, resulting in unreliable calibrated weights. Thus we employ LASSO as our assisting model as it can simultaneously prevent model over-fitting through variable selection and perform parameter estimation.

1.5.1 Definition and notations

Linear LASSO is a regression of $\mathbf{y}_{n\times 1}$ on $\mathbf{X}_{n\times p}$, where the regression coefficients $\boldsymbol{\beta}_{p\times 1}$ are subject to the constraint:

$$\sum_{j=1}^{p-1} |\beta_j| \leq s \tag{1.5.1.1}$$

The intercept coefficient, β_0, is not part of the constraint. Mathematically, linear LASSO regression coefficients minimize the sum of squares plus the Lagrange multiplier of the constraint (1.5.1.1):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\underbrace{\sum_{i\in s_A} \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2}_{L} + \underbrace{\lambda_n \sum_{j=1}^{p-1} |\beta_j|}_{P} \right) \tag{1.5.1.2}$$

The L term of equation (1.5.1.2) denotes a Loss function, and P is known as the Penalty term. In LASSO, we try to find parameter estimates that minimize the Loss function subject to a penalty. Therefore LASSO is also called a penalized regression method (*Fu*, 1998).

Because we are restricting the absolute value of β_j (instead of squared or other powers of β_j) in equation (1.5.1.1), LASSO falls under L-1 regularization. When the penalty is $\sum_{j=1}^{p-1} \beta_j^2$ instead, it is under L-2 regularization, and the $\hat{\beta}$ is the solution to the regularized regression known as ridge regression. The parameter λ_n is a penalty parameter that optimizes a model-fitness measure (e.g., AIC, BIC), and is often calculated by cross-validation. The subscript n emphasizes that λ_n depends on sample size. When sample size is small, λ_n tends to be large to prevent model over-fitting by setting coefficients to zero. When sample size is large, there is less chance of model over-fitting, thus λ_n tends to 0, and $\hat{\boldsymbol{\beta}}$ resembles ordinary least square (OLS) solutions. Since the solution of $\boldsymbol{\beta}$ in LASSO regression may contain zero(s), LASSO is also used as a variable selection method. In logistic LASSO (with binary outcome $\mathbf{y}$), we try to find $\boldsymbol{\beta}$ that minimizes the negative log-likelihood function:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\underbrace{\sum_{i \in s_A} [-y_i \mathbf{x}_i' \boldsymbol{\beta} + ln(1 + exp(-\mathbf{x}_i' \boldsymbol{\beta}))]}_{L} + \underbrace{\lambda_n \sum_{j=1}^{p-1} |\beta_j|}_{P} \right) \tag{1.5.1.3}$$

1.5.2 Oracle property and adaptive LASSO

Suppose the parameters in a full regression model have both zero and non-zero components, without loss of generality, let the first p be non-zero and the last q zero:

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

The model has the oracle property if it meets the two following criteria (*Fan and Li*, 2001):

- The probability of estimating 0 for zero-valued parameters tends to one: $Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \rightarrow 1$.
- The estimates of non-zero parameters are as good as if the true sub-model is known:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \rightarrow N\left(\mathbf{0}, \mathbf{C}\right)$$

where $\mathbf{C} = \Sigma(\boldsymbol{\beta}^{(1)})$ is the covariance matrix of $\boldsymbol{\beta}^{(1)}$ under linear model, and $\mathbf{C} = I^{-1}(\boldsymbol{\beta}^{(1)})$ is the inverse of Fisher information matrix of $\boldsymbol{\beta}^{(1)}$ under generalized linear model. With the oracle property, a model not only "selects out" zero-valued parameters by setting them to 0, it also provides accurate estimates to the non-zero model parameters.

While LASSO performs both estimation and variable selection, it has been shown that in order for LASSO to have the oracle property, the regression design matrix has to satisfy fairly strict conditions, called "Irrepresentable Condition" (*Zhao and Yu*, 2006). The condition requires that covariates corresponding to the zero components of the regression parameters not contributing meaningfully to the estimation of the non-zero parameters. An example of a regression matrix satisfying irrepresentable condition is a matrix where the correlation between covariates are constant r, and there exists a constant $c > 0$ such that $0 < r \leq 1/(1 + cq)$, where q is the number of zero-valued parameters. In practice, data gathered from surveys seldom have a set of covariates with well-defined structures to satisfy irrepresentatble condition. Thus from model consistency point view, LASSO is not practical in survey research.

Many variants of LASSO have been developed. One in particular, adaptive LASSO (*Zou*, 2006), can satisfy the oracle property for both correctly specified and miss-specified models, even without the irrepresentable condition (*Zhao and Yu*, 2006). The adaptive LASSO regression coefficients are obtained by adding a weight parameter, α_j, to the penalty term:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda_n \sum_{j=1}^{p} \alpha_j^\gamma \left| \beta_j \right| \right) \qquad (1.5.2.1)$$

Similarly for adaptive logistic LASSO:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} \left[-y_i \left(\mathbf{x}_i^T \beta \right) + log \left(1 + exp \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) \right) \right] + \lambda_n \sum_{j=1}^{p} \alpha_j^\gamma \left| \beta_j \right| \right) \qquad (1.5.2.2)$$

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: $\alpha_j \propto 1 \big/ \left| \beta_j \right|$. Common choices of α_j: $\alpha_j = 1 \big/ \left| \hat{\beta}_j^{MLE} \right|$, where $\hat{\beta}_j^{MLE}$ is the maximum likelihood estimate of β_j, or $\alpha_j = 1 \big/ \left| \hat{\beta}_j^{RIDGE} \right|$, where $\hat{\beta}_j^{RIDGE}$ is the ridge regression estimates of β_j mentioned in Section 1.5.

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. For example, if we still want LASSO to favor large effect covariates when the sample size is small, we should set γ small. If we want to de-emphasize effect sizes further, we should set γ large. *Zou* (2006) has shown that the oracle property is satisfied only when:

$$\lambda_n \big/ \left(\sqrt{n} / (\sqrt{n})^\gamma \right) \to \infty \quad \text{and} \quad \lambda_n \big/ \sqrt{n} \to 0$$

The conditions require that λ_n grow at least at the rate of $\sqrt{n}/(\sqrt{n})^\gamma$, but not faster than $\sqrt{n}$. In practice, we do not observe the theoretical rate of growth of λ_n, unless a large number of samples with different sample sizes are collected from the same population. Choices of λ_n and γ are determined by the modeler. In R *glmnet* implementation (*Friedman et al.*, 2010), a range of λ_n is determined by the following scheme:

(1) Set $\gamma = 0$.

(2) Determine λ_n^{max} by finding the smallest λ_n that sets all coefficients to 0.

(3) If sample size n is larger than the number of parameters in the regression model, set $\lambda_n^{min} = 0.0001\lambda_n^{max}$. If sample size n is smaller than the number of parameters, set $\lambda_n^{min} = 0.01\lambda_n^{max}$ (to set parameters to 0 sooner).

(4) Generate a grid of λ_n, typically 100 equally spaced points between λ_n^{min} and λ_n^{max}.

The initial range of values of λ_n is determined independently of γ. With an initial range of values of λ_n, a modeler can use data-driven techniques, such as cross-validation, to find λ_n given a γ. Choices of γ are less data-driven. Some modelers choose one of $\gamma = 0.1, 0.5, 1, 2$. We can also perform cross-validation for each pair of (λ_n, γ), given a model-fitness metric (e.g. mean-absolute-error, area under curve, etc.).

1.6 Outline of chapters

The organization of the dissertation is as follows: Chapter II establishes the theoretical framework for LASSO-assisted calibration, given population auxiliary data. We derive the estimator of population total with LASSO calibrated weights, and asymptotic expectation and variance estimators for the total. The estimator and variance estimates are evaluated through simulation under different types of populations and sampling schemes. The root-mean-square of the estimator is compared with an unadjusted estimator as well as traditional calibration estimator of totals for both continuous and binary outcome variables. In Chapter III, we extend LASSO calibration to cases where the benchmark data is a probability-based sample. We introduce the estimated-control LASSO calibration estimator, ECLASSO, to estimate population totals. Asymptotic expectation and variance estimates are derived. We evaluate ECLASSO under simulation with National Health Interview Survey 2013 data as the population, given different levels of sample and benchmark sizes. The root-mean-square error of ECLASSO is compared to traditional calibration estimator, GREG, estimated-control generalized regression estimator, ECGREG, and the propensity-score weighted estimates of total, PSCORE. In Chapter IV, we apply ECLASSO to an actual non-probability internet-based election polling data. Given the actual election results, we compare root-mean-square error of election forecasts by unweighted estimate, UNWT, ECLASSO, ECGREG, and PSCORE. The final chapter, Chapter V, provides the summary, implications, and limitations of current research, as well as potential extensions for future research.

CHAPTER II

Calibration with LASSO

2.1 Introduction

For many survey agencies, adjusting survey weights to known auxiliary information is the final and most crucial step in the weight construction process. *Särndal and Deville* (1992) introduced the term "calibrated weights" as the adjusted weights that are as close as possible to the original design weights while adhering to a set of constraints. The constraints are known population totals for a set of auxiliary variables in the survey. The calibrated weights ensure that the weighted sum of each auxiliary variable equals to its corresponding total in the population. Calibration plays an important role in official statistics because it can generate weights such that the weighted demographic estimates across different surveys are consistent. Examples of large-scale surveys producing calibrated weights include Consumer Expenditure Survey (*Jayasuriya and Valiant*, 1996), Canadian Labour Force Survey (*Singh et al.*, 2001), and Survey of Health Aging and Retirement in Europe (*Börsch-Supan et al.*, 2013).

In probability samples, when design weights equal to the inverse of selection probabilities, weighted estimates of totals are design-unbiased for the population total. A main objective of calibration is to correct sample undercoverage by adjusting subgroups of the sample to their known population totals. For large samples, the final calibrated weights can be applied to all variables in the survey, because they maintain the unbiased property of original design weights. In non-probability samples, however, there are no selection probabilities to construct initial design weights that can produce unbiased estimates. Thus there is no guarantee that the traditional calibrated weights can work for all variables in the non-probability sample. To make inference from non-probability samples, one practical approach is to construct a set of weights that can lower the root-mean-square error (RMSE) of weighted estimates with respect to a specific outcome of interest. Model-assisted calibration provides the framework to construct calibrated weights targeting an outcome variable, given a model that can approximate the expected values of the outcome (*Wu and Sitter*, 2001). The key to successful model-assisted calibration is a model with strong predictive properties: model parameters estimated from one sample can be used to reliably predict values in a different sample of the same population.

The Least Angle Shrinkage and Selection Operator, LASSO, is a regularized regression that can perform both variable selection and parameter estimation (*Tibshirani*, 1996). *Kamarianakis et al.* (2012) found success with LASSO in predicting average traffic speed in the presence of severe multi-collinearity due to aggregated area-level regressors. *Kato and Uemura* (2012) applied LASSO to predict the signal of a star being observable in the sky, given a large set of periodic amplitude values. The non-signal amplitudes are considered as noise, and LASSO was successful in filtering out noise to detect the true signal. In the fields of genetics and finance, LASSO is also used in prediction modeling given hundreds or thousands of predictors (*Wu et al.*, 2009; *Wang and Zhu*, 2010). A wide range of applications have demonstrated that LASSO is effective in preventing model over-fitting by automatically selecting more accurate and parsimonious models. In survey research, under traditional calibration, *McConville* (2011) has developed the theoretical framework to show approximate design unbiasedness and consistency of LASSO calibration estimator of total for a continuous outcome variable, given LASSO regression parameter estimates. More recently, *McConville et al.* (2015) examined the use of LASSO under the model-assisted calibration framework in a simulation study that can extend LASSO calibration to estimating totals of non-continuous outcomes, and showed empirically (not theoretically) that model-assisted LASSO calibration can result in much smaller RMSE than traditional calibration. Although model-assisted calibration with LASSO holds great promise in constructing a set of weights that can result in small RMSE of weighted estimates for an outcome variable in a non-probability sample, there is no theoretical framework established for the bias and consistency properties of model-assisted LASSO calibration estimators. The main objectives of this chapter are:

(1) Develop the theoretical framework for model-assisted calibration with LASSO for both continuous and binary outcome variables: derive the point estimate of total, its asymptotic expectation, and asymptotic theoretical variance estimate.

(2) Investigate relative performances, in terms of root-mean-square-error, of LASSO calibration to traditional calibration under different outcome types, sampling schemes, sample sizes, and calibration variable covariance structures. The aim is to understand the situations where LASSO calibration can work well.

The framework for non-probability-based sampling is equivalent, except we assume the initial design weights are obtained under simple-random-sampling (SRS) regardless of how the samples are formed. The theoretical framework of LASSO calibration allows for estimating population totals from a non-probability sample with small root-mean-square-error. When the outcome of interest is binary, the LASSO calibration estimator of the total can have large gains in RMSE relative to traditional calibration estimator of the total, because traditional calibration assumes a linear relationship between the outcome and calibrated auxiliary variables. Many variants of LASSO have been developed since LASSO's introduction nearly two decades ago. The adaptive LASSO (*Zou*, 2006), in particular, has shown to have model-consistency properties, i.e., selecting the correct variables and providing unbiased estimates of parameters under mild conditions. Thus we employ the adaptive LASSO as the assisting model in model-assisted calibration. To simplify naming, we refer to adaptive LASSO simply as LASSO for the remainder of this chapter.

The organization of the chapter is as follows: Sections 2.2 and 2.3 provide the definition and notations of calibration and LASSO regression. Section 2.4 develops the LASSO calibration estimator of population total, its asymptotic estimator, and asymptotic variances. Sections 2.5 and 2.6 describes the simulation and results for evaluating the root-mean-square-error and variance estimates of $\hat{T}_y^{LASSO}$. The chapter ends with Section 2.8 summarizing the findings.

2.2 Calibration

2.2.1 Traditional calibration

For an analytical sample s_A (the sample which requires weight calibration) of size n_A drawn from sample design $\mathcal{A}$ with design weights $\underset{n_A \times 1}{\mathbf{d}}$, and the diagonal matrix of design weights $\mathbf{D}$, calibrated weights $\underset{n_A \times 1}{\mathbf{w}}$ minimize a distance measure:

$$E_{\mathcal{A}}\left[\sum_{i \in s_A} g(w_i, d_i)/q_i\right] \tag{2.2.1.1}$$

under the constraint:

$$\sum_{i \in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}} \tag{2.2.1.2}$$

where $g(w_i, d_i)$ is a differentiable function with respect to w_i, strictly convex on an interval containing d_i, and $g(d_i, d_i) = 0$. The constant q_i is independent of design weight d_i. We focus on the most common distance measure used, the chi-square distance: $g(w_i, d_i) = (w_i - d_i)^2/d_i$ with $q_i = 1$. Under this distance measure:

$$\mathbf{w}^{GREG} = \mathbf{d} + \mathbf{DX}\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)^T \tag{2.2.1.3}$$

where $\mathbf{T}^X$ is a row vector of known population totals of sample calibration variables $\mathbf{X}$. The estimate of population total of outcome $\mathbf{y}$ based on calibrated weights:

$$\begin{aligned}\hat{T}_y^{GREG} &= \mathbf{w}^T\mathbf{y}\\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}\\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}} \end{aligned} \tag{2.2.1.4}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}$ is the weighted least square estimate of the linear regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, given weights $\mathbf{D}$. The calibrated weights defined in equation (2.2.1.3) do not rely on any outcome variable. Thus the same set of weights can be applied to all variables in the survey. The weighted total expressed in equation (2.2.1.4) corresponds to the generalized regression estimator (GREG) of total, thus we denote the weights $\mathbf{w}^{GREG}$ and the estimator $\hat{T}_y^{GREG}$. In GREG, an implicit linear relationship is assumed. The linear model, $E\left[y_i \middle| \mathbf{x}_i, \boldsymbol{\beta}\right] = \mathbf{x}_i^T\boldsymbol{\beta}$, is referred to as the working model for GREG. Although $\hat{T}_y^{GREG}$ is asymptotically design-unbiased for T_y, when the relationship between $\mathbf{y}$ and $\mathbf{X}$ is non-linear, such as in the case when $\mathbf{y}$ is binary, the variance of $\hat{T}_y^{GREG}$ can be larger than the variance of pure-design based estimator of total (an estimator not using auxiliary totals). Model-assisted calibration estimators can have significant advantage over $\hat{T}_y^{GREG}$ in reducing variance of estimates of totals, because model-assisted calibration allows for non-linear models to assist in the construction of calibrated weights. In the next section, we briefly describe the framework for model-assisted calibration.

2.2.2 Model-assisted calibration

In model-assisted calibration, we assume a relationship between an outcome $\mathbf{y}$ and $\mathbf{X}$ through first two moments (*Wu and Sitter*, 2001):

$$E_{\xi}(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta}), V_{\xi}(y_k|\mathbf{x}_k) = \nu_k^2\sigma^2 \tag{2.2.2.1}$$

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 $\boldsymbol{\beta}$, 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 (or census) estimate of $\boldsymbol{\beta}$ (i.e., the quasilikelihood estimator of $\boldsymbol{\beta}$ based on the entire finite population), and $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$. The model-assisted calibrated weights $\mathbf{w}$ then minimize a distance measure:

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

under the constraints:

$$\begin{aligned} \sum_{i \in s_A} w_i &= N \\ \sum_{i \in s_A} w_i \hat{\mu}_i &= \sum_{k=1}^{N} \hat{\mu}_k \end{aligned} \tag{2.2.2.2}$$

The main conceptual difference between traditional calibration and model-assisted calibration is that in model-assisted calibration, the constraints are based on two quantities: (1) population total, and (2) population total of predicted values $\hat{\mu}_k$. In traditional calibration, the constraint is a vector of population totals of $\mathbf{X}$ (see equation (2.2.1.2)). Under chi-square distance measure with $q_i = 1$, the model-assisted calibrated weights are:

$$\mathbf{w}^{MC} = \mathbf{d} + \mathbf{DM}\left(\mathbf{M}^T\mathbf{DM}\right)^{-1}\left(\mathbf{T}^M - \mathbf{d}^T\mathbf{M}\right)^T \tag{2.2.2.3}$$

where $\mathbf{T}^M = \left[N, \sum_{k=1}^N \hat{\mu}_k\right]$ and $\mathbf{M} = [\mathbf{d}, (\hat{\mu}_i)_{i\in s_A}]$. The estimate for the population total based on model-assisted calibrated weights is then:

$$\begin{aligned}\hat{T}_y^{MC} &= \left(\mathbf{w}^{MC}\right)^T \mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy} \\ &= \mathbf{d}^T\mathbf{y} + \left(\sum_{k=1}^N \hat{\mu}_k - \sum_{i\in s_A} d_i\hat{\mu}_i\right)\hat{B}^{MC} \qquad (2.2.2.4)\end{aligned}$$

where $\hat{B}^{MC}$ is the calibration slope to satisfy the calibration constraints (different from the model parameter estimates $\hat{\mathbf{B}}$):

$$\begin{aligned}\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i(\hat{\mu}_i - \bar{\hat{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} d_i(\hat{\mu}_i - \bar{\hat{\mu}})^2} \\ \bar{\hat{\mu}} &= \sum_{i\in s_A} d_i\hat{\mu}_i \Big/ \sum_{i\in s_A} d_i \\ \bar{y} &= \sum_{i\in s_A} d_i y_i \Big/ \sum_{i\in s_A} d_i\end{aligned}$$

Unbiasedness and small variances of $\hat{T}_y^{MC}$ both rely on how well the $\hat{\mu}_i$ approximates the true expected value of y_i.

2.3 LASSO

2.3.1 Definition and parameters

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \left| \beta_j \right| \right) \tag{2.3.1.1}$$

Similarly for logistic adaptive LASSO:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} \left[-y_i \left(\mathbf{x}_i^T \beta \right) + log \left(1 + exp \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) \right) \right] + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \left| \beta_j \right| \right) \tag{2.3.1.2}$$

Given λ_n and γ, we can calculate $\hat{\boldsymbol{\beta}}$ through some iterative procedures.

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: $\alpha_j = 1/\left|\hat{\beta}_j^{MLE}\right|$, 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. For example, if we still want LASSO to favor large effect covariates when the sample size is small, we should set γ small. If we want to de-emphasize effect sizes further, we should set γ large.

2.3.2 Oracle property

An important concept in measuring the performance of a model selection and estimation method is called the “oracle property”. The optimal method selects the correct variables and provides unbiased estimates to selected parameters. Suppose the parameters in a full regression model have both zero and non-zero components, without loss of generality, let the first p be non-zero and the last q zero:

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

A regression model has the oracle property if it satisfies the following conditions *Fan and Li* (2001):

- The probability of estimating 0 for zero-valued parameters tends to one: $Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \to 1$.
- The estimates of non-zero parameters are as good as if the true sub-model is known:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \to N\left(\mathbf{0}, \mathbf{C}\right)$$

where $\mathbf{C} = \Sigma(\boldsymbol{\beta}^{(1)})$ is the covariance matrix of $\boldsymbol{\beta}^{(1)}$ under linear model, and $\mathbf{C} = I^{-1}(\boldsymbol{\beta}^{(1)})$ is the inverse of Fisher information matrix of $\boldsymbol{\beta}^{(1)}$ under generalized linear model. For finite-population inference, suppose ν indexes a population with size N_ν, let $\mathbf{B}$ be the quasilikelihood estimates of $\boldsymbol{\beta}$ in population ν, and $\hat{\mathbf{B}}$ is the estimate of $\mathbf{B}$ based on a sample with size $n_\nu \leq N_\nu$, the finite-population equivalent of the oracle property is:

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

where $\mathbf{C}_\nu = \Sigma(\mathbf{B}^{(1)})$ is the covariance matrix of $\mathbf{B}^{(1)}$ under linear model, and $\mathbf{C} = I^{-1}(\mathbf{B}^{(1)})$ is the inverse of Fisher information matrix of $\mathbf{B}^{(1)}$ under generalized linear model. For convenience, we omit ν from the notations. It is assumed that N and n are sequences of numbers, both grow to infinity as $\nu \to \infty$. We write $\mathbf{B} \to \boldsymbol{\beta}$ to mean that $\mathbf{B}$ approaches $\boldsymbol{\beta}$ as both sample and population sizes grow.

Zou (2006) has shown that if:

$$\lambda_n / \left(\sqrt{n}/(\sqrt{n})^\gamma\right) \to \infty \quad \text{and} \quad \lambda_n/\sqrt{n} \to 0$$

then the adaptive LASSO satisfies the oracle property. The conditions require that λ_n grow at least at the rate of $\sqrt{n}/(\sqrt{n})^\gamma$, but not faster than $\sqrt{n}$. We discuss the choice of λ_n and γ in the next section.

2.3.3 Determining parameter values and estimates

In practice, we do not observe the theoretical rate of growth of λ_n, unless we have obtained many samples of the same population with various sample sizes. Given a sample, the choices of λ_n and γ depend on the modeler. In R *glmnet* implementation (*Friedman et al.*, 2010), a range of λ_n is determined by the following scheme:

(1) Set $\gamma = 0$.

(2) Determine λ_n^{max} by finding the smallest λ_n that sets all coefficients to 0.

(3) If sample size n is larger than the number of parameters in the regression model, set $\lambda_n^{min} = 0.0001\lambda_n^{max}$. If sample size n is smaller than the number of parameters, set $\lambda_n^{min} = 0.01\lambda_n^{max}$ (to set parameters to 0 sooner).

(4) Generate a grid of λ_n, typically 100 equally spaced points between λ_n^{min} and λ_n^{max}.

The initial range of values of λ_n is determined independently of γ. Choices of γ is less data-driven. Some modelers choose one of $\gamma = 0.1, 0.5, 1, 2$. In this chapter, we determine (λ_n, γ) through cross-validation as follows:

Step 1. Obtain $\alpha_j = 1/\left|\hat{\beta}_j^{MLE}\right|$

Step 2. Determine 100 equally spaced values of λ_n based on R *glmnet*'s implementation.

Step 3. For each pair (λ_n, γ), λ_n from Step 2, and $\gamma = 0.1, 0.5, 1, 2$, split data into 5 folds. Use 4 folds to obtain $\hat{\boldsymbol{\beta}}$.

Step 4. Apply $\hat{\boldsymbol{\beta}}$ to the last fold not used to estimate $\hat{\boldsymbol{\beta}}$ and calculate a metric. For continuous $\mathbf{y}$, we calculate the mean-absolute-error (MAE), $\sum_{i \in s_{A(k)}} |\hat{\mu}_i - y_i|$. For binary $\mathbf{y}$, we calculate the area under curve (AUC) (calculated through R *glmnet* :: *auc* function).

Step 5. Average the 5 metrics for each pair of (λ_n, γ), and choose the pair with the best average metric: minimum MAE for continuous $\mathbf{y}$, maximum AUC for binary $\mathbf{y}$.

The adaptive LASSO coefficient estimates are then obtained by solving (2.3.1.1) or (2.3.1.2) given the selected (λ_n, γ). The R code used to perform cross-validation in this dissertation is in Appendix A.2.

2.4 LASSO calibration

This section develops the main theoretical framework of this chapter. We derive the analytical formula for LASSO estimator of total, its asymptotic expectation, and asymptotic linearized variance estimates. We make the following assumptions in the theoretical framework:

A. The samples are drawn from a single-stage sample design $\mathcal{A}$, allowing for unequal probabilities of selection. The selection probability for unit i is denoted by π_i^A, and the joint selection probability of units i and j is 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 $\mathbf{D}^A$.

B. Population-level auxiliary data are known, denoted by $\mathbf{X} = (\mathbf{x}_k^T),\ k = 1, \cdots, N$.

C. A superpopulation model is assumed, as is described in section 2.3.3:

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

D. The true superpopulation parameters are a subset of the full regression model for LASSO: $\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}$

E. The full-range of $\mathbf{X}$ in the population has non-zero probability of being observed in the analytical sample.

2.4.1 Point estimate: $\hat{T}_y^{LASSO}$

The LASSO calibration estimate of total can be obtained following the steps:

Step 1. Obtain LASSO regression coefficients $\hat{\mathbf{B}}$ as described in section 2.3. We use R package *glmnet* (*Friedman et al.*, 2010) to obtain LASSO coefficients for both linear and glm models, given a pair of (λ_n, γ) selected by cross-validation. Linear LASSO:

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

Logistic LASSO:

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

Step 2. Use $\hat{\mathbf{B}}$ to calculate $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$ in the sample and the population.

Step 3. Define $\mathbf{T}^M = \left(N, \sum_{k=1}^N \hat{\mu}\right)$ and $\mathbf{M} = \left[\mathbf{d}^A, (\hat{\mu}_i)_{i\in s_A}\right]$, under chi-square distance measure with $q_i = 1$:

$$\mathbf{w}^{LASSO} = \mathbf{d}^A + \mathbf{D}^A\mathbf{M}\left(\mathbf{M}^T\mathbf{D}^A\mathbf{M}\right)^{-1}\left(\mathbf{T}^M - (\mathbf{d}^A)^T\mathbf{M}\right)^T \tag{2.4.1.1}$$

Step 4. LASSO calibration estimator of total:

$$\begin{aligned}\hat{T}_y^{LASSO} &= \left(\mathbf{w}^{LASSO}\right)^T\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{D}^A\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}^A\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\sum_{k=1}^N \hat{\mu}_k - \sum_{i\in s_A} d_i^A\hat{\mu}_i\right)\hat{B}^{MC}\end{aligned} \tag{2.4.1.2}$$

where $\hat{B}^{MC}$ is the calibration slope to satisfy the calibration constraints (different from the model parameter estimates $\hat{\mathbf{B}}$):

$$\begin{aligned}\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})^2} \\ \hat{\bar{\mu}} &= \sum_{i\in s_A} d_i^A\hat{\mu}_i / \sum_{i\in s_A} d_i^A \\ \bar{y} &= \sum_{i\in s_A} d_i^A y_i / \sum_{i\in s_A} d_i^A\end{aligned}$$

2.4.2 Asymptotic estimator of total

Wu and Sitter (2001) established the conditions to derive asymptotic model-assisted calibration estimator. We state the conditions here with slight modification in notations to be consistent with the current research. Let $\boldsymbol{\beta}$ be the true superpopulation parameter for the model defined in equation (2.2.2.1), and $\mathbf{B}$ be the finite-population quasilikelihood estimator of $\boldsymbol{\beta}$, the following conditions are used for deriving LASSO calibration asymptotic estimator of total:

(2.4.2.i) $\hat{\mathbf{B}} = \mathbf{B} + O_p(1/\sqrt{n})$, $\mathbf{B}$ is the finite-population regression slope of $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$.

(2.4.2.ii) For each $\mathbf{x}_i$, $\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}$ is continuous in $\mathbf{t}$, and $max_i\, |\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}| \leq h(\mathbf{x}_i, \boldsymbol{\beta})$ for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\beta}$, and $N^{-1}\sum_{i\in U} h(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.

(2.4.2.iii) For each $\mathbf{x}_i$, $\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}\partial\mathbf{t}^T$ is continuous in $\mathbf{t}$, and $max_{j,k}\, |\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial t_j \partial t_k| \leq k(\mathbf{x}_i, \boldsymbol{\beta})$ for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\beta}$, and $N^{-1}\sum_{i\in U} k(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.

(2.4.2.iv) The Horvitz-Thompson estimators of certain population means are asymptotically normally distributed.

(2.4.2.v) $\lambda_n \big/ (\sqrt{n}/(\sqrt{n})^{\gamma}) \to \infty$ and $\lambda_n \big/ \sqrt{n} \to 0$.

Remark II.1. The certain means in condition (2.4.2.iv) are means of first and second derivatives of $\mu(\mathbf{x}_i, \mathbf{t})$ in the Taylor series expansion of $\mu(\mathbf{x}_i, \mathbf{t})$ evaluated at a neighborhood around $\mathbf{B}$, which is a vector of values if $\mathbf{B}$ has more than one parameter. The condition requires that the Horvitz-Thompson estimates of the means are bounded element-wise.

Lemma II.2. *Assume the superpopulation model:*

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

Let $\mathbf{B}$ *be the finite-population quasilikelihood estimate of* $\boldsymbol{\beta}$*,* $\mathbf{B} \to \boldsymbol{\beta}$*. Under conditions (2.4.2.i)-(2.4.2.vi), the model-assisted asymptotic estimator of population total is:*

$$\hat{T}_y^{MC} = \sum_{i\in s_A} d_i^A (y_i - \mu_i B^{MC}) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{2.4.2.1}$$

where

$$\begin{aligned}\mu_i &= \mu(\mathbf{x}_i, \mathbf{B})\\ B^{MC} &= \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}\end{aligned}$$

The proof for Lemma II.2 is left in Appendix A.1.1. Given Lemma II.2, we can derive the asymptotic LASSO estimator of total: $\hat{T}_y^{LASSO}$. We later use $\hat{T}_y^{LASSO}$ to derive asymptotic expectation as well as asymptotic linearized variance estimates LASSO calibration estimator of total.

Theorem II.3. *Suppose the parameters in a full regression model have both zero and non-zero components, without loss of generality, let the first p be non-zero and the last q be zero:* $\boldsymbol{\beta}^F = \begin{pmatrix}\boldsymbol{\beta}^{(1)}_{(p\times 1)}\\ \boldsymbol{\beta}^{(2)}_{(q\times 1)}\end{pmatrix}$, $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)}$, *under conditions (2.4.2.i)-(2.4.2.v), the asymptotic LASSO calibration estimator of total is:*

$$\hat{T}_y^{LASSO} = \sum_{i\in s_A} d_i^A (y_i - \mu_i B^{MC}) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{2.4.2.2}$$

Proof. Under condition (2.4.2.v), the adaptive LASSO regression satisfies the oracle property through Theorems 1 and 4 in (*Zou*, 2006):

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

where $\mathbf{C} = \Sigma(\mathbf{B})$ is the covariance matrix of $\mathbf{B}$ under linear model, and $\mathbf{C} = I^{-1}(\mathbf{B})$ is the inverse of Fisher information matrix of $\mathbf{B}^{(1)}$ under generalized linear model. By Slutsky's theorem, the oracle property implies $\hat{\mathbf{B}}^{(1)} = \mathbf{B} + O_p(1/\sqrt{n})$. By condition

(2.4.2.i) and Lemma II.2:

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

□

Theorem II.4. $\hat{T}_y^{LASSO}$ *is asymptotically model-unbiased.*

Proof. Under the assumption of our theoretical framework, the superpopulation parameters are a subset of the full LASSO regression parameters, we can prove the asymptotic unbiasedness of $\hat{T}_y^{LASSO}$ by taking expectations with respect to model ξ. First note that:

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

Thus

$$\begin{aligned}E_\xi\left[\hat{T}_y^{LASSO} - T\right] &\approx E_\xi\left[\sum_{i\in s_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^N \mu_i B^{MC} - \sum_{i=1}^N y_i\right] \\ &= \sum_{i\in s_A} d_i^A(\mu_i - \mu_i) + \sum_{i=1}^N \mu_i - \sum_{i=1}^N \mu_i \quad \text{(since } E_\xi\left[B^{MC}\right] = 1\text{)} \\ &= 0\end{aligned}$$

□

Thus as long as LASSO regression parameters include the superpopulation parameters, $\hat{T}_y^{LASSO}$ is model-unbiased regardless of design weights. This property is essential in non-probability samples, where there are no initial design weights to guarantee unbiasedness.

2.4.3 Asymptotic design variance of $\hat{T}_y^{LASSO}$

To complete our theoretical development, we derive the linearized asymptotic variance estimate of $\hat{T}_y^{LASSO}$ by taking variance of the asymptotic LASSO estimator of total, $\hat{T}_y^{LASSO}$:

$$\begin{aligned} v_{\mathcal{A}}(\hat{T}_y^{LASSO}) &= V_{\mathcal{A}}\left(\sum_{i \in s_A} d_i^A \left(y_i - \mu_i B^{MC}\right) + \sum_{i=1}^{N} \mu_i B^{MC}\right) \\ &= V_{\mathcal{A}}\left(\sum_{i \in s_A} d_i^A \left(y_i - \mu_i B^{MC}\right)\right) \end{aligned} \tag{2.4.3.1}$$

(since $\mathcal{A}$ is single-stage probability-based sampling)

$$\begin{aligned} &= \sum_{i \in U} \left(\frac{y_i - \mu_i B^{MC}}{\pi_i}\right)^2 \pi_i(1-\pi_i) + \\ &\quad \sum_{i \in U}\sum_{j \neq i} (\pi_{ij} - \pi_i\pi_j) \frac{(y_i - \mu_i B^{MC})}{\pi_i}\frac{(y_j - \mu_j B^{MC})}{\pi_j} \end{aligned} \tag{2.4.3.2}$$

Equation (2.4.3.2) is consistent with equation (3.30) derived for the variance of traditional LASSO calibration estimator of total in *McConville* (2011). We use sample estimates for population quantities in (2.4.3.2):

$$\begin{aligned} v_{\mathcal{A}}(\hat{T}_y^{LASSO}) &= \sum_{i \in s_A} \left(\frac{y_i - \hat{\mu}_i \hat{B}^{MC}}{\pi_i}\right)^2 (1-\pi_i) + \\ &\quad \sum_{i \in s_A}\sum_{j \neq i} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{(y_i - \hat{\mu}_i \hat{B}^{MC})}{\pi_i}\frac{(y_j - \hat{\mu}_j \hat{B}^{MC})}{\pi_j} \end{aligned} \tag{2.4.3.3}$$

An alternative linearized variance estimate, suggested by (*Särndal et al.*, 1989), multiplies $(y_i - \hat{\mu}_i \hat{B}^{MC})$ by g-weights, which are the ratios of calibrated weights to the original design weights:

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

$$v.g_{\mathcal{A}}(\hat{T}_y^{LASSO}) = \sum_{i \in s_A} \left(\frac{g_i \left(y_i - \hat{\mu}_i \hat{B}^{MC} \right)}{\pi_i} \right)^2 (1 - \pi_i) + \sum_{i \in s_A} \sum_{j \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{g_i(y_i - \hat{\mu}_i \hat{B}^{MC})}{\pi_i} \frac{g_j(y_j - \hat{\mu}_j \hat{B}^{MC})}{\pi_j} \qquad (2.4.3.4)$$

To simplify notations, we refer to $v_{\mathcal{A}}(\hat{T}_y^{LASSO})$ as v^{LASSO} and $v.g_{\mathcal{A}}(\hat{T}_y^{LASSO})$ as v_g^{LASSO}.

2.5 Simulation setup

We design a simulation to evaluate LASSO calibration estimator of total: $\hat{T}_y^{LASSO}$, developed in Section 2.4.1, and linearized variance estimates of $\hat{T}_y^{LASSO}$: v^{LASSO} and v_g^{LASSO}, developed in section 2.4.3. Since both linearized variance estimates are based on asymptotic LASSO calibration estimate of the total, they might not perform well for small sample sizes. We also apply naive bootstrap variance estimates to obtain v_{boot}^{LASSO} by drawing 500 samples with replacement from each simulation sample to obtain a bootstrap variance estimate of $\hat{T}_y^{LASSO}$.

To simulate non-probability samples, we generate samples with unequal selection probabilities, but set design weights $\mathbf{d}^A = N/n$. The simulation is based on an artificial population, for which we know the true population regression coefficients. The calibration estimator that incorporates true regression coefficients (as opposed to coefficient estimates based on samples) is denoted by $\hat{T}_y^{ORACLE}$. We denote $\hat{T}_y^{GREG}$ as the traditional calibration estimator of total (see equation 2.2.1.4). The goal of the simulation is to compare bias, variance, and root-mean-square-error (RMSE) of $\hat{T}_y^{ORACLE}$, $\hat{T}_y^{LASSO}$, $\hat{T}_y^{GREG}$, and $\hat{T}_y^{HT}$ (pure design-based Horvitz-Thompson estimator) under different experimental designs. We evaluate v^{LASSO}, v_g^{LASSO}, and v_{boot}^{LASSO} through their coverage rates. Because $\hat{T}_y^{LASSO}$ performs both variable selection and estimation, we implement backward stepwise selection to select the working model for GREG. Although there is no theoretical justification for using stepwise variable selection, citeskinner1997variable has shown that given two auxiliary variables, a stepwise procedure can result in improved efficiency of GREG estimator. We are interested in knowing the performance of each estimator under (1) populations with different signal-to-noise-ratios (SNR), (2) independent, informative, and biased sampling schemes, and (3) small and large sample sizes. The signal-to-noise ratio is calculated according to definitions in (*Czanner et al.*, 2008). Let $\mathbf{X}^{(1)}$ be the covariate matrix corresponding to non-zero regression parameters $\boldsymbol{\beta}^{(1)}$, and $\mathbf{X}^{(2)}$ be the covariate matrix corresponding to zero-valued regression parameters $\boldsymbol{\beta}^{(2)}$, $\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{bmatrix}$. For a continuous outcome,

$$SNR = \frac{SSResidual(\mathbf{y}, \mathbf{X}^{(2)}, \mathbf{B}^{(2)}) - SSResidual(\mathbf{y}, \mathbf{X}, \mathbf{B})}{SSResidual(\mathbf{y}, \mathbf{X}, \mathbf{B})}$$

where *SSResidual* is the sum of squares of residuals of the linear regression. For a binary outcome,

$$SNR = \frac{Dev(\mathbf{y}, \mathbf{X}^{(2)}, \mathbf{B}^{(2)}) - Dev(\mathbf{y}, \mathbf{X}, \mathbf{B})}{Dev(\mathbf{y}, \mathbf{X}, \mathbf{B})}$$

where *Dev* is the deviance of the logistic regression model. We set two levels of correlations (low/high) between covariates, and two levels of effect sizes (low/high) of the covariates, resulting in 4 populations: low/low, low/high, high/low, and high-/high correlations. We set the low/high and high/low populations to have the same SNR in order to understand the influence of correlation and effect size on estimator's performance given the same SNR. Three sampling schemes are used to draw samples: simple-random-sampling without replacement, SRS, Poisson sampling with selection probabilities proportional to covariates, POI(X), Poisson sampling with selection probabilities proportional to covariates and the outcome, POI(X+Y). POI(X+Y) sampling simulates self-selection bias of non-probability samples, where the propensity of a respondent to participate in a study relates to the analysis variable. We consider two sample sizes: 250 and 1000. Thus we have a total of 24 experimental groups for a continuous outcome variable, and 24 experimental groups for a binary outcome variable.

2.5.1 Population

Population outcome and covariates. We follow a common type of population covariance structure used in LASSO-related simulations (*Tibshirani*, 1996; *Wang and Leng*, 2008; *Meier et al.*, 2008) – auto-decay correlation structure:

$$cor(X_i, X_j) = \rho^{|i-j|}, \quad \Sigma^\rho = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^p \\ \rho & 1 & \rho & \cdots & \rho^{p-1} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^p & \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{bmatrix}$$

We generate population covariate matrix with $N = 100,000$ from a multivariate normal distribution with mean $\mathbf{0}_{(p\times 1)}$ and covariance Σ^ρ, $p = 40$. The continuous outcome variable is generated by the regression model:

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

The binary outcome variable is generated by the logistic regression model:

$$\phi_i = expit(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{40} x_{i40}), \qquad expit(u) = (1 + exp(u))^{-1}$$
$$y_i = bernoulli(\phi_i)$$

Signal-to-Noise-Ratio parameters. We set $\rho = 0.15$ for low correlation population, and $\rho = 0.73$ for high correlation population. For both continuous and binary outcome variables:

$$\text{Low effect-size } \boldsymbol{\beta}^{(1)} := \beta_{12} \cdots \beta_{19}, \beta_{32} \cdots \beta_{39} = 0.45$$

$$\text{High effect-size } \boldsymbol{\beta}^{(1)} := \beta_{12} \cdots \beta_{19}, \beta_{32} \cdots \beta_{39} = 0.74$$

For continuous $\mathbf{y}$: $\beta_0 = 1$, for binary $\mathbf{y}$: $\beta_0 = 0.4$. The rest of $\beta_i = 0$. Out of 41 regression parameters, 16 are non-zero and 25 are zero.

2.5.2 Sampling schemes

Three sampling schemes are used to generate the sample:

(1) Simple-Random-Sampling (SRS): selection probabilities $= n/N$.

(2) Poisson sampling with probabilities proportional to $\mathbf{X}$, $POI(X)$.

$$\begin{cases} \text{continuous } \mathbf{y}: & \pi_i \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} \\ \text{binary } \mathbf{y}: & logit(\pi_i) = 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} \end{cases}$$

(3) Poisson sampling with probabilities proportional to $\mathbf{X}$ and $\mathbf{y}$, $POI(X + Y)$.

$$\begin{cases} \text{continuous } \mathbf{y}: & \pi_i \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} + 0.5y_i \\ \text{binary } \mathbf{y}: & logit(\pi_i) \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} + y_i \end{cases}$$

Recall that $\mathbf{X}_{15}$ and $\mathbf{X}_{35}$ have non-zero regression coefficients in generating $\mathbf{y}$. Table 2.1 summarizes the population quantities.

Table 2.1: Simulation parameters

Population size $N = 100,000$		SNR		T	
ρ	effect	binary y	continuous y	binary y	continuous y
low ($\rho = 0.15$)	low ($\beta^{(1)} = 0.4$)	0.51	0.47	56,175	100,707
low ($\rho = 0.15$)	high ($\beta^{(1)} = 0.74$)	1.10	1.26	54,472	101,369
high ($\rho = 0.73$)	low ($\beta^{(1)} = 0.4$)	1.10	1.26	54,184	101,772
high ($\rho = 0.73$)	high ($\beta^{(1)} = 0.74$)	2.75	3.41	52,782	103,120

Sampling scheme	outcome	selection probability	design weights
SRS	binary y	$\pi_i = n/N$	N/n
	continuous y	$\pi_i = n/N$	N/n
POI(X)	binary y	$logit(\pi_i) \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35}$	N/n
	continuous y	$\pi_i \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35}$	N/n
POI(X+Y)	binary y	$logit(\pi_i) \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} + y_i$	N/n
	continuous y	$\pi_i \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} + 0.5y_i$	N/n

	small	large
Sample size n	250	1,000

2.5.3 Evaluation metrics

Point estimates and variance. We evaluate empirical bias, variance, and RMSE for each estimator of total. Let S be the number of simulation iterations. We define:

$$\hat{\theta} = \hat{T}_y^{HT}, \quad \hat{T}_y^{GREG}, \quad \hat{T}_y^{LASSO}, \quad \hat{T}_y^{ORACLE}$$

$$\begin{aligned}
bias\left(\hat{\theta}\right) &= \frac{1}{S}\sum_{j=1}^{S}\left(\hat{\theta}_j - \theta\right) \\
var\left(\hat{\theta}\right) &= \frac{1}{S-1}\sum_{j=1}^{S}\left(\hat{\theta}_j - \bar{\hat{\theta}}_j\right)^2, \bar{\hat{\theta}} = \frac{1}{S}\sum_{j=1}^{S}\hat{\theta}_j \\
rmse\left(\hat{\theta}\right) &= \sqrt{bias^2(\theta_j) + var(\hat{\theta}_j)} \\
\theta &= \sum_{k=1}^{N} y_k
\end{aligned}$$

Relative performance. The most relevant experimental groups to non-probability samples are based on POI(X) and POI(X+Y). In these sampling schemes, the samples are off-balanced because the selection favors cases with higher covariate values. In the case of POI(X+Y), the analysis variable is part of the selection to mimic self-selection of non-probability sample members. We compare traditional calibration estimator, GREG, with LASSO calibration estimator under POI(X) and POI(X+Y) in terms of bias ratio (BR) and percent-relative-rmse (relrmse):

$$BR = \frac{bias\left(\hat{\theta}\right)}{\sqrt{var\left(\hat{\theta}\right)}}, \quad \%relrmse = 100\frac{rmse(\hat{T}_y^{LASSO})}{rmse(\hat{T}_y^{GREG})}$$

$$\hat{\theta} = \hat{T}_y^{GREG}, \hat{T}_y^{LASSO}$$

A bias ratio with value $|BR| > 0.4$ indicates potential problem in coverage of confidence intervals.

Variance estimates. We evaluate the linearized variance estimates and bootstrap variance estimates by the their 95% nominal coverage and %bias relative to empirical variance. We use normal approximation to generate confidence intervals. Let s index simulation number. We construct a confidence interval for each method, $v = v^{LASSO}, v_g^{LASSO}, v_{boot}^{LASSO}$:

$$CI_s = \hat{T}_{ys}^{LASSO} \pm 1.96\sqrt{v_s}$$

$$I_s = \begin{cases} 1, & \text{if } T \in CI_s \\ 0, & \text{otherwise} \end{cases}$$

where I_s is the indicator for whether the confidence interval covers the true population total. The nominal 95% coverage is: $100 \sum_{s=1}^{S} I_s/S$. We calculate %bias as:

$$\%bias = 100 \left[v - var\left(\hat{T}_y^{LASSO}\right)\right] / var\left(\hat{T}_y^{LASSO}\right)$$

where $var\left(\hat{T}_y^{LASSO}\right)$ is the empirical variance obtained from the simulation samples.

2.6 Simulation results

The simulation results are based on $S = 1,000$ simulated samples per each experimental group. Table 2.2 lists the numerical results of bias, variance, and root-mean-square-error of each estimator under different experimental designs for estimating the total of a continuous outcome variable. Table 2.3 lists the numerical results for estimating the total of a binary outcome variable. For bias and RMSE, we make the following systematic comparisons:

(1) LASSO relative to ORACLE: to see under what situations does LASSO approach the optimal performance, i.e., attains oracle property.

(2) Pure design-based HT estimator relative calibration-based estimators LASSO,

GREG: to evaluate the effectiveness of models in reducing bias and RMSE.

(3) LASSO relative to GREG under experimental groups that mimic non-probability samples: to compare the results under POI(X) and POI(X+Y), where we rely on working models to compensate the unavailability of initial design weights.

We follow the evaluations of bias, variance, and RMSE with comparisons of different variance estimates of LASSO calibration estimator of total. In section 2.6.4, we compare two asymptotic linearized variance estimates, v^{LASSO} and v_g^{LASSO} along with naive-bootstrap variance estimate, v_{boot}^{LASSO} in terms of coverage and bias.

2.6.1 LASSO relative to ORACLE

LASSO RMSE is closer to ORACLE RMSE as sample size increases. The gap between LASSO and ORACLE is also smaller as signal-to-noise-ratio grows; this is especially evident for estimating binary outcomes. It seems LASSO attains oracle property sooner when estimating continuous outcomes. The performance loss of LASSO relative to ORACLE is largely due to variance rather than bias is expected, since LASSO has to perform model selection per sample while the ORACLE estimator does not. Under the same SNR with sample size 250, LASSO is closer to ORACLE when the covariates exhibit high correlations. It seems that variable effect sizes have less impact on LASSO performance when the sample size is small. When the sample size is 1,000, for both binary and continuous outcomes, LASSO and ORACLE have very similar RMSEs. There is evidence that LASSO has oracle property as the sample size grows.

2.6.2 HT relative to calibration-based estimators

Under SRS, the initial design weights are correct. This is the only type of experimental group where HT produces roughly unbiased results. However, the calibration-based estimators still outperform HT for both continuous and binary outcomes and all sample sizes. There is strong evidence that assisting models can reduce RMSE over pure-design-based estimators. When the assisting-model is miss-specified, such as GREG estimator for binary outcomes, we still observe gains in RMSE over HT. Relative to GREG, LASSO's gain over HT is more substantial when estimating totals of binary variables. This can be largely attributed to LASSO employing a logistic-type model for binary outcome variables rather than a linear model. When the sampling is informative or biased (POI(X) and POI(X+Y)), HT performs poorly because the design weights are incorrect. Under high correlation and high effect sizes, HT bias is as large as the true population totals for continuous variables. The results are in-line with the theoretical development in Section 2.4.2: when the correct model is a subset of the full regression model, model-assisted estimator of total can still be approximately unbiased without the correct design weights. One interesting pattern is the bias of continuous outcome under POI(X+Y) sampling – for both GREG and LASSO, the bias is smaller for smaller sample sizes. This is likely due to additive bias under continuous outcome when the assisting model has not completely removed the sample bias. For estimating binary outcome totals, LASSO is effective in reducing the bias to roughly 1% under POI(X+Y).

2.6.3 LASSO relative to GREG under POI(X) and POI(X+Y)

Evaluating LASSO estimator without correct design weights under POI(X) and POI(X+Y) sampling schemes is at the heart of this research. POI(X) and POI(X+Y) induce biased samples by selecting cases with larger covariate values with higher probabilities. Under POI(X+Y), the selection also favors cases with larger outcome values. Tables 2.4 and 2.5 list bias ratio of GREG and LASSO and percent-relative-RMSE of LASSO to GREG. Except for continuous outcome under POI(X) sampling, both GREG and LASSO absolute bias ratios tend to be larger under sample size 1,000 than 250, suggesting persistent bias remains in estimating population totals in both sampling schemes. Under POI(X) sampling, GREG and LASSO have absolute bias ratios under 0.4 for both continuous and binary outcomes. Thus the bias may not be significant enough to cause issues in coverage under informative sampling. Under POI(X+Y) sampling, absolute bias ratios of GREG and LASSO are under or close to 0.4 for the continuous outcome, but consistently greater than 0.4 for the binary outcome. The absolute bias ratio of LASSO decreases as SNR increases, whereas the bias ratio of GREG shows only slight improvement as SNR increases. There is evidence that the coverage of LASSO estimator of total for a binary outcome improves as either correlation or effect size of covariates increases. The sample bias likely remains significant with GREG in estimating totals of binary outcomes. In terms of RMSE, LASSO has marginal improvement over GREG for estimating totals of continuous outcome variables. The improvement is slightly noticeable, about 3%, when there are highly correlated predictors in the model. Under POI(X+Y) sampling, LASSO calibration shows distinct advantage for both bias and RMSE over GREG. The advantage grows as SNR increases. Under Low/High and High/Low population types, the SNR is the same, thus the difference in performance between LASSO and GREG is attributed to correlation or effect size. LASSO performs better in both bias and RMSE in High/Low population type, suggesting that LASSO has stronger advantage over GREG when there are highly correlated predictors in the model. This suggests that LASSO has better variable selection capability in the presence of multicollinearity relative to stepwise variable selection procedure used in GREG. With high levels of correlation, LASSO improves over GREG by more than 22% in terms of RMSE while reducing bias by roughly 50% or more. When SNR increases, the relationship between $\mathbf{y}$ and $\mathbf{X}$ is stronger. One would expect that LASSO advantage over GREG, in terms of using the correctly specified working model, is less evident as SNR increases. We observe the opposite in the simulation. LASSO, under higher SNR has improved RMSE over GREG than under lower SNR. When stepwise is able to select better working models under higher SNR for GREG, LASSO is able to select more accurate models.

Figures 2.1 and 2.2 provide visual comparisons of LASSO and GREG for all experimental groups. For SRS and POI(X), both LASSO and GREG look approximately unbiased. The performances of LASSO and GREG are nearly the same when estimating continuous outcome totals. When estimating binary outcome totals, under POI(X+Y), LASSO approaches the true value as signal-to-noise-ratio increases, while GREG remains distant from the true value. There is evidence that LASSO can achieve better bias reduction than traditional calibration under the situations: (1) some predictors are highly correlated, (2) samples are prone to self-selection bias, (3) the inference is on population totals of binary outcomes.

2.6.4 Variance estimates

Tables 2.6 and 2.7 list the 95% nominal coverage and percent-bias for each of the two linearized asymptotic variance estimators developed in this research, as well as the naive bootstrap variance estimate of the LASSO calibration estimator. We explore bootstrap variance estimates because the rate of convergence to asymptotic variance may be slow.

For continuous outcome, bootstrap variances have coverages that are consistently close to 95% under SRS and POI(X) sampling schemes for both sample sizes. Under POI(X+Y) sampling scheme, the coverages are lower, 92%-93%. The linearized variances have coverages that are sensitive to both sample size and sampling scheme. With sample size 250, the coverages are consistently 91%-92% under SRS and POI(X), and 90%-91% under POI(X+Y). With sample size 1,000, the linearized variances have improved coverages: 93%-94% under SRS and POI(X), 90%-92% under POI(X+Y). The difference in coverage of linearized variance estimates between small and large sample sizes is expected, since the variance estimates are asymptotic and improves

Table 2.2: Simulation summary for continuous outcome

		sampling	**HT**			**GREG**			**LASSO**			**ORACLE**		
Population	**n**	**scheme**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**
		SRS	532	546,381,441	23,381	861	424,526,197	20,622	934	427,591,058	20,699	846	358,924,754	18,964
	250	POI(X)	12,382	524,741,577	26,039	-580	445,787,070	21,122	-356	441,273,103	21,010	-989	356,411,600	18,905
low/low		POI(X+Y)	19,397	518,926,811	29,920	4,596	443,164,082	21,547	4,713	430,951,163	21,288	3,910	353,111,827	19,194
T = 100,707		SRS	156	129,228,386	11,369	273	93,805,883	9,689	297	93,504,572	9,674	351	90,805,251	9,536
SNR = 0.47	1000	POI(X)	12,634	128,995,875	16,989	-131	91,169,298	9,549	-162	91,777,156	9,581	-425	88,288,602	9,406
		POI(X+Y)	19,728	128,307,969	22,749	4,912	91,348,523	10,746	4,957	90,968,340	10,749	4,473	88,219,229	10,403
		SRS	352	849,471,957	29,148	856	414,610,415	20,380	954	416,676,767	20,435	846	358,924,754	18,964
	250	POI(X)	21,070	818,411,393	35,529	-1,265	433,794,615	20,866	-917	432,230,564	20,810	-989	356,411,600	18,905
low/high		POI(X+Y)	31,709	817,283,474	42,694	3,749	427,131,827	21,004	4,023	426,946,783	21,051	3,871	354,097,194	19,212
T = 101,369		SRS	30	199,722,626	14,132	272	93,872,408	9,693	297	93,058,990	9,651	351	90,805,251	9,536
SNR = 1.26	1000	POI(X)	21,059	199,443,370	25,356	-133	91,201,517	9,551	-211	90,416,237	9,511	-425	88,288,602	9,406
		POI(X+Y)	31,684	196,263,286	34,643	4,865	90,576,516	10,688	4,825	89,406,049	10,615	4,435	87,756,620	10,364
		SRS	62	940,500,518	30,668	960	421,417,670	20,551	1,029	399,385,645	20,011	853	359,046,433	18,968
	250	POI(X)	50,237	894,823,134	58,469	-681	433,556,452	20,833	-1,634	402,421,619	20,127	-1,118	363,311,654	19,093
high/low		POI(X+Y)	57,774	872,312,893	64,886	4,052	435,016,719	21,247	3,002	399,327,736	20,207	3,510	361,871,205	19,344
T = 101,772		SRS	22	218,139,275	14,770	284	93,674,279	9,683	293	92,973,285	9,647	349	90,805,664	9,536
SNR = 1.26	1000	POI(X)	50,594	209,736,746	52,626	-102	93,432,544	9,667	-539	91,379,357	9,574	-489	89,854,380	9,492
		POI(X+Y)	58,167	209,105,092	59,937	4,670	95,183,110	10,816	4,183	92,237,356	10,475	4,164	90,079,904	10,364
		SRS	-421	1,896,806,029	43,554	849	435,725,344	20,891	1,020	406,548,977	20,189	853	359,046,433	18,968
	250	POI(X)	83,393	1,826,143,138	93,704	-765	434,899,932	20,868	-1,541	406,073,024	20,210	-1,118	363,311,654	19,093
high/high		POI(X+Y)	96,471	1,778,575,637	105,286	3,723	428,204,930	21,025	3,040	403,711,035	20,321	3,359	362,307,680	19,328
T = 103,120		SRS	-191	444,307,136	21,079	282	93,144,269	9,655	303	92,995,701	9,648	349	90,805,664	9,536
SNR = 3.41	1000	POI(X)	83,577	424,227,983	86,077	-235	93,110,279	9,652	-510	91,063,458	9,556	-489	89,854,380	9,492
		POI(X+Y)	96,880	422,928,909	99,039	4,372	94,084,518	10,639	4,084	91,842,347	10,417	4,019	89,819,107	10,294

Table 2.3: Simulation summary for binary outcome

		sampling	**HT**			**GREG**			**LASSO**			**ORACLE**		
Population	**n**	**scheme**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**
		SRS	7	10,166,643	3,189	10	7,151,995	2,674	45	7,041,549	2,654	77	6,104,936	2,472
	250	POI(X)	2,582	9,955,930	4,077	181	8,022,014	2,838	122	7,756,078	2,788	24	5,914,888	2,432
low/low		POI(X+Y)	4,856	9,820,367	5,780	1,980	8,134,486	3,472	1,810	7,840,279	3,334	1,529	6,036,561	2,894
T = 56,175		SRS	-6	2,678,458	1,637	13	1,740,499	1,319	11	1,621,735	1,274	24	1,562,244	1,250
SNR = 0.51	1000	POI(X)	2,493	2,415,532	2,938	2	1,773,634	1,332	-32	1,708,100	1,307	-39	1,513,083	1,231
		POI(X+Y)	4,741	2,315,032	4,979	1,767	1,770,841	2,212	1,643	1,704,963	2,098	1,489	1,555,240	1,942
		SRS	-20	10,841,939	3,293	21	6,074,082	2,465	87	5,385,316	2,322	72	4,399,422	2,099
	250	POI(X)	3,022	10,206,245	4,397	134	6,138,442	2,481	83	5,775,727	2,405	31	4,105,432	2,026
low/high		POI(X+Y)	5,297	9,825,517	6,155	1,556	6,215,811	2,939	1,319	5,822,126	2,750	1,081	4,165,828	2,310
T = 54,472		SRS	-11	2,651,591	1,628	33	1,287,543	1,135	34	1,048,220	1,024	28	1,017,130	1,009
SNR = 1.10	1000	POI(X)	2,944	2,371,990	3,323	9	1,420,417	1,192	-60	1,234,151	1,113	-19	1,085,200	1,042
		POI(X+Y)	5,184	2,292,359	5,400	1,420	1,407,244	1,850	1,100	1,224,615	1,560	1,049	1,124,686	1,492
		SRS	-23	10,297,721	3,209	38	5,769,837	2,402	98	4,915,369	2,219	96	4,085,485	2,024
	250	POI(X)	6,614	9,641,798	7,307	271	6,160,573	2,497	-217	4,801,426	2,202	-63	4,038,766	2,011
high/low		POI(X+Y)	8,618	9,327,405	9,143	1,750	6,280,660	3,057	940	4,936,623	2,413	1,020	4,241,811	2,298
T = 54,184		SRS	-24	2,498,438	1,581	14	1,231,678	1,110	24	997,263	999	21	950,476	975
SNR = 1.10	1000	POI(X)	6,557	2,194,907	6,722	180	1,373,329	1,186	-223	1,091,859	1,068	-123	1,054,915	1,034
		POI(X+Y)	8,536	2,128,354	8,660	1,630	1,357,195	2,003	898	1,072,196	1,371	957	1,073,504	1,410
		SRS	-121	10,171,546	3,192	-21	5,209,275	2,282	59	3,774,442	1,944	21	2,728,195	1,652
	250	POI(X)	7,139	9,824,255	7,797	288	5,708,252	2,407	-243	3,649,966	1,926	4	2,740,594	1,655
high/high		POI(X+Y)	9,114	9,379,607	9,615	1,516	5,716,313	2,831	543	3,707,496	2,001	692	2,841,778	1,822
T = 52,782		SRS	-55	2,540,285	1,595	-16	1,073,393	1,036	3	647,693	805	0	604,681	778
SNR = 2.75	1000	POI(X)	7,090	2,191,534	7,243	188	1,228,264	1,124	-213	745,018	889	-30	674,859	822
		POI(X+Y)	9,055	2,108,114	9,170	1,440	1,205,997	1,811	532	727,537	1,005	672	684,669	1,066

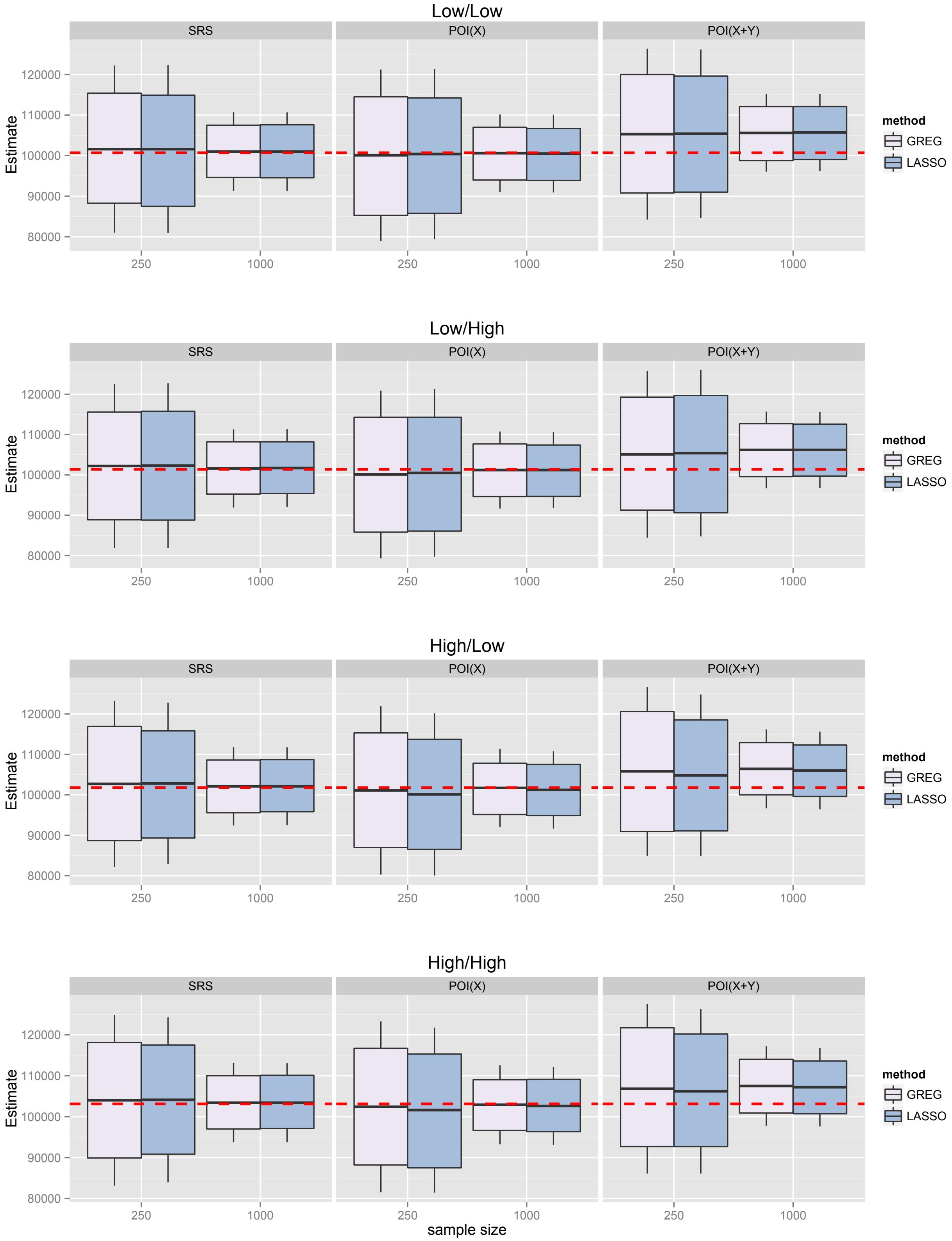
Table 2.4: Bias ratio of LASSO and GREG under POI(X) and POI(X+Y)

		POI(X)				**POI(X+Y)**			
		Continuous		**Binary**		**Continuous**		**Binary**	
Population	**n**	**GREG**	**LASSO**	**GREG**	**LASSO**	**GREG**	**LASSO**	**GREG**	**LASSO**
low/low	250	-0.03	-0.02	0.06	0.04	0.22	0.23	0.69	0.65
	1000	-0.01	-0.02	0.00	-0.02	0.51	0.52	1.33	1.26
low/high	250	-0.06	-0.04	0.05	0.03	0.18	0.19	0.62	0.55
	1000	-0.01	-0.02	0.01	-0.05	0.51	0.51	1.20	0.99
high/low	250	-0.03	-0.08	0.11	-0.10	0.19	0.15	0.70	0.42
	1000	-0.01	-0.06	0.15	-0.21	0.48	0.44	1.40	0.87
high/high	250	-0.04	-0.08	0.12	-0.13	0.18	0.15	0.63	0.28
	1000	-0.02	-0.05	0.17	-0.25	0.45	0.43	1.31	0.62

Table 2.5: Relative RMSE of LASSO to GREG under POI(X) and POI(X+Y)

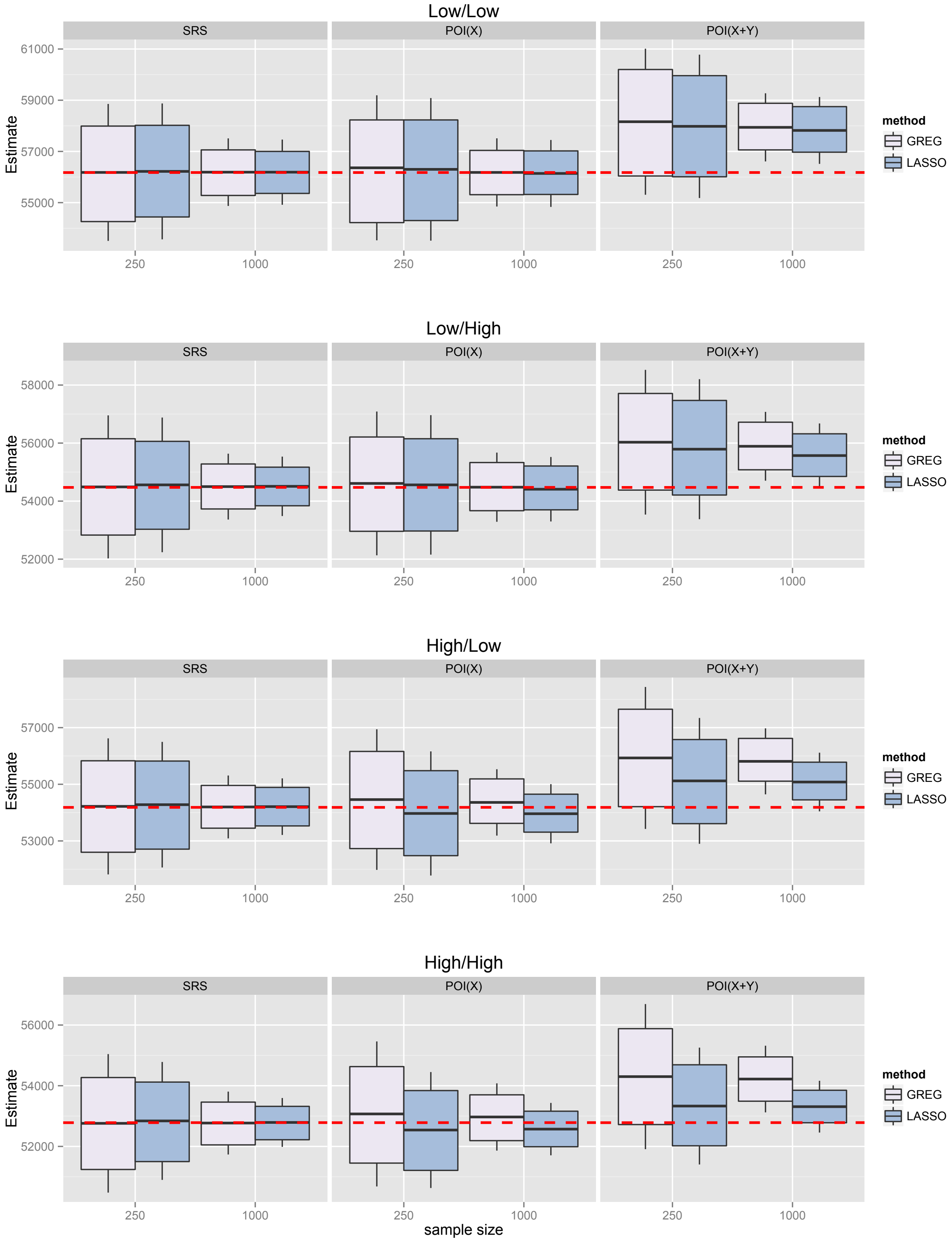
		POI(X)		**POI(X+Y)**	
		Continuous	**Binary**	**Continuous**	**Binary**
Population	**n**	**%relrmse**	**%relrmse**	**%relrmse**	**%relrmse**
low/low	250	99.5%	98.2%	98.8%	96.0%
	1000	100.3%	98.2%	100.0%	94.9%
low/high	250	99.7%	96.9%	100.2%	93.6%
	1000	99.6%	93.3%	99.3%	84.3%
high/low	250	96.6%	88.2%	95.1%	78.9%
	1000	99.0%	90.1%	96.8%	68.4%
high/high	250	96.8%	80.0%	96.7%	70.7%
	1000	99.0%	79.1%	97.9%	55.5%

Figure 2.1: Boxplot continuous outcome



Box = 25%-75% quantile, horizontal line in box = mean, vertical line extending from box = mean ± sd, dotted red line = True T

Figure 2.2: Boxplot binary outcome



Box = 25%-75% quantile, horizontal line in box = mean, vertical line extending from box = mean ± sd, dotted red line = True T over larger samples. There is slight improvement for linearized variance coverage as SNR increases under POI(X+Y). In terms of bias, there is evidence that bias improves as SNR increases. With the same SNR, both linearized and bootstrap variances have smaller bias given predictors with high correlations relative to predictors with high effect sizes. Linearized variances tend to underestimate the empirical variance, especially when the sample size is small. Overall, there is very little difference between the two linearized variance estimates. Bootstrap variance tends to overestimate the empirical variance, but the bias is generally smaller than linearized variance estimates'.

For binary outcome, both linearized and bootstrap variance estimates are sensitive to sample size, sampling scheme, and SNR. Bootstrap variance coverages are consistently close to 95% under SRS and POI(X) for both sample sizes and all population types, but coverages range from 75% to 94% under POI(X+Y). Under POI(X+Y), the bootstrap variance coverages are better with sample size 250 than with sample size 1,000, and better with high-correlation populations than with low-correlation populations. In terms of coverage, linearized variances show a similar trend under POI(X+Y) as bootstrap: better coverage with smaller samples than bigger samples, and better coverage with high-correlation populations than with low-correlation populations. It is likely that sample bias persists under POI(X+Y), especially for low-correlation populations, thus the coverage is low at larger sample sizes due to more biases being added together. Under SRS and POI(X), linearized variance coverage improves as sample size increases: from 85%-90% to 90%-94%. In terms of bias, both bootstrap and linearized variances have smaller bias with larger sample sizes. Under the same sample size, linearized variance estimates have larger bias as SNR increases. The same trend is not observed in bootstrap variance estimates. Similar to continuous outcome results, linearized variance tends to underestimate the empirical variance, especially when the sample size is small. Unlike continuous outcome results, there is evidence that the g-weighted linearized variance estimates have better bias-property than unweighted linearized variance estimate. Bootstrap variance tends to overestimate the empirical variance. However, the biases are much smaller than linearized variance estimates'.

Overall, bootstrap variances have better coverage than linearized variances, although in settings with large sample sizes and low correlation between covariates, nominal coverage is still poor. Bootstrap variance estimates also have smaller bias, and they are almost always positive, which is more desirable than the negative bias of linearized variance estimates. Under continuous outcome, if the sample size is sufficiently large, linearized asymptotic variances may potentially be used. In this simulation, the linearized variance coverages under sample size 1,000 for continuous outcomes range from 90% to 95%, with bias generally around -5%. For all other situations: small sample size, binary outcome type, and all correlation/effect size combinations, we recommend bootstrap variance estimates over linearized variance estimates where possible.

Table 2.6: 95% nominal coverage and %bias of variance estimates for LASSO

Continuous outcome			**coverage**			**%bias**		
Population	**n**	**scheme**	v^{LASSO}	v_g^{LASSO}	v_{boot}^{LASSO}	v^{LASSO}	v_g^{LASSO}	v_{boot}^{LASSO}
low/low	250	SRS	91.7%	91.8%	95.4%	-22.6%	-22.3%	2.9%
		POI(X)	91.2%	91.2%	96.1%	-25.1%	-24.5%	5.7%
		POI(X+Y)	89.6%	89.9%	95.4%	-23.5%	-22.8%	7.9%
	1000	SRS	93.2%	93.2%	93.8%	-7.3%	-7.2%	-0.3%
		POI(X)	94.0%	93.9%	95.5%	-5.7%	-5.3%	6.6%
		POI(X+Y)	90.0%	90.1%	92.1%	-4.9%	-4.4%	7.9%
low/high	250	SRS	91.5%	91.5%	95.7%	-22.6%	-22.3%	6.2%
		POI(X)	90.9%	91.2%	96.4%	-25.4%	-24.9%	8.8%
		POI(X+Y)	90.0%	90.2%	95.1%	-24.5%	-23.7%	9.9%
	1000	SRS	93.4%	93.5%	94.3%	-6.6%	-6.5%	-0.1%
		POI(X)	94.1%	94.2%	95.9%	-4.0%	-3.5%	7.6%
		POI(X+Y)	90.7%	90.7%	92.7%	-2.9%	-2.3%	9.6%
high/low	250	SRS	92.3%	92.2%	95.4%	-17.4%	-17.1%	2.0%
		POI(X)	92.5%	92.6%	95.8%	-17.9%	-16.1%	6.4%
		POI(X+Y)	91.2%	91.8%	96.5%	-17.4%	-15.4%	7.1%
	1000	SRS	93.5%	93.5%	94.4%	-6.5%	-6.4%	-0.9%
		POI(X)	94.1%	94.0%	95.4%	-5.0%	-3.1%	5.7%
		POI(X+Y)	91.9%	92.3%	93.4%	-6.0%	-3.9%	5.0%
high/high	250	SRS	92.3%	92.3%	95.2%	-19.6%	-19.3%	2.2%
		POI(X)	92.0%	92.3%	96.1%	-19.6%	-17.8%	7.4%
		POI(X+Y)	91.2%	91.8%	95.6%	-19.1%	-16.9%	8.3%
	1000	SRS	93.4%	93.4%	94.5%	-6.5%	-6.4%	-0.7%
		POI(X)	94.0%	94.5%	95.6%	-4.7%	-2.8%	6.7%
		POI(X+Y)	92.2%	92.4%	93.4%	-5.6%	-3.3%	6.1%

Table 2.7: 95% nominal coverage and %bias of variance estimates for LASSO

Binary outcome				**coverage**			**%bias**	
Population	**n**	**scheme**	v^{LASSO}	v_g^{LASSO}	v_{boot}^{LASSO}	v^{LASSO}	v_g^{LASSO}	v_{boot}^{LASSO}
low/low	250	SRS	89.8%	90.0%	95.9%	-28.1%	-27.8%	9.2%
		POI(X)	88.1%	88.6%	96.7%	-37.3%	-35.3%	9.2%
		POI(X+Y)	79.0%	79.9%	91.2%	-38.7%	-35.9%	8.0%
	1000	SRS	92.8%	92.8%	93.5%	-11.9%	-11.8%	-3.5%
		POI(X)	92.0%	92.8%	95.7%	-17.9%	-15.5%	1.0%
		POI(X+Y)	68.6%	69.6%	74.6%	-18.5%	-14.9%	0.5%
low/high	250	SRS	86.8%	87.0%	94.9%	-37.7%	-37.3%	11.3%
		POI(X)	85.4%	86.1%	95.5%	-42.9%	-41.2%	14.4%
		POI(X+Y)	78.7%	80.1%	92.6%	-44.0%	-41.3%	14.4%
	1000	SRS	94.4%	94.3%	95.2%	-5.5%	-5.4%	5.8%
		POI(X)	91.8%	92.1%	94.9%	-20.5%	-18.6%	-1.8%
		POI(X+Y)	76.8%	77.8%	82.9%	-20.4%	-16.9%	-1.3%
high/low	250	SRS	89.2%	89.1%	94.4%	-28.5%	-28.1%	0.4%
		POI(X)	89.0%	90.1%	95.5%	-31.9%	-25.3%	12.7%
		POI(X+Y)	85.7%	88.4%	93.8%	-33.9%	-25.4%	10.9%
	1000	SRS	93.9%	93.9%	95.6%	-6.3%	-6.2%	3.5%
		POI(X)	92.6%	93.4%	94.8%	-16.5%	-9.2%	1.9%
		POI(X+Y)	83.3%	85.4%	88.1%	-15.0%	-5.0%	5.2%
high/high	250	SRS	82.8%	82.8%	93.8%	-44.6%	-44.3%	-6.4%
		POI(X)	83.6%	85.5%	95.1%	-44.3%	-39.4%	3.8%
		POI(X+Y)	82.9%	85.1%	93.8%	-45.1%	-38.4%	4.6%
	1000	SRS	94.3%	94.4%	96.1%	-7.8%	-7.6%	6.3%
		POI(X)	91.3%	92.2%	94.0%	-20.0%	-13.8%	0.2%
		POI(X+Y)	86.3%	88.6%	91.5%	-18.1%	-9.2%	2.8%

2.7 Application to National Health Interview Survey (NHIS)

2.7.1 NHIS and ACS Data

We next apply LASSO calibration to National Health Interview Survey (NHIS) 2013 to estimate the total number of adults (age 18 or older) diagnosed with cancer in the population. National Health Interview Survey is a nationally representative sample of non-institutionalized civilian households collected by a multi-stage area-probability sampling. Each month, health-related data on people in selected households are obtained by face-to-face interviews. The data provides pseudo-primary-sampling-unit (PSU), pseudo-strata, and sampling weights to allow for weighted estimates with complex survey design. In addition to health-related measures, NHIS also collects family income data to supplement the core interviews.

To calibrate NHIS on a set of demographic and income-related variables, we use the American Community Survey (ACS) 2013 micro-data as the benchmark data. ACS samples are households selected through multi-stage area-probability sampling from 3,143 counties of the U.S. The design of ACS is to improve estimates of small areas between the decennial census long-form samples. Around three million households are selected each year, with measures on household types and individual demographics within the households. ACS also collects data from group-quarters, which are excluded from this analysis. For ACS 2013, the sample size for adults is 2,317,301. The NHIS 2013 sample size is 34,201 after removing 242 cases with missing values on demographic variables. The weighted total number of persons is 234,810,075 in the ACS benchmark data, and 234,950,584 in the NHIS sample. The total weighted number of individuals between the two samples are very close. Thus the removal of 234 cases from the NHIS sample should have minimal impact on our analysis. As ACS sample is significantly larger, we treat weighted estimates of calibration variables from ACS as known population totals.

2.7.2 Estimators

The outcome variable of interest is whether a person has been diagnosed with cancer. Define the binary indicator for the outcome variable:

$$y_i = \begin{cases} 1 : & \text{if person } i \text{ has been diagnosed with cancer} \\ 0 : & \text{otherwise} \end{cases}$$

We first use the NHIS 2013 sampling weights, $\mathbf{w}^{NHIS}$, and design variables to obtain an unbiased estimate of the population total, $T_y = \sum_{i=1}^{N} y_i$. Then we assume that the NHIS 2013 sample is collected from a simple-random-sampling, with initial design weights $\mathbf{d}^A = N/n$, where N is the population total obtained from ACS, and n is the sample size of NHIS. We calibrate $\mathbf{d}^A$ by a set of demographic and income variables with traditional calibration and LASSO calibration. Thus we generate four estimates:

(1) $\hat{T}_y^{NHIS} = \sum_{i \in s_A} w_i^{NHIS} y_i$: Estimate obtained with NHIS weights.

(2) $\hat{T}_y^{HTSRS} = \sum_{i \in s_A} (N/n) y_i$: Estimate obtained with weights $\mathbf{d}^A = N/n$.

(3) $\hat{T}_y^{GREG} = \sum_{i \in s_A} w_i^{GREG} y_i$: Estimate obtained by calibrating $\mathbf{d}^A$ with GREG and backward stepwise variable selection.

(4) $\hat{T}_y^{LASSO} = \sum_{i \in s_A} w_i^{LASSO} y_i$: Estimate obtained by calibrating $\mathbf{d}^A$ with LASSO.

The variance of $\hat{T}_y^{NHIS}$ is the linearized variance estimate of total, accounting for sampling-stratum, primary-sampling-units, and survey weights in the NHIS 2013 sample. Variances of HTSRS and GREG are linearized variance estimates with weights $\mathbf{d}^A$ and $\mathbf{w}^{GREG}$ respectively. We obtain the variance of LASSO estimator through naive bootstrap as in the simulation study of section 2.5.

2.7.3 Working models

Table 2.8 lists calibration variable names, labels, values, and distributions in this analysis. The first row is the unweighted distribution of variables in the NHIS sample. The second row contains variable distributions in the NHIS sample, weighted by $\mathbf{w}^{NHIS}$ person-level weights. The third row is the distribution of variables in the population obtained from the ACS benchmark data. Missing income category is included as a separate category to capture the difference in missing patterns between NHIS and ACS. Including a missing category also allows us to maintain the analytical sample size. Relative to ACS, the unweighted NHIS sample has higher proportions of females, widowed/divorced/separated, and fewer proportion of non-Hispanic whites. After weighting, the NHIS distributions of gender and race are close to the benchmark's, and only marital status categories show some differences. Overall, NHIS sample composition is relatively close to the population's. However, we can still have substantial bias in estimating the total number of adults diagnosed with cancer due to different distributions of variables not collected in both NHIS and ACS, such as smoking status and other health-risk factors.

We use an unweighted linear model with backward-stepwise variable selection to determine the working model for GREG. The final variables included in the model for GREG are:

$$Age, Education, Race, Employed, Family\ income$$

For LASSO calibration, we use all variables in Table 2.8 for the LASSO regression model. Table 2.9 lists the model parameter estimates. For GREG, the parameters are obtained from a linear regression model, $E[y_i] = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, while LASSO parameters estimates are obtained from the logistic LASSO regression as described in section 2.3:

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

Because the number of parameters is not large, LASSO sets only one category to 0 – region[2]. For variables that are included in both models, besides the constant intercept, the parameter estimates agree in direction and show similar trend in effect sizes. For example, the risk for cancer increases as age increases. Age is the strongest predictor for risk of cancer in both models. We anticipate that GREG and LASSO calibrations to have similar impact on reducing sample bias when estimating the total number of individuals diagnosed with cancer.

2.7.4 Results

Table 2.10 lists the estimates, standard errors (SE), and percent-deviate from the NHIS estimate: $\%deviate = 100(\hat{T} - \hat{T}_y^{NHIS})/\hat{T}_y^{NHIS}$. We treat NHIS estimate as the unbiased estimate because it is calculated with probability-based sampling weights provided by NHIS. Without any weighting adjustment, HTSRS shows a significant positive bias. GREG estimator reduces the bias from 5.94% to 1.97%, while LASSO estimator reduces the bias to 0.88%. For population totals, 5.94% can be a substantial error. In this analysis, if NHIS were a non-probability sample, without weighting adjustment, we would have over-counted the number of adults with cancer by 1.18 million. With traditional calibration, the error is reduced to an over-count of 392,276. LASSO calibration further reduces the over-count to 175,344.

As expected, the standard error of the NHIS estimate is the largest, as it properly incorporates complex survey design. If the calibration working model correctly captures the relationship between the outcome variable and the calibration variables, we anticipate that the calibration estimator standard errors to be smaller than HTSRS

Table 2.8: Calibration variables

		Region				**Age)**							**Gender**		**Education**					**Race**				**Marital status**			**Employed**		**Family income**				
		(region)				(agegrp)							(gender)		(educ2)					(race)				(marst)			(employed)		(faminc_q)				
		Northeast	Midwest	South	West	18-29	30-39	40-49	50-59	60-69	70-79	80+	Male	Female	Less than high school	High school or less	Some college	College graduate	Post-graduate	non-Hispanic white	non-Hispanic black	Hispanic	Other	Married/partnered	Widowed/divorced/separated	Never married	No	Yes	1st quartile	2nd quartile	3rd quartile	4th quartile	missing
sample	**weights**	1	2	3	4	1	2	3	4	5	6	7	1	2	0	1	2	3	4	1	2	3	4	1	2	3	0	1	0	1	2	3	9
NHIS	none	16%	20%	37%	26%	19%	17%	16%	17%	15%	9%	6%	45%	55%	16%	26%	20%	29%	10%	60%	15%	17%	8%	49%	27%	24%	35%	65%	22%	20%	21%	21%	17%
NHIS	person-level	18%	23%	37%	23%	21%	17%	18%	18%	14%	8%	4%	48%	52%	14%	26%	20%	30%	10%	66%	12%	15%	7%	60%	18%	22%	33%	67%	15%	17%	22%	28%	19%
ACS	person-level	18%	21%	37%	23%	21%	17%	18%	18%	14%	8%	5%	52%	48%	13%	28%	23%	25%	10%	66%	12%	15%	7%	52%	20%	28%	39%	61%	19%	20%	20%	19%	22%

Table 2.9: Working model parameter estimates

	Dependent variable: cancer	
	GREG	LASSO
region[2]		0
region[3]		0.048
region[4]		−0.0780
agegrp[2]	0.010**	0.484
agegrp[3]	0.022***	0.985
agegrp[4]	0.077***	1.997
agegrp[5]	0.132***	2.460
agegrp[6]	0.198***	2.837
agegrp[7]	0.232***	2.984
gender[2]		0.004
educ2[1]	0.015***	0.121
educ2[2]	0.021***	0.210
educ2[3]	0.020***	0.211
educ2[4]	0.034***	0.357
race[2]	−0.042***	−0.595
race[3]	−0.043***	−0.811
race[6]	−0.051***	−0.771
marst[2]		−0.017
marst[3]		−0.093
employed[1]	−0.034***	−0.426
faminc_q[1]	0.001	−0.055
faminc_q[2]	0.005	0.004
faminc_q[3]	0.001	0.005
faminc_q[9]	−0.014***	−0.228
Constant	0.043***	−3.90
Observations	34,201	34,201
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2.10: Results for estimating total number of individuals with cancer

Estimator	$\hat{T}$	**SE**	**%deviate from NHIS**
NHIS	19,889,327	492,263	0.00%
HTSRS	21,070,498	362,883	5.94%
GREG	20,281,603	367,900	1.97%
LASSO	20,064,671	347,586	0.88%

%deviate is the difference to NHIS estimate divide by the NHIS estimate

estimator's. This is not the case for the GREG estimator, where the standard error is larger than HTSRS's. For LASSO calibration, we do observe a smaller standard error than HTSRS's, even with the bootstrap variance estimate that tends to overestimate. Without using the correct design weights, LASSO calibration is able to accurately produce an estimate of a population total while providing the smallest standard error among the estimators in this application.

2.8 Conclusion

In this chapter, we developed the LASSO calibration estimator of population totals, $\hat{T}_y^{LASSO}$, given population auxiliary data. We also derived the asymptotic linearized variance estimates for $\hat{T}_y^{LASSO}$. Simulation results show that the point estimates are approximately unbiased under simple-random sampling and informative sampling. For sample selections that are related to analysis variables, LASSO was able to significantly reduce sample bias even without the correct design weights. The performance of $\hat{T}_y^{LASSO}$ is near the optimal of $\hat{T}_y^{ORACLE}$ in terms of RMSE for both binary and continuous outcome variables when the sample size is large. LASSO tends to outperform stepwise-selected working models when there are high correlations among the choices of covariates. For analysis with many categorical variables, where there are natural correlations between the categories, LASSO calibration estimator can perform well over traditional calibration estimators, even if the effect sizes are small. The improvement is even more significant when the outcome of interest is binary. We have demonstrated theoretically and through simulations that LASSO calibration holds great promise in making unbiased inference of population totals from non-probability samples. Although asymptotic linearized variance estimates did not produce very accurate nominal coverage, naive bootstrap is a viable alternative approach. In an application to estimate population total of individuals diagnosed with cancer, without correct design weights, the LASSO calibration estimator was able to produce an estimate that is the closest to the estimate based on correct survey weights. LASSO calibration estimator also has the smallest standard error or all the estimators considered, even when using bootstrap variance estimate that tends to have positive bias. The application shows that LASSO calibration can generate inference to the population for a specific outcome variable, and the inference is both more accurate and precise than traditional calibration estimators.

CHAPTER III

Calibration with LASSO to Estimated Control

3.1 Introduction

Amidst declining response rates and rapidly increasing cost of probability-based sampling, the resurgence of more cost-effective non-probability sampling has prompted survey researchers to explore different adjustment methods for non-probability samples. The most recent development in post-survey adjustment of non-probability samples is propensity-score weighting (*Lee*, 2004; *Schonlau et al.*, 2004). Propensity-score weighting combines probability and non-probability samples to generate pseudo-selection-weights for non-probability sample respondents. The method requires a probability-sampling-based data, and all variables related to propensity of a respondent being in the non-probability sample. In practice, researchers who turn to non-probability-sampling for data collection are trying to avoid the cost to conduct probability-based-sampling. Usually only a small probability-based reference sample is obtained, which can result in highly variable propensity weights when all relevant variables are included in the propensity model (*Valliant and Dever*, 2011). Even if an appropriate probability-based reference sample is collected, there is no systematic way to determine the correct variables that can generate weights to fix all errors of a non-probability sample (*Schonlau et al.*, 2009). An alternative approach is to adjust the non-probability sample so that the weighted sample totals of a set of variables, known as calibration variables, equal to their Census benchmark totals. Although the method does not require specialized probability-sampling-based data, the set of calibration variables is small due to limited Census benchmark information. The resulting calibrated weights can only correct the imbalance with respect to the calibration variables, which is insufficient for adjusting all errors of a non-probability sample. The current approach to create one single set of survey weights to correct all imbalance within a non-probability sample is not practical.

This research focuses on making generalizable inference from non-probability samples by constructing an outcome-specific set of weights designed to reduce the bias and variance of a weighted total. We combine both approaches of propensity-score weighting and calibration by utilizing a probability-based benchmark sample similar to the probability-based reference sample used for propensity-score weighting. However, unlike the propensity-score weighting adjustment, the aim is not to include all possible variables in creating an one-size-fits-all weight to correct the imbalance within the non-probability sample. Instead, we 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. The approach falls under the model-assisted calibration framework, and has clear advantages over the existing methods. First, the method requires one set of calibration variables that can predict the outcome variable well, which reduces the need for large sample sizes of the probability-based data. From both modeling and data-collection perspective, model-assisted calibration is much more practical than propensity-score adjustment. Furthermore, the set of calibration variables can be much more diverse than the available Census totals, which allows for calibration to address a wider range of outcome of interests than traditional calibration methods. Despite the promising features of model-assisted calibration, there are two important gaps in the framework:

(1) **Limited choice of models**. The effectiveness of model-assisted calibration relies on models with strong predictive property: model parameters estimated from one sample can be used to reliably predict values in a different sample of the same population. Specifically, a model that prevents over-fitting can be most successful. Currently, there is no framework established for using models outside of traditional maximum likelihood-based models. Best-subset regression, or traditional regressions with stepwise procedures are the most commonly used assisting models. These methods have no theoretical justification in determining the best set of variables for calibration.

(2) **Lack of framework to account for benchmark sample uncertainties**. The current model-assisted calibration framework assumes that the benchmark sample is larger than the analytical sample (the sample that requires weighting adjustment). However, it is much more likely that the non-probability samples have significantly larger sizes than the probability-based benchmark data. The theoretical framework to incorporate uncertainties from smaller benchmark samples has not been established for model-assisted calibration. When benchmark data is small, without accounting for the uncertainties in the benchmark, such as induced by sample weights, inferences based on model-assisted calibrations would have underestimated standard errors.

This chapter addresses the two missing pieces of model-assisted calibration to enable the method as a practical post-survey adjustment tool for non-probability samples. Specifically, we employ a widely used modern statistical model, 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 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). For the ease of notation, we use adaptive LASSO and LASSO interchangeably. 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.

The organization of the chapter is as follows. Section 3.2 provides background and notation for traditional post-survey weighting schemes used for non-probability samples. Section 3.3 provides background and notation for model-assisted calibration. Section 3.4 develops the main theoretical framework for this research: we formulate the ECLASSO estimator for a population total of continuous and binary outcome variables, $\hat{T}_y^{ECLASSO}$, derive its asymptotic expectation and derive asymptotic linearized variance estimates. The theoretical framework is defined under probability-based sampling. The non-probability-based sampling is equivalent, except we assume the initial design weights are based on simple-random-sampling regardless of how the samples are formed. Section 3.5 describes the simulation used to evaluate $\hat{T}_y^{ECLASSO}$ and the asymptotic linearized variance estimates. We compare ECLASSO estimator with estimates from traditional weighting adjustment methods. The simulation results are discussed in section 3.6. The chapter ends with summaries in Section 3.7.

3.2 Weighting non-probability samples

3.2.1 Propensity-score weighting

Suppose a non-probability sample and a probability-based reference sample are available, with a common set of measures, $\mathbf{X}$. The objective of propensity-score weighting is to estimate the conditional probability that respondent i is a non-probability sample respondent given $\mathbf{X}$:

$$Z_i = \begin{cases} 1, & \text{if respondent } i \text{ is a non-probability sample respondent} \\ 0, & \text{otherwise} \end{cases}$$

$$p_i = Pr\left(Z_i = 1 \middle| \mathbf{X}\right) \tag{3.2.1.1}$$

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 $\mathbf{X}$: $P(Z = 1|\mathbf{X}, Y) = P(Z = 1|\mathbf{X})$ for almost all $\mathbf{X}$ and Y. In words, after controlling for $\mathbf{X}$, Y in the non-probability samples have the same distribution as the Y in probability samples. In missing data mechanism, this falls under Missing At Random (MAR) (*Rubin*, 1976): Y observed in the non-probability samples can be used to infer the unobserved Y in probability samples after controlling for $\mathbf{X}$. Thus the weights can correctly "inflate" the non-probability sample measures to make inference on the population. The key for propensity-score weighting is the set of $\mathbf{X}$ to achieve conditional independence. In practice, it is recommended that all possible variables to be used in $\mathbf{X}$, which can be impractical if the reference sample is small. Note that if Y itself is related to Z, we cannot achieve conditional independence, and propensity-score weighting would fail. This situation matches Not Missing At Random (NMAR) missing data mechanism. The most common type of model used to calculate p_i is logistic regression:

$$Pr(Z_i|\mathbf{X}, \boldsymbol{\beta}) = p_i$$
$$p_i = expit(\mathbf{X}\boldsymbol{\beta}), \quad expit(u) = (1 + exp(-u))^{-1}$$

In this chapter, we will refer to the logistic model used for propensity-score estimates as the working model for propensity-score weighting. The estimator of total based on propensity-score weights, given by:

$$\hat{T}_y^{PSCORE} = \sum_{i \in s_A} w_i^{PSCORE} y_i \tag{3.2.1.2}$$

where s_A is the non-probability sample. The remainder of this chapter focuses on calibration-based weighting adjustments that utilize an external probability-based data. Calibration refers to the probability-based data as benchmark samples, while propensity-score adjustment uses the term reference sample. For the remainder of this chapter, we will call the probability-based external data "benchmark samples".

3.2.2 Traditional calibration

For an analytical sample s_A (the sample which requires weight calibration) of size n_A drawn from sample design $\mathcal{A}$ with design weights $\underset{n_A \times 1}{\mathbf{d}^A}$, and the diagonal matrix of design weights $\mathbf{D}^A$, calibrated weights $\underset{n_A \times 1}{\mathbf{w}}$ minimize a distance measure:

$$E_{\mathcal{A}}\left[\sum_{i \in s_A} g(w_i, d_i^A)/q_i\right] \tag{3.2.2.1}$$

under the constraint:

$$\sum_{i \in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}} \tag{3.2.2.2}$$

where $g(w_i, d_i^A)$ 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 constant q_i is independent of design weight d_i^A. We focus on the most common distance measure used, the chi-square distance: $g(w_i, d_i^A) = (w_i - d_i^A)^2/d_i^A$ with $q_i = 1$. Under this distance measure:

$$\mathbf{w}^{GREG} = \mathbf{d}^A + \mathbf{D}^A\mathbf{X}\left(\mathbf{X}^T\mathbf{D}^A\mathbf{X}\right)^{-1}\left(\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)^T \tag{3.2.2.3}$$

where $\mathbf{T}^X$ is a row vector of known population totals of $\mathbf{X}$. The estimate of population total of outcome $\mathbf{y}$ based on calibrated weights:

$$\begin{aligned}\hat{T}_y^{GREG} &= \mathbf{w}^T\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{D}^A\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}^A\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)\hat{\boldsymbol{\beta}} \end{aligned} \tag{3.2.2.4}$$

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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, given weights $\mathbf{D}^A$. The calibrated weights defined in equation (3.2.2.3) do not rely on any outcome variable. Thus the same set of weights can be applied to all variables in the survey. The weighted total expressed in equation (3.2.2.4) corresponds to the generalized regression estimator (GREG) of total, thus we denote the weights $\mathbf{w}^{GREG}$ and the estimator $\hat{T}_y^{GREG}$. In GREG, an implicit linear relationship is assumed. The linear model, $E\left[y_i \middle| \mathbf{x}_i, \boldsymbol{\beta}\right] = \mathbf{x}_i^T\boldsymbol{\beta}$, is referred to as the working model for GREG. When the relationship between $\mathbf{y}$ and $\mathbf{X}$ is non-linear, such as in the case when $\mathbf{y}$ is binary, variance of $\hat{T}_y^{GREG}$ can be larger than the variance of pure-design based estimator of total (an estimator not using auxiliary totals).

To incorporate uncertainties from benchmark totals, *Dever* (2008) introduced estimated-control calibration. The framework replaces known population totals $\mathbf{T}^{\mathbf{X}}$ in equation (3.2.2.3) by estimated totals from the benchmark $\hat{\mathbf{T}}^{\mathbf{X}}$ (*Dever and Valliant*, 2010, 2016):

$$\mathbf{w}^{ECGREG} = \mathbf{d}^A + \mathbf{D}^A\mathbf{X}\left(\mathbf{X}^T\mathbf{D}^A\mathbf{X}\right)^{-1}\left(\hat{\mathbf{T}}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)^T \tag{3.2.2.5}$$

The resulting estimator of population total:

$$\begin{aligned}\hat{T}_y^{ECGREG} &= \mathbf{w}^T\mathbf{y}\\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\hat{\mathbf{T}}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{D}^A\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}^A\mathbf{y}\\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\hat{\mathbf{T}}^{\mathbf{X}} - (\mathbf{d}^A)^T\mathbf{X}\right)\hat{\boldsymbol{\beta}} \end{aligned} \tag{3.2.2.6}$$

The estimate-control calibration estimator 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.

One of the key assumptions in traditional calibration weighting is that the outcome of interest within the cells defined by calibration variables have the same distributions between sample and non-sample members. From the missing data literature, we assume that the unobserved cases in cells defined by $\mathbf{X}$ are missing at random. Thus "inflating" the observed measures within the calibration cells can effectively compensate the unobserved measures. This is similar to the assumption made for propensity-score weighting. If MAR is violated, we would be weighting the sample incorrectly, and the resulting weighted analysis can produce biased inference. Under probability-based-sampling, where the collected responses are based on a randomized sample of the population, the outcome of interest may already be similar among sample and non-sample members for large subgroups of the data. Thus we may not need to have many calibration variables to correctly weight the sample to the population. For non-probability samples, however, the participation probabilities can be very complex (*Valliant and Dever*, 2011). A large number of cells are needed to satisfy the MAR assumption. For calibration, this translates to a large number of totals to control to, which can greatly increase the risk of sparse cells that result in unstable calibrated weights, i.e., weights that produce large variance of weighted estimates. The problem is made worse in ECGREG, where the benchmark sample is small. Researchers often eliminate or collapse variable categories in the process of constructing calibrated weights (*Liu et al.*, 2012), which can easily violate the MAR assumption. An alternative calibration framework, model-assisted calibration, controls to the total of predicted outcome values instead of totals in **X**. Under this framework, the impact of sparse calibration cells is reduced. The inference no longer relies heavily on the MAR assumption. Instead, bias and variance property of weighted estimates in model-assisted calibration rely on the model's capability to estimate expected values of an outcome given the calibration variables. Thus model-assisted calibration is a practical approach to calibrate non-probability samples. In the next section, we provide background and notations of model-assisted calibration and our choice of assisting-model in this research: Least Angle Shrinkage and Selection Operator, LASSO (*Tibshirani*, 1996).

3.3 Model-assisted calibration

Similar to traditional calibration, we have an analytical sample s_A with size n_A, drawn from sample design $\mathcal{A}$ with design weights $\underset{n_A \times 1}{\mathbf{d}^A}$, and the diagonal matrix of design weights $\mathbf{D}^A$.

3.3.1 Background and notations

In model-assisted calibration, we assume a relationship between an outcome $\mathbf{y}$ and $\mathbf{X}$ through first two moments:

$$E_\xi(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta}), V_\xi(y_k|\mathbf{x}_k) = \nu_k^2\sigma^2 \tag{3.3.1.1}$$

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 $\boldsymbol{\beta}$, 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 $\boldsymbol{\beta}$ (i.e., the quasilikelihood estimator of $\boldsymbol{\beta}$ based on the entire finite population), and $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$. The model-assisted calibrated weights $\mathbf{w}$ minimize a distance measure:

$$E_A\left[\sum_{i\in s_A} g(w_i, d_i^A)/q_i\right] \tag{3.3.1.2}$$

under the constraints:

$$\begin{aligned}\sum_{i\in s_A} w_i &= N \\ \sum_{i\in s_A} w_i\hat{\mu}_i &= \sum_{k=1}^{N}\hat{\mu}_k.\end{aligned} \tag{3.3.1.3}$$

Equations (3.3.1.2) and (3.3.1.3) are defined for single-stage sampling designs. *Kennel* (2013) further extends model-assisted framework for clustered samples. The main conceptual difference between traditional calibration and model-assisted calibration is that in model-assisted calibration, the constraints are based on two quantities: (1) population total, and (2) population total of predicted values $\hat{\mu}_k$. In traditional calibration, the constraint is a vector of population totals of $\mathbf{X}$ (see equation (3.2.2.2)). Under chi-square distance measure with $q_i = 1$, the model-assisted calibrated weights are:

$$\mathbf{w}^{MC} = \mathbf{d}^A + \mathbf{D}^A\mathbf{M}\left(\mathbf{M}^T\mathbf{D}^A\mathbf{M}\right)^{-1}\left(\mathbf{T}^M - (\mathbf{d}^A)^T\mathbf{M}\right)^T \tag{3.3.1.4}$$

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

$$\begin{aligned}\hat{T}_y^{MC} &= \left(\mathbf{w}^{MC}\right)^T \mathbf{y} \\ &= (\mathbf{d}^A)^T \mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - (\mathbf{d}^A)^T \mathbf{X}\right) \left(\mathbf{X}^T \mathbf{D}^A \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D}^A \mathbf{y} \\ &= (\mathbf{d}^A)^T \mathbf{y} + \left(\sum_{k=1}^{N} \hat{\mu}_k - \sum_{i \in s_A} d_i^A \hat{\mu}_i\right) \hat{B}^{MC} \qquad (3.3.1.5)\end{aligned}$$

where $\hat{B}^{MC}$ is the calibration slope to satisfy the calibration constraints (different from the model parameter estimates $\hat{\mathbf{B}}$):

$$\begin{aligned}\hat{B}^{MC} &= \frac{\sum_{i \in s_A} d_i^A (\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{\sum_{i \in s_A} d_i^A (\hat{\mu}_i - \hat{\bar{\mu}})^2} \\ \hat{\bar{\mu}} &= \sum_{i \in s_A} d_i^A \hat{\mu}_i \Big/ \sum_{i \in s_A} d_i^A \\ \bar{y} &= \sum_{i \in s_A} d_i^A y_i \Big/ \sum_{i \in s_A} d_i^A .\end{aligned}$$

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: $\hat{\mathbf{T}}^M = (\sum_{i \in s_B} d_i^B, \sum_{i \in s_B} \hat{\mu}_i)$, where s_B denotes the benchmark sample and d_i^B is the probability-based design weights of the benchmark sample:

$$\hat{T}_y^{ECMC} = (\mathbf{d}^A)^T \mathbf{y} + \left(\sum_{i \in s_B} d_i^B \hat{\mu}_i - \sum_{i \in s_A} d_i^A \hat{\mu}_i\right) \hat{B}^{MC}. \qquad (3.3.1.6)$$

The asymptotic properties and variance of $\hat{T}_y^{ECMC}$ have not been established in the literature. For non-probability samples, it is essential that uncertainties of benchmark samples are included in the estimation formulas, because the probability-based samples with appropriate set of calibration variables are almost certainly smaller than the non-probability sample. Without accounting for the benchmark sample uncertainties, we would be making inferences with underestimated standard errors.

Wu and Sitter (2001) has shown that $\hat{T}_y^{MC}$ is asymptotically design unbiased, even when the model is miss-specified. As long as the original design weights produce unbiased estimates, $\hat{T}_y^{MC}$ is approximately unbiased when the sample size is large. Under non-probability-based-sampling, there are no initial design weights to guarantee unbiasedness of weighted estimates. Thus we rely on models that can approximate the expected value of y_i closely to compensate the lack of design weights. In the expression for $\mathbf{T}^M$, we use model parameters obtained under the analytical sample s_A to make predictions of $\sum_{k=1}^{N}\hat{\mu}_k$ in the population. Thus the key to successful model-assisted calibration is the model's ability to predict values on an independent dataset accurately. For both practical and theoretical reasons, there has been very little development in the use of assisting models outside of maximum-likelihood-based models. The choices of calibration variables are typically made by substantive interest, or through some stepwise procedures in traditional regression, or largely left to the discretion of a modeler. In this research, we introduce LASSO regression into model-assisted calibration framework to guide the selection of calibration variables in order to accurately approximate population total of $\hat{\mu}_i$. In the next section, we give an overview of LASSO regression, and how to determine the LASSO tuning parameters used for this research.

3.3.2 Assisting model - LASSO

3.3.2.1 Background and notations

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \left| \beta_j \right| \right). \qquad (3.3.2.1)$$

Similarly for logistic adaptive LASSO:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} \left[-y_i \left(\mathbf{x}_i^T \beta \right) + log \left(1 + exp \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) \right) \right] + \lambda_n \sum_{j=1}^{p} \alpha_j^\gamma \left| \beta_j \right| \right). \quad (3.3.2.2)$$

Given λ_n and γ, we can calculate $\hat{\boldsymbol{\beta}}$ through some iterative procedures.

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: $\alpha_j = 1/\left|\hat{\beta}_j^{MLE}\right|$, 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. For example, if we still want LASSO to favor large effect covariates when the sample size is small, we should set γ small. If we want to de-emphasize effect sizes further, we should set γ large. λ_n and γ are closely related to the model-consistency property of adaptive LASSO, known as the oracle property. Suppose the parameters in a full regression model have both zero and non-zero components, without loss of generality, let the first p be non-zero and the last q zero:

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

Zou (2006) has shown that if:

$$\lambda_n / \left(\sqrt{n} / (\sqrt{n})^\gamma \right) \to \infty \quad \text{and} \quad \lambda_n / \sqrt{n} \to 0$$

then the adaptive LASSO satisfies the oracle property:

- The probability of estimating 0 for zero-valued parameters tends to one: $Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \to 1$.

- The estimates of non-zero parameters are as good as if the true sub-model is known:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \rightarrow N\left(\mathbf{0}, \mathbf{C}\right)$$

where $\mathbf{C} = \Sigma(\boldsymbol{\beta}^{(1)})$ is covariance matrix of $\boldsymbol{\beta}$ under linear model, and $\mathbf{C} = I^{-1}(\boldsymbol{\beta}^{(1)})$ is the inverse of Fisher information matrix of $\boldsymbol{\beta}$ under generalized linear model. For finite-population inference, suppose ν indexes a population with size N_ν. Let $\mathbf{B}$ be the quasilikelihood estimates of $\boldsymbol{\beta}$ in population ν, and $\hat{\mathbf{B}}$ is the estimate of $\mathbf{B}$ based on a sample with size $n_\nu \leq N_\nu$. The finite-population equivalent of the oracle property is:

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

where $\mathbf{C}_\nu = \Sigma(\mathbf{B}^{(1)})$ is covariance matrix of $\mathbf{B}^{(1)}$ under linear model, and $\mathbf{C} = I^{-1}(\mathbf{B}^{(1)})$ is the inverse of Fisher information matrix of $\mathbf{B}^{(1)}$ under generalized linear model. For convenience, we omit ν from the notations. It is assumed that N and n are sequences of numbers, both grow to infinity as $\nu \rightarrow \infty$. We write $\mathbf{B} \rightarrow \boldsymbol{\beta}$ to mean that $\mathbf{B}$ approaches $\boldsymbol{\beta}$ as both sample and population sizes grow.

For adaptive LASSO to achieve the oracle property, the conditions require that λ_n grow at least at the rate of $\sqrt{n}/(\sqrt{n})^\gamma$, but not faster than $\sqrt{n}$. We discuss the choice of λ_n and γ in the next section.

3.3.2.2 Determining parameter values and estimates

In practice, we do not observe the theoretical rate of growth of λ_n, unless we have obtained many samples of the same population with various sample sizes. Given a sample, the choices of λ_n and γ depend on the modeler. Thus λ_n and γ are also called tuning parameters for LASSO regression. In R *glmnet* implementation (*Friedman et al.*, 2010), a range of λ_n is determined by the following scheme:

(1) Set $\gamma = 0$.

(2) Determine λ_n^{max} by finding the smallest λ_n that sets all coefficients to 0.

(3) If sample size n is larger than the number of parameters in the regression model, set $\lambda_n^{min} = 0.0001\lambda_n^{max}$. If sample size n is smaller than the number of parameters, set $\lambda_n^{min} = 0.01\lambda_n^{max}$ (to set parameters to 0 sooner).

(4) Generate a grid of λ_n, typically 100 equally spaced points between λ_n^{min} and λ_n^{max}.

The initial range of values of λ_n is determined independently of γ. Choices of γ is less data-driven. Some modelers choose one of $\gamma = 0.1, 0.5, 1, 2$. In this chapter, we determine (λ_n, γ) through cross-validation as follows:

Step 1. Obtain $\alpha_j = 1/\left|\hat{\beta}_j^{MLE}\right|$.

Step 2. Determine 100 equally spaced values of λ_n based on R *glmnet*'s implementation.

Step 3. For each pair (λ_n, γ), λ_n from Step 2, and $\gamma = 0.1, 0.5, 1, 2$, split data into 5 folds. Use 4 folds to obtain $\hat{\boldsymbol{\beta}}$.

Step 4. Apply $\hat{\boldsymbol{\beta}}$ to the last fold not used to estimate $\hat{\boldsymbol{\beta}}$ and calculate a metric. For continuous $\mathbf{y}$, we calculate the mean-absolute-error (MAE), $\sum_{i \in s_{A(k)}} |\hat{\mu}_i - y_i|$. For binary $\mathbf{y}$, we calculate the area under curve (AUC) (calculated through R *glmnet* :: *auc* function).

Step 5. Average the 5 metrics for each pair of (λ_n, γ), and choose the pair with the best average metric: minimum MAE for continuous $\mathbf{y}$, maximum AUC for binary $\mathbf{y}$.

The adaptive LASSO coefficient estimates are obtained given the selected (λ_n, γ). The R code used to perform cross-validation in this dissertation is in Appendix A.2.

3.4 Estimated control LASSO calibration

This section develops the main theoretical framework for our proposed method: Estimated Control LASSO (ECLASSO) calibration. Estimated controls are benchmark sample estimates rather than known population quantities. We develop the analytical formula for ECLASSO estimator of total, its asymptotic expectation, and asymptotic linearized variance estimates. We make the following assumptions in the theoretical framework:

A. The analytical samples, s_A with size n_A, are drawn from a single-stage sampling design $\mathcal{A}$, allowing for unequal probabilities of selection. The selection probability for unit i is denoted by π_i^A, and the joint selection probability of units i and j is 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 $\mathbf{X}^A$.

B. The benchmark samples, s_B with size n_B, are drawn from a single-stage sampling design $\mathcal{B}$, 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 $\mathbf{X}^B$.

C. A superpopulation model is assumed, as is described in section 3.3:

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

D. The true superpopulation parameters are a subset of the full regression model for LASSO: $\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}$

E. The full-range of $\mathbf{X}$ in the population has non-zero probability of being observed in both analytical and benchmark samples.

3.4.1 Point estimate: $\hat{T}_y^{ECLASSO}$

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

Step 1. Obtain LASSO regression coefficients $\hat{\mathbf{B}}$ as described in section 3.3.2. We use R package *glmnet* (*Friedman et al.*, 2010) to obtain LASSO coefficients for both linear and glm models, given a pair of (λ_n, γ) selected by cross-validation. For linear LASSO:

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

For logistic LASSO:

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

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

Step 3. Define $\hat{\mathbf{T}}^M = \left(\sum_{i\in s_B} d_i^B, \sum_{i\in s_B}\hat{\mu}_i\right)$ and $\mathbf{M} = \left[\mathbf{d}^A, (\hat{\mu}_i)_{i\in s_A}\right]$, under chi-square distance measure with $q_i = 1$:

$$\mathbf{w}^{LASSO} = \mathbf{d}^A + \mathbf{D}^A\mathbf{M}\left(\mathbf{M}^T\mathbf{D}^A\mathbf{M}\right)^{-1}\left(\hat{\mathbf{T}}^M - (\mathbf{d}^A)^T\mathbf{M}\right)^T \tag{3.4.1.1}$$

Step 4. ECLASSO calibration estimator of total:

$$\begin{aligned}
\hat{T}_y^{ECLASSO} &= \left(\mathbf{w}^{ECLASSO}\right)^T\mathbf{y} \\
&= (\mathbf{d}^A)^T\mathbf{y} + \left(\sum_{i\in s_B} d_i^B\hat{\mu}_i - (\mathbf{d}^A)^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{D}^A\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}^A\mathbf{y} \\
&= (\mathbf{d}^A)^T\mathbf{y} + \left(\sum_{i\in s_B} d_i^B\hat{\mu}_i - \sum_{i\in s_A} d_i^A\hat{\mu}_i\right)\hat{B}^{MC} \qquad (3.4.1.2)
\end{aligned}$$

where $\hat{B}^{MC}$ is the calibration slope to satisfy the calibration constraints (different from the model parameter estimates $\hat{\mathbf{B}}$):

$$\begin{aligned}
\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})^2} \\
\bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A\hat{\mu}_i / \sum_{i\in s_A} d_i^A \\
\bar{y} &= \sum_{i\in s_A} d_i^A y_i / \sum_{i\in s_A} d_i^A
\end{aligned}$$

Note that the main difference between ECLASSO calibration and the model-assisted calibration is the use of benchmark sample weights in calculating $\hat{\mathbf{T}}^M$.

3.4.2 Asymptotic estimator of total

In this section, we derive the asymptotic estimated-control model-assisted LASSO calibration (ECLASSO) estimator of total, $\hat{T}_y^{ECLASSO}$. The asymptotic ECLASSO estimator is later used to derive asymptotic expectation and asymptotic linearized variance estimates of $\hat{T}_y^{ECLASSO}$. We assume that the finite-population and sample size are sequences of numbers indexed by ν: N_ν and n_ν. Both N_ν and n_ν grow to infinity. For simplicity, ν is omitted from the suffix of N and n. We first derive the asymptotic estimated-control model-assisted calibration (ECMC) estimator of population mean, then apply additional conditions to derive the asymptotic ECLASSO estimator of population total. Unless stated otherwise, n refers to the analytical sample size. The following conditions are necessary to derive the asymptotic estimators:

(3.4.2.i) $\hat{\mathbf{B}} = \mathbf{B} + O_p(1/\sqrt{n})$, $\mathbf{B}$ is the finite-population regression slope of $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$.

(3.4.2.ii) For each $\mathbf{x}_i$, $\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}$ is continuous in $\mathbf{t}$ and bounded: $max_i\,|\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}| \leq h(\mathbf{x}_i, \boldsymbol{\beta})$ for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\beta}$, and $N^{-1}{\sum}_{i\in U} h(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.

(3.4.2.iii) For each $\mathbf{x}_i$, $\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}\partial\mathbf{t}^T$ is continuous in $\mathbf{t}$ and bounded: $max_{j,k}\,|\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial t_j \partial t_k| \leq$ $k(\mathbf{x}_i, \boldsymbol{\beta})$ for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\beta}$, and $N^{-1}{\sum}_{i\in U} k(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.

(3.4.2.iv) The Horvitz-Thompson estimators of certain population means are asymptotically normally distributed for $\mathbf{d}^A$.

(3.4.2.v) The Horvitz-Thompson estimators of certain population means are asymptotically normally distributed for $\mathbf{d}^B$.

(3.4.2.vi) $\lambda_n \big/ \left(\sqrt{n}/(\sqrt{n})^{\gamma}\right) \to \infty \quad \text{and} \quad \lambda_n \big/ \sqrt{n} \to 0.$

Remark III.1. The mean in conditions in (3.4.2.iv) and (3.4.2.v) are the means of first and second derivatives of $\mu(\mathbf{x}_i, \mathbf{t})$ in the Taylor series expansion of $\mu(\mathbf{x}_i, \mathbf{t})$ evaluated at a neighborhood around $\mathbf{B}$, which is a vector of values if $\mathbf{B}$ has more than one parameter. The conditions require that the Horvitz-Thompson estimates of the means are bounded element-wise.

Lemma III.2. *Let s_B be a probability-based benchmark sample with size n_B and design weights $\mathbf{d}^B$, and s_A be an analytical sample with size n_A and design weights*

$\mathbf{d}^A$, and N be a known population total derived from a sample bigger than s_B and s_A. Assume the superpopulation model:

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

Let $\mathbf{B}$ be the finite-population quasilikelihood estimate of $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. Under conditions (3.4.2.i)-(3.4.2.v), the asymptotic estimated-control calibration estimator of population total is:

$$\hat{T}_y^{ECMC} = \mathbf{d}^A\mathbf{y} + \left(\sum_{i\in s_B} d_i^B\mu_i - \sum_{i\in s_A} d_i^A\mu_i\right) B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right)$$

$$\begin{aligned} n^* &= min(n_A, n_B) \\ B^{MC} &= \frac{\sum_{i=1}^N(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^N(\mu_i - \bar{\mu})^2} \\ \bar{\mu} &= N^{-1}\sum_{i=1}^N \mu_i, \quad \bar{y} = N^{-1}\sum_{i=1}^N y_i \end{aligned}$$

Proof. We begin by deriving the asymptotic model-assisted estimator for a population mean, $\hat{\bar{y}}^{ECMC} = N^{-1}\hat{T}_y^{ECMC}$ (see equation (3.4.2.11)). By conditions (3.4.2.ii) and (3.4.2.iii), the second order Taylor series expansion of $\mu(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ around $\mathbf{B}$ is:

$$\mu(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) = \mu(\mathbf{x}_i, \mathbf{B}) + \left\{\frac{\mu(\mathbf{x}_i, \mathbf{t})}{\partial\mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}}\right\}^T (\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \left\{\frac{\partial^2\mu(\mathbf{x}_i, \mathbf{t})}{\partial\mathbf{t}\partial\mathbf{t}^T}\bigg|_{\mathbf{t}=\mathbf{B}^*}\right\} (\hat{\mathbf{B}} - \mathbf{B}) \tag{3.4.2.1}$$

for $\mathbf{B}^* \in (\hat{\mathbf{B}}, \mathbf{B})$ or $(\mathbf{B}, \hat{\mathbf{B}})$. Let

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

Note that $\mathbf{h}$ is a vector of length m and $\mathbf{k}$ is a matrix of size $m \times m$, where m is the number of parameters in $\boldsymbol{\beta}$. By conditions (3.4.2.ii) and (3.4.2.iii),

$$max_i \left|\mathbf{h}(\mathbf{x}_i, \mathbf{B})\right| \leq h(\mathbf{x}_i, \mathbf{B}) \qquad (3.4.2.2)$$

$$max_{k,j} \left|\mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\right| \leq k(\mathbf{x}_i, \mathbf{B}^*) \qquad (3.4.2.3)$$

The population mean of (3.4.2.1) based on sample s_B:

$$\begin{aligned} N^{-1}\sum_{i \in s_B} d_i^B \mu(\mathbf{x}_i, \hat{\mathbf{B}}) &= N^{-1}\sum_{i \in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i \in s_B} d_i^B \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T (\hat{\mathbf{B}} - \mathbf{B}) + \\ &\quad O_p\left(\frac{1}{\sqrt{n_B}}\right) O_p\left(\frac{1}{\sqrt{n_B}}\right) \\ &= N^{-1}\sum_{i \in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i \in s_B} d_i^B \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T (\hat{\mathbf{B}} - \mathbf{B}) + O_p\left(\frac{1}{n_B}\right) \end{aligned} \qquad (3.4.2.4)$$

Similarly, the population mean of (3.4.2.1) based on sample s_A:

$$N^{-1}\sum_{i \in s_A} d_i^A \mu(\mathbf{x}_i, \hat{\mathbf{B}}) = N^{-1}\sum_{i \in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i \in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T (\hat{\mathbf{B}} - \mathbf{B}) + O_p\left(\frac{1}{n_A}\right) \qquad (3.4.2.5)$$

By conditions (3.4.2.i), (3.4.2.iv), (3.4.2.v), and equations (3.4.2.4) and (3.4.2.5):

$$
\begin{aligned}
&N^{-1}\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \hat{\mathbf{B}}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \hat{\mathbf{B}}) \\
&\quad = N^{-1}\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(\frac{1}{\sqrt{n^*}}\right) + O_p\left(\frac{1}{n^*}\right) \\
&\quad = N^{-1}\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(\frac{1}{\sqrt{n^*}}\right) \qquad (3.4.2.6)
\end{aligned}
$$

where $n^* = min(n_A, n_B)$. Note that,

$$
\begin{aligned}
\bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \hat{\mathbf{B}}) \Big/ \sum_{i\in s_A} d_i^A \\
&= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \Big(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)(\hat{\mathbf{B}} - \mathbf{B})\Big)
\end{aligned}
$$

(by conditions (3.4.2.i) and (3.4.2.iii))

$$
\begin{aligned}
&= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \Big(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B})\Big) + O_p(1/n_A) \\
&= \bar{\mu} + \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + O_p(1/n_A)
\end{aligned}
$$

(by conditions (3.4.2.i) and (3.4.2.2))

$$
\begin{aligned}
&= \bar{\mu} + O_p(1/\sqrt{n_A}) + O_p(1/n_A) \\
&= \bar{\mu} + O_p(1/\sqrt{n_A}) \qquad (3.4.2.7)
\end{aligned}
$$

Then from (3.4.2.1) and (3.4.2.7),

$$N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})$$
$$= N^{-1}\sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i,\mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i,\mathbf{B})(\hat{\mathbf{B}}-\mathbf{B}) + (\hat{\mathbf{B}}-\mathbf{B})^T\mathbf{k}(\mathbf{x}_i,\mathbf{B}^*)(\hat{\mathbf{B}}-\mathbf{B}) - \hat{\mu}\right)$$
$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + N^{-1}\sum_{i\in s_A}\mathbf{h}^T(\mathbf{x}_i,\mathbf{B})(\hat{\mathbf{B}}-\mathbf{B}) +$$
$$N^{-1}\sum_{i\in s_A}(\hat{\mathbf{B}}-\mathbf{B})^T\mathbf{k}(\mathbf{x}_i,\mathbf{B}^*)(\hat{\mathbf{B}}-\mathbf{B}) - O_p(1/\sqrt{n_A})$$

(by conditions (3.4.2.i) and (3.4.2).iii)

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + N^{-1}\sum_{i\in s_A}\mathbf{h}^T(\mathbf{x}_i,\mathbf{B})(\hat{\mathbf{B}}-\mathbf{B}) + O_p(1/n_A) - O_p(1/\sqrt{n_A})$$

(by conditions (3.4.2.i) and (3.4.2.2))

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + O_p(1/\sqrt{n_A}) + O_p(1/n_A) - O_p(1/\sqrt{n_A})$$
$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + O_p(1/\sqrt{n_A}) \qquad (3.4.2.8)$$

$$N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})^2 = N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + (O_p(1/\sqrt{n_A}))^2$$
$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + O_p(1/n_A) \qquad (3.4.2.9)$$

From (3.4.2.8) and (3.4.2.9) we have:

$$\hat{B}^{MC} = \frac{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})^2} = \frac{N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})^2}$$
$$= \frac{\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})(y_i - \bar{y}) + O_p\left(\frac{1}{\sqrt{n_A}}\right)}{\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + O_p\left(\frac{1}{n_A}\right)}$$
$$\to B^{MC} \quad \text{as } n_A \to \infty \qquad (3.4.2.10)$$

Thus $\hat{B}^{MC} = B^{MC} + o_p(1)$, and we have:

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

where $n^* = min(n_A, n_B)$. Since $N = O_p(N)$, we have $N \cdot o_P(1/\sqrt{n^*}) = O_p(N)o_p(1/\sqrt{n^*}) = o_p(N/\sqrt{n^*})$. Thus,

$$\begin{aligned}
\hat{T}_y^{ECMC} &= N\hat{\bar{y}}^{ECMC} \\
&= N\left(N^{-1}\mathbf{d}^A\mathbf{y} + \left(N^{-1}\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} \mu(\mathbf{x}_i, \mathbf{B}) \right) B^{MC} + o_p\left(\frac{1}{\sqrt{n^*}}\right) \right) \\
&= \mathbf{d}^A\mathbf{y} + \left(\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) - \sum_{i\in s_A} \mu(\mathbf{x}_i, \mathbf{B}) \right) B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right) \qquad (3.4.2.11)
\end{aligned}$$

□

We are now ready to derive the asymptotic ECLASSO estimator of total.

Theorem III.3. *Suppose the parameters in a full regression model have both zero and non-zero components, without loss of generality, let the first p be non-zero and the last q be zero:* $\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}$, $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)}$. *Let* s_B *be a probability-based benchmark sample with design weights* $\mathbf{d}^B$, *and* s_A *be an analytical sample with design weights* $\mathbf{d}^A$, *and* N *be a known population total derived from a sample bigger than* s_B *and* s_A, *assume the superpopulation model:*

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

Let $\mathbf{B}$ *be the finite-population quasilikelihood estimate of* $\boldsymbol{\beta}$*,* $\mathbf{B} \to \boldsymbol{\beta}$*, under conditions (3.4.2.i)-(3.4.2.vi), the asymptotic ECLASSO calibration estimator of population total is:*

$$\hat{T}_y^{ECLASSO} = \sum_{i \in s_A} d_i^A (y_i - \mu_i B^{MC}) + \sum_{i \in s_B} d_i^B \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right) \tag{3.4.2.12}$$

where $n^* = min(n_A, n_B)$ *and* $\mu_i = \mu(\mathbf{x}_i, \mathbf{B})$*.*

Proof. Under condition (3.4.2.vi), the adaptive LASSO regression satisfies the oracle property through Theorems 1 and 4 in (*Zou*, 2006):

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

where $\mathbf{C} = \Sigma(\mathbf{B})$ is the covariance matrix of $\mathbf{B}$ under linear model, and $\mathbf{C} = I^{-1}(\mathbf{B})$ is the inverse of Fisher information matrix of $\mathbf{B}$ under generalized linear model. By Slutsky's theorem, the oracle property implies: $\hat{\mathbf{B}} = \mathbf{B} + O_p(1/\sqrt{n_A})$. Since LASSO regression satisfies condition (3.4.2.i), it is asymptotically equivalent to estimated-control model-assisted calibration estimator of total. By Lemma III.2:

$$\begin{aligned}\hat{T}_y^{ECLASSO} &= \hat{T}_y^{ECMC} \\ &= \sum_{i \in s_A} d_i^A \left(y_i - \mu(\mathbf{x}_i, \mathbf{B}) B^{MC}\right) + \sum_{i \in s_B} d_i^B \mu_i(\mathbf{x}_i, \mathbf{B}) B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right)\end{aligned} \tag{3.4.2.13}$$

□

With the asymptotic ECLASSO estimator of total, we can derive the asymptotic expectation of $\hat{T}_y^{LASSO}$.

Theorem III.4. *$\hat{T}_y^{ECLASSO}$ is asymptotically design and model-unbiased.*

Proof. Under the assumption of our theoretical framework, the superpopulation parameters are a subset of the full LASSO regression parameters, and that the sample design $\mathcal{B}$ is probability-based with design weights $\mathbf{d}^B$, we can prove the asymptotically design unbiasedness of $\hat{T}_y^{ECLASSO}$ by taking expectations with respect to model ξ and sample design $\mathcal{B}$. First note that:

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

Thus

$$\begin{aligned}
E_{\mathcal{B}}\left[\hat{T}_y^{ECLASSO} - T\right] &\approx E_{\mathcal{B}}\left[\sum_{i\in s_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i\in s_B} d_i^B \mu_i B^{MC} - \sum_{i=1}^N y_i\right] \\
&= E_{\mathcal{B}}\left[E_\xi\left[\sum_{i\in s_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i\in s_B} d_i^B \mu_i B^{MC} - \sum_{i=1}^N y_i\right]\right] \\
&\quad \text{(since } E_\xi[y_i] = \mu_i \text{ and } E_\xi\left[B^{MC}\right] = 1\text{)} \\
&= E_{\mathcal{B}}\left[\sum_{i\in s_A} d_i^A(\mu_i - \mu_i) + \sum_{i\in s_B} d_i^B \mu_i - \sum_{i=1}^N \mu_i\right] \\
&\quad \text{(since } \mathbf{d}^B \text{ is probability-sampling-based design weights)} \\
&= \sum_{i=1}^N \mu_i - \sum_{i=1}^N \mu_i \\
&= 0
\end{aligned}$$

□

As long as LASSO regression parameters include the superpopulation parameters (right set covariates are available to LASSO), and benchmark sample weights are probability-based, $\hat{T}_y^{ECLASSO}$ is unbiased regardless of analytical design weights. This property is essential in non-probability samples, where there are no initial design weights to guarantee unbiasedness.

3.4.3 Asymptotic design variance of $\hat{T}_y^{ECLASSO}$

We derive the asymptotic linearized variance estimate by taking the variance of equation (3.4.2.12) directly:

$$\begin{aligned}
v_{\mathcal{A}}(\hat{T}_y^{ECLASSO}) &= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i-\mu_i B^{MC}\right)+\sum_{i\in s_B} d_i^B \mu_i B^{MC}\right)\\
&= V_{\mathcal{A}}\left(E_{\mathfrak{B}}\left[\sum_{i\in s_A} d_i^A\left(y_i-\mu_i B^{MC}\right)+\sum_{i\in s_B} d_i^B \mu_i B^{MC}\right]\right)+\\
&\quad E_{\mathcal{A}}\left[V_{\mathfrak{B}}\left(\sum_{i\in s_A} d_i^A\left(y_i-\mu_i B^{MC}\right)\sum_{i\in s_B} d_i^B \mu_i B^{MC}\right)\right]\\
&= V_{\mathcal{A}}E_{\mathfrak{B}}+E_{\mathcal{A}}V_{\mathfrak{B}}
\end{aligned}$$

We derive each component $V_{\mathcal{A}}E_{\mathfrak{B}}$, $E_{\mathcal{A}}V_{\mathfrak{B}}$ separately:

$$\begin{aligned}
V_{\mathcal{A}}E_{\mathfrak{B}} &= V_{\mathcal{A}}\left(E_{\mathfrak{B}}\left[\sum_{i\in s_A} d_i^A\left(y_i-\mu_i B^{MC}\right)+\sum_{i\in s_B} d_i^B \mu_i B^{MC}\right]\right)\\
&= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i-\mu_i B^{MC}\right)+\sum_{i=1}^{N} \mu_i B^{MC}\right)
\end{aligned} \tag{3.4.3.1}$$

(since $\mathcal{A}$ is single-stage probability-based sampling)

$$\begin{aligned}
&= \sum_{i\in U}\left(\frac{y_i-\mu_i B^{MC}}{\pi_i^A}\right)^2 \pi_i^A(1-\pi_i^A)+\\
&\quad \sum_{i\in U}\sum_{j\neq i}\left(\pi_{ij}^A-\pi_i^A\pi_j^A\right)\frac{\left(y_i-\mu_i B^{MC}\right)}{\pi_i^A}\frac{\left(y_j-\mu_j B^{MC}\right)}{\pi_j^A}
\end{aligned} \tag{3.4.3.2}$$

We use sample estimates for population quantities in (3.4.3.2):

$$\widehat{V_{\mathcal{A}}E_{\mathcal{B}}} = \sum_{i \in s_A} \left(\frac{y_i - \hat{\mu}_i \hat{B}^{MC}}{\pi_i^A} \right)^2 (1 - \pi_i^A) + \\ \sum_{i \in s_A} \sum_{j \neq i} \frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_{ij}^A} \frac{\left(y_i - \hat{\mu}_i \hat{B}^{MC}\right)}{\pi_i^A} \frac{\left(y_j - \hat{\mu}_j \hat{B}^{MC}\right)}{\pi_j^A} \tag{3.4.3.3}$$

Now for the second component:

$$\begin{aligned} E_{\mathcal{A}}V_{\mathcal{B}} &= E_{\mathcal{A}} \left[V_{\mathcal{B}} \left(\sum_{i \in s_A} d_i^A \left(y_i - \mu_i B^{MC} \right) + \sum_{i \in s_B} d_i^B \mu_i B^{MC} \right) \right] \\ &= E_{\mathcal{A}} \left[V_{\mathcal{B}} \left(\sum_{i \in s_B} d_i^B \mu_i B^{MC} \right) \right] \end{aligned} \tag{3.4.3.4}$$

(since sample design $\mathcal{B}$ is single-stage probability-based sampling)

$$= \sum_{i \in U} \left(\frac{\mu_i B^{MC}}{\pi_i^B} \right)^2 \pi_i^B (1 - \pi_i^B) + \\ \sum_{i \in U} \sum_{j \neq i} \left(\pi_{ij}^B - \pi_i^B \pi_j^B \right) \frac{\mu_i B^{MC}}{\pi_i^B} \frac{\mu_j B^{MC}}{\pi_j^B} \tag{3.4.3.5}$$

We use sample estimates for population quantities in (3.4.3.5):

$$\widehat{E_{\mathcal{A}}V_{\mathcal{B}}} = \sum_{i \in s_B} \left(\frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \right)^2 (1 - \pi_i^B) + \\ \sum_{i \in s_B} \sum_{j \neq i} \frac{\pi_{ij}^B - \pi_i^B \pi_j^B}{\pi_{ij}^B} \frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \frac{\hat{\mu}_j \hat{B}^{MC}}{\pi_j^B} \tag{3.4.3.6}$$

Finally, the asymptotic linearized variance estimator of $\hat{T}_y^{ECLASSO}$ is:

$$\begin{aligned}
v_{\mathcal{A}}(\hat{T}_y^{ECLASSO}) &\approx \widehat{V_{\mathcal{A}}E_{\mathcal{B}}} + \widehat{E_{\mathcal{A}}V_{\mathcal{B}}} \\
&= \sum_{i \in s_A} \left(\frac{y_i - \hat{\mu}_i \hat{B}^{MC}}{\pi_i^A} \right)^2 (1 - \pi_i^A) + \\
&\quad \sum_{i \in s_A} \sum_{j \neq i} \frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_{ij}^A} \frac{(y_i - \hat{\mu}_i \hat{B}^{MC})}{\pi_i^A} \frac{(y_j - \hat{\mu}_j \hat{B}^{MC})}{\pi_j^A} + \\
&\quad \sum_{i \in s_B} \left(\frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \right)^2 (1 - \pi_i^B) + \\
&\quad \sum_{i \in s_B} \sum_{j \neq i} \frac{\pi_{ij}^B - \pi_i^B \pi_j^B}{\pi_{ij}^B} \frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \frac{\hat{\mu}_j \hat{B}^{MC}}{\pi_j^B}. \qquad (3.4.3.7)
\end{aligned}$$

An alternative linearized variance estimate, suggested by (*Särndal et al.*, 1989), multiplies $(y_i - \hat{\mu}_i \hat{B}^{MC})$ by g-weights, which are the ratios of calibrated weights to the original design weights:

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

$$\begin{aligned}
v.g_{\mathcal{A}}(\hat{T}_y^{ECLASSO}) &= \sum_{i \in s_A} \left(\frac{g_i \left(y_i - \hat{\mu}_i \hat{B}^{MC} \right)}{\pi_i^A} \right)^2 (1 - \pi_i^A) + \\
&\quad \sum_{i \in s_A} \sum_{j \neq i} \frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_{ij}^A} \frac{g_i(y_i - \hat{\mu}_i \hat{B}^{MC})}{\pi_i^A} \frac{g_j(y_j - \hat{\mu}_j \hat{B}^{MC})}{\pi_j^A} + \\
&\quad \sum_{i \in s_B} \left(\frac{g_i \left(\hat{\mu}_i \hat{B}^{MC} \right)}{\pi_i^B} \right)^2 (1 - \pi_i^B) + \\
&\quad \sum_{i \in s_B} \sum_{j \neq i} \frac{\pi_{ij}^B - \pi_i^B \pi_j^B}{\pi_{ij}^B} \frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \frac{\hat{\mu}_j \hat{B}^{MC}}{\pi_j^B}. \qquad (3.4.3.8)
\end{aligned}$$

To simplify notations, we refer to $v_{\mathcal{A}}(\hat{T}_y^{ECLASSO})$ as $v^{ECLASSO}$ and $v.g_{\mathcal{A}}(\hat{T}_y^{ECLASSO})$ as $v_g^{ECLASSO}$.

The theoretical framework is complete. We have developed the point estimate of ECLASSO calibration estimator of total and its asymptotic linearized variance estimates.

3.5 Simulation design

We design a simulation to evaluate ECLASSO calibration estimator of total: $\hat{T}_y^{ECLASSO}$, developed in Section 3.4.1, and linearized variance estimates of $\hat{T}_y^{LASSO}$: $v^{ECLASSO}$ and $v_g^{ECLASSO}$, developed in section 3.4.3. Since both linearized variance estimates are based on asymptotic LASSO calibration estimate of total, they might not perform well for small sample sizes. 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 $\hat{T}_y^{ECLASSO}$. We repeat the process 500 times per simulation sample to obtain the bootstrap variance estimate for the simulation sample. To simulation non-probability samples, we draw samples from the population with unequal probabilities, but set the design weights to 1.

3.5.1 Estimators

In addition to $\hat{T}_y^{ECLASSO}$, we will generate estimates based on traditional weighting schemes in the simulation. The estimators that are evaluated are:

1. Pure-design based Horvitz-Thompson estimator of total, assume SRS, HT: $\hat{T}_y^{HT} = (N/n)\sum_{i \in s_A} y_i$.

2. Traditional calibration estimator of total, GREG: $\hat{T}_y^{GREG}$ (see (3.2.2.4)).

3. Traditional estimated-control calibration estimator of total, ECGREG: $\hat{T}_y^{ECGREG}$ (see (3.2.2.6)).

4. Propensity-score weighting, PSCORE: $\hat{T}_y^{PSCORE}$ (see (3.2.1.2)).

3.5.2 Data and experimental groups

The data used as the population for this simulation is National Health Interview Survey (NHIS), 2013 adults data. The NHIS 2013 adults data is merged with NHIS 2013 family income data to obtain income-related variables. After removing respondents with missing values on demographics, income, and health indicators, the population size is $N = 31,914$. 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 can construct a model predicting *internet_use*, with *internet_health* as a predictor. The predicted probabilities 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. If the outcome of interest is associated with the general health well-being of a respondent, our samples are prone to selection bias. We choose the outcome of interest:

$$y_i = \begin{cases} 1, & \text{if respondent } i \text{ does not have health insurance coverage} \\ 0, & \text{if respondent } i \text{ does have health insurance coverage} \end{cases}$$

The goal is to predict the total number of individuals in the population without health insurance, $T = \sum_{i=1}^{N} y_i = 5,432$. Table 3.1 lists the variables that are used in the simulation.

The main goal of the simulation is to evaluate $\hat{T}_y^{ECLASSO}$ under different levels of sample and benchmark sizes:

- Analytical sample $n = 250, 500, 1000$

Table 3.1: Variables used in the working models

Used in working models			
Variable	**Name in model**	**Categories**	**Values**
Age	agegrp	18-30	0
		31-40	1
		41-50	2
		51-60	3
		61-70	4
		71-80	5
		81+	6
Age 65 or older	agegrp	No	0
		Yes	1
Gender	sex	Male	1
		Female	2
Race/Ethnicity	race	non-Hispanic white	1
		non-Hispanic black	2
		Hispanic	3
		Other	4
Education	educ2	Less than HS	0
		HS	1
		Some college	2
		Associate/bachlors	3
		Post-graduate	4
Currently employed	employed	No	0
		Yes	1
Seen health professional last 12 months	sathc	No	0
		Yes	1
Diagnosed with cancer	cancer	No	0
		Yes	1
Family income	faminc_q	1st quartile	0
		2nd quartile	1
		3rd quartile	2
		4th quartile	3
Additional variables used in generating sample selection probabilities			
Use internet	internet_use	No	0
		Yes	1
Region	region	Northeast	1
		Midwest	2
		South	3
		West	4
Marital status	marst	Married/partnered	1
		Widowed/divorced/seperated	2
		Never married	3
Work at a private firm	wrk_private	No	0
		Yes	1
Looked up health on internet	internet_health	No	0
		Yes	1

- Benchmark sample $n = 250, 500, 1000, 2000, 4000, 8000, 16000$

There are a total of $3 \times 7 = 21$ experimental groups.

3.5.3 Working models

Five sets of working models are defined for the estimators:

- 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:

$$\mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{region} + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ2}$$

- Trimmed:

$$\mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ2} + \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{employed}$$

- Partial:

$$\begin{aligned}\mathbf{x}_i^T \boldsymbol{\beta} =& \beta_0 + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ2} + \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{employed} + \\ & \beta_{k[i]}^{sex} \times \beta_{k[i]}^{age65} + \beta_{k[i]}^{race} \times \beta_{k[i]}^{age65}\end{aligned}$$

- Full:

$$\begin{aligned}\mathbf{x}_i^T \boldsymbol{\beta} =& \beta_0 + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ2} + \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{employed} + \\ & \beta_{k[i]}^{sex} \times \beta_{k[i]}^{age65} + \beta_{k[i]}^{race} \times \beta_{k[i]}^{age65} + \beta_{k[i]}^{race} \times \beta_{k[i]}^{faminc\_q}\end{aligned}$$

All variables are categorical. We denote $k[i]$ as the category respondent i belongs to for a given variable. For example, model coefficient for respondents age 31-40 is denoted by $\beta_{1[i]}^{agegrp}$. If respondent i is not in age group 31-40, then $\beta_{1[i]}^{agegrp} = 0$. The first value for each variable in Table 3.1 is used as the reference group. Depending on the estimator, the $\hat{\boldsymbol{\beta}}$ is obtained differently. For GREG and ECGREG, $\hat{\boldsymbol{\beta}}$ is obtained from a linear regression: $y_i = \mathbf{x}_i\boldsymbol{\beta}$. For PSCORE, $\hat{\boldsymbol{\beta}}$ is the solution to the logistic regression: $logit(z_i|\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i\boldsymbol{\beta}$. And for ECLASSO, $\hat{\boldsymbol{\beta}}$ is obtained through LASSO regression described in Section 3.3.2. To see the relationship of each variable to the binary outcome variable y_i, we fit 5 logistic regressions on the population. Table 3.2 lists the logistic 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.

Models Demographics1 and Demographics2 are the working models for traditional calibration estimators. We denote GREG1 and GREG2 to be the estimators using Demographics1 and Demographics2 respectively. We anticipate GREG1 to perform worse than estimators using other models, because the Demographics1 has the worst model-fitness measure for the population. The addition of education variable to Demographics1 improves model-fitness substantially (see Demographics2). Thus we expect GREG2 to do better than GREG1.

Models Trim, Partial, and Full represent three levels of complexity. ECLASSO uses the Full model in all experimental groups. There are two propensity-score estimators in this simulation. First one, PSCORE1, and ECGREG use the same working models. Because the larger models cannot be estimated in a stable manner from the small datasets, we use the following:

- When the minimum of analytical and benchmark sample size is 250, ECGREG and PSCORE1 use the Trimmed model.

- When the minimum of analytical and benchmark sample size is 500, ECGREG and PSCORE1 use the Partial model.
- When the minimum of analytical and benchmark sample size is 1,000, ECGREG and PSCORE1 use the Full model.

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.

3.5.4 Sample generation

To generate selection probabilities, we fit the logistic regression model on the whole population:

$$\begin{aligned} E\left[I(internet\_use = 1)\right] &= \pi_i^A \\ logit(\pi_i^A) &= \beta_0 + \beta_{k[i]}^{region} + \beta_{k[i]}^{sex} + \beta_{k[i]}^{agegrp} + \beta_{k[i]}^{race} + \beta_{k[i]}^{educ2} \\ &\quad \beta_{k[i]}^{faminc\_q} + \beta_{k[i]}^{marst} + \beta_{k[i]}^{sathc} + \beta_{k[i]}^{wrk\_private} + \beta_{k[i]}^{internet\_health} \end{aligned}$$

We use Poisson sampling with the predicted probabilities, $\hat{\pi}_i^A$, as the basis of our selection probabilities. For each analytical sample size n, the probabilities are rescaled to generate a sample size close to n on expectation: $\hat{\pi}_i^{A*} = n\hat{\pi}_i^A / \sum_{i=1}^N \hat{\pi}_i^A$. The selection probabilities simulate a person's propensity to be in a non-probability internet-based sample. The same type of sample generation was used in (*Valliant and Dever*, 2011) to simulate web-volunteer samples. The benchmark sample is a simple-random-sample of the population. Table 3.3 displays two model outputs. The output on the left is the fit to predict internet use, which generated the selection probabilities. The output on the right is the fit of the same variables to predict the outcome of interest. Income

Table 3.2: Working models fit on population

	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***
educ2[1]		−0.414***	−0.266***	−0.262***	−0.263***
educ2[2]		−0.833***	−0.588***	−0.585***	−0.592***
educ2[3]		−1.187***	−0.674***	−0.672***	−0.677***
educ2[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***
Observations	31,914	31,914	31,914	31,914	31,914
Log Likelihood	−12,819.310	−12,319.740	−10,198.670	−10,187.870	−10,173.000
Akaike Inf. Crit.	25,666.620	24,675.490	20,441.330	20,427.730	20,415.990
Note:					*p<0.1; **p<0.05; ***p<0.01

and education are highly significant in both models with large effect sizes. We expect weighting adjustments that control to these variables to do well. The variable that is strongly associated with *internet_use*, *internet_health*, is mildly significant with insurance coverage. Table 3.4 tabulates the total number of people without health insurance in the population, by quintile groups of selection probabilities. The group with highest selection probabilities also has the lowest count of total number without insurance. Assuming SRS, the estimate of total number of people without health insurance would be underestimating the true population total.

Table 3.5 lists the summary of simulation parameters.

3.5.5 Evaluation metrics

Point estimates and variance. We evaluate empirical bias, variance, and RMSE for each estimator of total. Let S be the number of simulation iterations. We define:

$$\hat{\theta} = \hat{T}_y^{HT}, \quad \hat{T}_y^{GREG1}, \quad \hat{T}_y^{GREG2}, \quad \hat{T}_y^{PSCORE1}, \quad \hat{T}_y^{PSCORE2}, \quad \hat{T}_y^{ECLASSO}$$

$$\begin{aligned} bias\left(\hat{\theta}\right) &= \frac{1}{S}\sum_{j=1}^{S}\left(\hat{\theta}_j - \theta\right) \\ var\left(\hat{\theta}\right) &= \frac{1}{S-1}\sum_{j=1}^{S}\left(\hat{\theta}_j - \bar{\hat{\theta}}_j\right)^2, \bar{\hat{\theta}} = \frac{1}{S}\sum_{j=1}^{S}\hat{\theta}_j \\ rmse\left(\hat{\theta}\right) &= \sqrt{bias^2(\theta_j) + var(\hat{\theta}_j)} \\ \theta &= \sum_{k=1}^{N} y_k. \end{aligned}$$

Relative performance. We compare each weighting adjustment estimator that utilizes benchmark samples to the estimators that do not use benchmark samples.

Table 3.3: Selection probabilities model

	Sample selection variables	
	(internet_use)	(no_coverage)
region[2]	-0.107^{*}	-0.044
region[3]	-0.133^{**}	0.392^{***}
region[4]	0.029	0.305^{***}
sex[2]	0.097^{***}	0.022
race[2]	-0.377^{***}	0.185^{***}
race[3]	-0.786^{***}	0.722^{***}
race[4]	-0.299^{***}	-0.041
employed[1]	0.492^{***}	0.305^{***}
agegrp[2]	-0.641^{***}	0.093^{*}
agegrp[3]	-1.248^{***}	0.008
agegrp[4]	-1.743^{***}	-0.205^{***}
agegrp[5]	-2.047^{***}	-1.152^{***}
agegrp[6]	-2.587^{***}	-2.800^{***}
agegrp[7]	-3.478^{***}	-2.676^{***}
faminc_q[1]	0.478^{***}	-0.271^{***}
faminc_q[2]	0.890^{***}	-1.062^{***}
faminc_q[3]	1.447^{***}	-2.221^{***}
marst[2]	0.043	-0.148^{***}
marst[3]	0.149^{***}	-0.286^{***}
educ2[1]	0.766^{***}	-0.234^{***}
educ2[2]	1.557^{***}	-0.538^{***}
educ2[3]	1.812^{***}	-0.602^{***}
educ2[4]	2.334^{***}	-1.089^{***}
cancer[1]	-0.081	-0.197^{**}
wrk_private[1]	-0.005	0.093^{**}
sathc[1]	-0.040	2.059^{***}
internet_health[1]	2.869^{***}	-0.069^{*}
Constant	-0.125	-1.183^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3.4: Outcome totals by selection probability quintile groups

Total uninsured	**average π_i**
1,183	0.173
1,640	0.581
1,053	0.889
1,038	0.979
518	0.995

Table 3.5: Simulation parameter summary

Population N	31,914
Population T	5,432
Number of samples	1,000
Sample n	250, 500, 1000
Benchmark n	250, 500, 1000, 2000, 4000, 8000, 16,000

Estimator		**working model**	**variables**
GREG1		Demographics1	Region, Age, Gender, Race
GREG2		Demographics2	Region, Age, Gender, Race, Education
ECGREG	min n = 250	Trim	Age, Gender, Race, Age, Age65, Education, Income, Seen health professional, Cancer status,
PSCORE1	min n = 500	Partial	Age, Gender, Race, Age, Age65, Education, Income, Seen health professional, Cancer status, Age65:Gender, Age65:Race
	min n = 1,000	Full	Age, Gender, Race, Age, Age65, Education, Income, Seen health professional, Cancer status, Age65:Gender, Age65:Race, Income:Race
PSCORE2		Internet	Region, Age, Gender Race, Education Employment, Income, Marital status, Cancer status, Work private, Seen health professional, Looked up health information online
ECLASSO		Full	Age, Gender, Race, Age, Age65, Education, Income, Seen health professional, Cancer status, Age65:Gender, Age65:Race, Income:Race

We calculate percent-root-mean-square-error:

$$\%relrmse = 100 \frac{rmse(\hat{\theta}^{bench})}{rmse(\hat{\theta}^{nobench})}$$

$$bench = PSCORE1, PSCORE2, ECGREG, ECLASSO$$

$$nobench = HT, GREG1, GREG2$$

Variance estimates. We evaluate the linearized variance estimates and bootstrap variance estimates by their 95% nominal coverage and %bias relative to empirical variance. We use normal approximation to generate confidence intervals. Letting s index simulation number, we construct a confidence interval for each method, $v = v^{ECLASSO}, v_g^{ECLASSO}, v_{boot}^{ECLASSO}$:

$$CI_s = \hat{T}_{ys}^{ECLASSO} \pm 1.96\sqrt{v_s}$$

$$I_s = \begin{cases} 1, & \text{if } T \in CI_s \\ 0, & \text{otherwise} \end{cases}$$

where I_s is the indicator whether the confidence interval covered the true population total. The nominal 95% coverage is: $100 \sum_{s=1}^{S} I_s/S$. To calculate %bias:

$$\%bias = 100 \left[v - var\left(\hat{T}_y^{ECLASSO}\right)\right] / var\left(\hat{T}_y^{ECLASSO}\right)$$

where $var\left(\hat{T}_y^{ECLASSO}\right)$ is the empirical variance obtained from 1,000 simulation samples. To understand how bootstrap variance estimates work with re-samplings of benchmark samples, we also generate $v_{boot}^{PSCORE1}$, $v_{boot}^{PSCORE2}$, v_{boot}^{ECGREG} and evaluate their 95% nominal coverage. We ignore the finite-population-correction factor in variance calculation, because the minimum sample size between the analytical and benchmark sample is at most 1,000, which is only roughly 3% of the population size.

3.6 Simulation results

3.6.1 Point estimates

The simulation results are based on 1,000 simulation samples. Table 3.6 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. As expected, assuming SRS without weighting adjustment, HT underestimates the true population total. Without a key calibration variable, GREG1 actually performed worse than HT with a sizable bias. When an important control variable is included, GREG2 has one of the smallest bias and RMSE among the estimators that are calculated in the simulation. This demonstrates that if population-level education control totals are not available, we would not have been able to estimate the population totals correctly without a smaller benchmark sample. In fact, we could have generated estimates that performed much worse than an estimator without weighting adjustment.

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, although the bias improves as benchmark size increases, when sample size increase, the bias gets worse. 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.

ECGREG and PSCORE1 use the same working models for all experimental groups. By comparing ECGREG and PSCORE1, there is evidence that calibration-based method is less sensitive to smaller external data in terms of variance. For benchmark sample size less 1,000, ECGREG has smaller variances than PSCORE1. However, PSCORE1 consistently has lower bias than ECGREG. One might attempt to combine both methods – generate propensity-score weights, then use the reference sample as the benchmark sample for calibration. This method is called propensity-score-poststratification, described in detail by (*Valliant and Dever*, 2011). However, (*Valliant and Dever*, 2011) found propensity-score-poststratification to be the most unreliable weighting adjustment in their simulation.

Table 3.7 lists the percentage of times each variable is selected by LASSO across 1,000 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. ECLASSO likely gains efficiency by setting these age categories to 0. As expected, the interaction terms do not seem to be highly important, thus ECLASSO further gains efficiencies over ECGREG under the full model.

Figure 3.1 provides a visual display of relative RMSE of each estimator that uses a benchmark sample to an estimator that does not. The blue colors mean better RMSE while red colors indicate worse 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 500 or larger. Thus ECLASSO is able to achieve the same performance as a calibration estimator controlled to a strong population-level variable, even with small benchmark samples. The same statement cannot be made for all other estimators that utilize a benchmark sample. When the key control variable is missing in population, traditional calibration can perform poorly, as in the case of GREG1. Nearly all weighting adjustment methods outperformed GREG1. 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. For the weighting adjustment method that uses the correct working model (same model that generated the samples), PSCORE2 was not able to remove sample bias completely. This suggests that the variables used in sample generation may not fully explain a person's tendency without health insurance.

3.6.2 Variance estimates

Table 3.8 lists the 95% nominal coverages and %bias for the asymptotic linearized variance estimates and naive bootstrap estimates of the ECLASSO estimator. The g-weighted variance estimate, v_g^{ECLASS}, have smaller bias relative to the unweighted variance estimate, v^{ECLASS}, although both methods consistently produce negative biases. The linearized variance estimates tend to have better bias and coverage when both analytical and benchmark sizes are small. This can be a side-effect of adding two biased estimates together. The bootstrap variance estimate, $v_{boot}^{ECLASSO}$, significantly

Table 3.6: Simulation summary, target is number of uninsured in the NHIS sample "population": $T = 5,432$

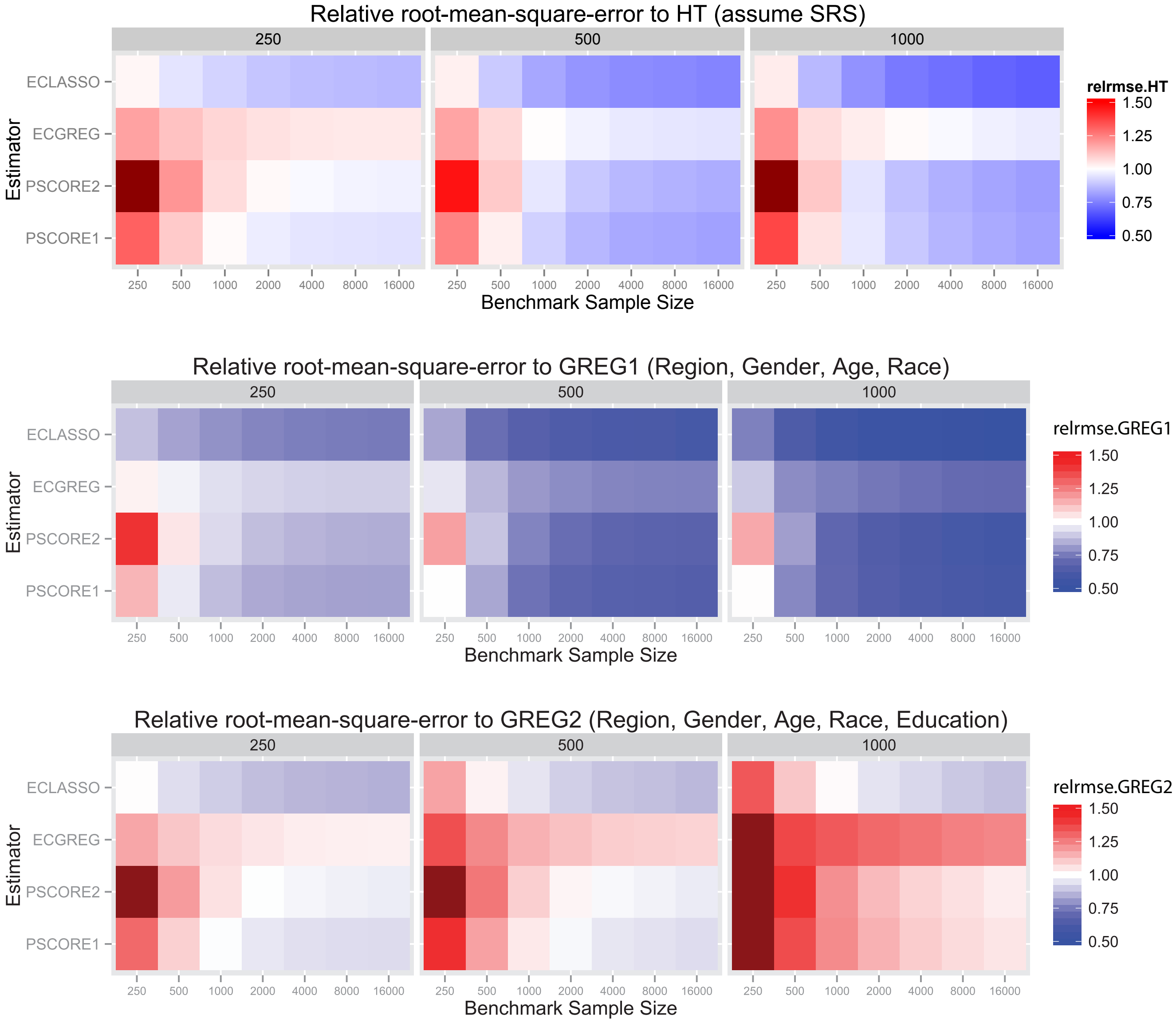
	HT			GREG1			GREG2		
sample n	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**
250	-383	539,696	828	-622	520,911	953	18	701,045	837
500	-378	270,501	643	-622	248,463	797	6	316,044	562
1,000	-355	137,074	513	-602	120,871	695	25	159,396	400

		PSCORE1			PSCORE2			ECGREG			ECLASSO		
sample n	**benchmark n**	**bis**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**	**bias**	**var**	**rmse**
250	250	260	1,107,072	1,084	442	1,607,227	1,343	344	840,761	979	20	707,219	841
250	500	165	805,848	913	192	960,659	999	348	735,850	926	28	614,884	785
250	1,000	118	683,108	835	109	769,714	884	343	681,759	894	28	573,320	758
250	2,000	108	627,301	799	86	691,357	836	345	649,934	877	29	540,198	736
250	4,000	90	612,221	788	62	667,272	819	337	637,954	867	19	524,209	724
250	8,000	90	604,779	783	59	652,658	810	337	632,807	864	18	516,230	719
250	16,000	93	601,555	781	59	647,378	807	339	628,826	862	19	509,372	714
500	250	258	571,430	799	365	753,605	942	328	465,819	757	-5	437,480	661
500	500	160	413,761	663	189	466,759	709	293	390,887	690	4	333,024	577
500	1,000	104	332,284	586	116	362,946	614	276	338,838	644	-3	283,775	533
500	2,000	94	296,795	553	99	320,949	575	281	314,649	628	0	260,722	511
500	4,000	79	280,539	535	82	300,933	555	274	304,088	616	-10	248,869	499
500	8,000	75	276,327	531	76	293,225	547	271	302,006	613	-12	245,002	495
500	16,000	74	270,378	525	74	286,701	541	272	298,008	610	-14	238,418	488
1,000	250	318	386,304	698	409	487,080	809	320	287,015	624	-17	282,471	532
1,000	500	236	237,164	541	256	255,922	567	289	215,426	547	-3	195,902	443
1,000	1,000	215	193,911	490	202	195,189	486	296	194,099	531	-9	163,316	404
1,000	2,000	209	166,298	458	194	166,493	452	304	175,362	517	-2	142,976	378
1,000	4,000	193	155,752	439	180	155,601	433	299	167,887	507	-6	136,351	369
1,000	8,000	187	148,255	428	173	149,077	423	295	161,660	498	-10	128,290	358
1,000	16,000	186	141,857	420	171	142,618	415	295	156,559	494	-11	123,956	352

Table 3.7: 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%

Figure 3.1: Relative Mean Square Errors



The blue color indicates a better RMSE (under 1), while the red color indicates a worse RMSE (over 1). Dark red color indicates a relative RMSE greater than 1.5.

over-estimates the empirical variance when the benchmark sample is small. As both analytical and benchmark sample size increase, $v_{boot}^{ECLASSO}$ improves in both bias and coverage. It does appear, however, that bootstrap variance estimates do not perform well with small benchmark samples. When benchmark sample sizes are 250, 500, or 1,000, there is evidence of large positive bias for bootstrap variance estimates. Since negative bias is less desirable, we would recommend bootstrap variance estimates where possible.

Table 3.9 lists the 95% nominal coverage of bootstrap variance estimates for the estimators that utilize benchmark samples. With small benchmark samples, PSCORE1 and PSCORE2 have large variances, which produce over-conservative confidence intervals that result in higher than expected 95% coverage rates. As benchmark sample size increases, nominal coverage of PSCORE1 and PSCORE2 bootstrap variance estimates get closer to 95%. 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 expected nominal coverage, around 90-91%. 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 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%. At sample size 250, the Trimmed model may still be too complex for ECGREG, thus the bootstrap variance estimates have high nominal coverage rates despite non-trivial biases. Among the estimators that use benchmark samples, ECLASSO is the least sensitive to both sample and benchmark sizes. As ECLASSO estimator is nearly unbiased for all experimental designs, stable nominal coverage rates also indicate that the ECLASSO consistently produces lower variances.

Table 3.8: Variance estimates 95% nominal coverage and %bias

		coverage			**%bias**		
sample n	**benchmark n**	$v^{ECLASSO}$	$v_g^{ECLASSO}$	$v_{boot}^{ECLASSO}$	$v^{ECLASSO}$	$v_g^{ECLASSO}$	$v_{boot}^{ECLASSO}$
250	250	88.6%	89.9%	97.4%	-27.0%	-21.4%	38.2%
250	500	88.4%	90.4%	96.8%	-28.2%	-21.9%	28.5%
250	1,000	88.9%	90.4%	96.4%	-29.6%	-23.0%	20.6%
250	2,000	89.3%	90.6%	95.7%	-28.8%	-21.8%	19.0%
250	4,000	88.8%	90.1%	96.3%	-28.5%	-21.4%	18.1%
250	8,000	89.4%	90.8%	95.8%	-28.3%	-21.1%	17.7%
250	16,000	89.6%	91.1%	95.9%	-27.8%	-20.5%	18.1%
500	250	92.4%	93.0%	97.0%	-11.1%	-7.0%	23.6%
500	500	91.9%	92.9%	96.8%	-13.5%	-8.1%	21.1%
500	1,000	92.3%	93.5%	96.0%	-16.2%	-10.1%	17.3%
500	2,000	91.9%	93.0%	96.4%	-18.5%	-11.9%	14.0%
500	4,000	91.5%	93.2%	96.3%	-19.7%	-13.0%	12.5%
500	8,000	91.2%	92.6%	96.2%	-21.1%	-14.3%	10.7%
500	16,000	91.2%	92.5%	96.2%	-20.2%	-13.2%	12.0%
1,000	250	93.0%	93.1%	96.1%	-10.9%	-8.1%	22.9%
1,000	500	93.4%	94.0%	96.5%	-11.9%	-7.9%	21.8%
1,000	1,000	92.8%	93.5%	96.6%	-18.7%	-14.0%	11.5%
1,000	2,000	92.4%	93.4%	96.2%	-21.0%	-15.6%	8.2%
1,000	4,000	90.6%	91.8%	95.8%	-24.4%	-18.9%	2.8%
1,000	8,000	91.3%	92.2%	95.6%	-23.6%	-17.7%	3.9%
1,000	16,000	92.1%	93.2%	95.9%	-22.9%	-16.9%	4.6%

Table 3.9: Bootstrap variance estimates and 95% nominal coverage

		coverage			
sample n	**benchmark n**	$v_{boot}^{PSCORE1}$	$v_{boot}^{PSCORE2}$	v_{boot}^{ECGREG}	$v_{boot}^{ECLASSO}$
250	250	99.0%	99.1%	97.1%	97.4%
250	500	98.5%	98.5%	96.5%	96.8%
250	1,000	97.6%	98.4%	96.9%	96.4%
250	2,000	97.3%	97.6%	96.4%	95.7%
250	4,000	97.2%	97.3%	97.2%	96.3%
250	8,000	96.7%	97.3%	96.7%	95.8%
250	16,000	96.8%	97.0%	96.7%	95.9%
500	250	98.9%	99.0%	96.7%	97.0%
500	500	98.4%	95.8%	91.3%	96.8%
500	1,000	97.1%	98.1%	90.3%	96.0%
500	2,000	97.4%	97.9%	90.8%	96.4%
500	4,000	97.1%	97.9%	91.0%	96.3%
500	8,000	97.2%	97.6%	91.2%	96.2%
500	16,000	97.0%	97.6%	91.2%	96.2%
1,000	250	98.7%	98.9%	95.9%	96.1%
1,000	500	98.3%	98.7%	96.4%	96.5%
1,000	1,000	98.2%	98.2%	97.1%	96.6%
1,000	2,000	97.2%	97.5%	97.1%	96.2%
1,000	4,000	96.6%	96.8%	96.9%	95.8%
1,000	8,000	96.9%	97.1%	97.1%	95.6%
1,000	16,000	96.6%	97.3%	97.1%	95.9%

3.7 Conclusion

In this chapter, we developed the full theoretical framework for ECLASSO calibration. We derived the point estimate formula under chi-square-distance measure, and proved that ECLASSO estimator of total is asymptotically design unbiased under the assumptions of our theoretical framework. Asymptotic linearized variance estimates are derived. We evaluate ECLASSO estimator of total and asymptotic linearized estimates through a simulation with an actual data. In terms of bias and RMSE, ECLASSO estimator uniformly outperforms traditional weighting adjustment methods that utilize the same benchmark data. The only estimator that has comparable bias and RMSE as ECLASSO estimator is the traditional calibration estimator controlled to the key variable in the population – education. If education-level population totals were not available, no estimators reached similar bias and RMSE as ECLASSO's. For variance estimates, neither the asymptotic linearized variance estimates nor bootstrap variance estimates gave satisfactory results. Linearized variance estimates have large negative bias while bootstrap variance estimates have significant positive bias. Compared to other estimators that use benchmark samples, however, ECLASSO bootstrap variance nominal coverage rates are the most stable and consistent across different sample and benchmark sizes.

CHAPTER IV

Application to Online Political Poll

4.1 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 with telephone samples that have been the staple of probability-based polling (*Holbrook et al.*, 2007; *Kohut et al.*, 2012). Recent research has shown the potential use of non-probability samples to predict election outcomes. *Wang et al.* (2014) 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. Major polling agencies within New York Times, CBS, and NBC have also turned to the more cost-effective non-probability sampling to collect large samples of potential voters within a short time period.

Without a well-defined sampling frame, non-probability-based election polls can have extremely off-balanced sample composition relative to the general voting population. *Wang et al.* (2014) found, for example, that the Xbox sample consisted of 75%

age 18-44 and over 90% male, compared to 50% age 18-44 and less than 50% male in the 2008 presidential election exit polls. Yet by making post-survey adjustments to match Xbox sample characteristics to the 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 benchmark data for the 2012 presidential election where president Obama ran for re-election. For most elections, however, no such large-scale benchmark exists. 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; *Esmer and Pettersson*, 2007; *Abramowitz*, 2008; *Healy et al.*, 2010), post-survey adjustments only to basic demographics are unlikely to remove all bias in imbalanced non-probability samples. Although adjusting non-probability samples to small benchmark samples with relevant variables can significantly reduce the bias, the small benchmark sample size can also greatly increase the standard errors of weighted estimates. To date, no research has shown whether small benchmark samples can effectively adjust non-probability election polls for accurate and precise election forecasts. In this chapter, we apply Estimated Control LASSO calibration (ECLASSO) to non-probability internet-based U.S. election polling data to adjust the polling sample with a small probability-based benchmark data. We aim to answer two key research questions:

(A) Can a small probability-based sample adjust a large non-probability internet-based election polling data for accurate election forecasts?

(B) Can the forecasts be precise given the small benchmark size?

Research question (A) concerns the bias property of an estimator, while (B) relates to the variance property of an estimator. ECLASSO is a promising approach to answer the research questions for two major reasons:

(1) *Leveraging information from benchmark.* By performing variable selection and parameter estimation simultaneously, ECLASSO can optimally leverage information from benchmark sample to allow for more effective bias reduction.

(2) *Estimating totals of binary outcome variables.* For bipartisan elections, the election outcome of interest is binary, i.e. Democrat or Republican. ECLASSO can result in smaller variance over traditional calibration where an implicit miss-specified linear relationship between the outcome and control variables is assumed.

4.2 Outcome of interest

For majority of the U.S., the winning political party is fairly predictable. From 1992-2012, 18 states have voted consecutively for Democratic presidential candidates while 13 states voted consecutively for Republican candidates. For these 31 states, it is not difficult to predict the winning political party for state-level elections. A measure of polling success thus relies on the predicted proportions of votes among the political parties. In particular, voting spread, the difference in proportions of votes between major political parties, has become an indicator for contentiousness of the election as well as a measure of polling accuracy. **We apply ECLASSO to predict the voting spread in the U.S. 2014 midterm election. We wish to estimate the proportion of Democratic votes minus the proportion of Republican votes (S_{D-R}) for each governor and senate race**. There are totals of 36 governor elections in 36 states, and 36 senate elections in 35 states. Since the actual election results are published, we can compare the bias and root-mean-square error of ECLASSO with traditional weighting adjustment methods. We provide details for each weighting method in the following section.

4.3 Estimation

For the outcome variable, we define a binary indicator:

$$y_i = \begin{cases} 1 & \text{if respondent } i \text{ voted for a Democratic candidate} \\ 0 & \text{if respondent } i \text{ voted for a Republican candidate} \end{cases}$$

For the rest of the chapter, we refer to the internet-based polling data as the analytical sample, denoted by s_A. Let $s_{A(r)}$ be the sample of respondents in state r, the voting spread in state r, $S_{D-R(r)}$, can be estimated by:

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

where w_i is the weight for respondent i. 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). Each weighting method constructs a set of weights by adjusting the initial design weights $\mathbf{w}_0$ based on information obtained from an external source. In the analysis, $\hat{S}_{D-R(r)}$ is calculated with five different set of weights:

$$\begin{aligned}
\mathbf{w}^{UNWT} &: \text{unadjusted weight} = \underset{n_A \times 1}{\mathbf{1}} \\
\mathbf{w}^{STATEWT} &: \text{State-level calibrated weights} \\
\mathbf{w}^{PSCORE} &: \text{Propensity-score adjusted weights} \\
\mathbf{w}^{ECGREG} &: \text{ECGREG calibrated weights} \\
\mathbf{w}^{ECLASSO} &: \text{ECLASSO calibrated weights}
\end{aligned}$$

Denote the estimated spread of each method by: $\hat{S}^{method}_{D-R(r)}$

$$method = UNWT, STATEWT, PSCORE, ECGREG, ECLASSO$$

Calibration-based weighting adjustment (STATEWT, ECGREG, ECLASSO) refers to the external source as benchmark data, while propensity-score weighting adjustment (PSCORE) refers to the external source as reference samples. To simplify the terminology, we refer to the external source as ”benchmark samples” in all weighting adjustment methods.

STATEWT method adjusts to state-level demographic totals that are derived from a sample much larger than the analytical sample, and thus is considered as adjustments to true population quantities. PSCORE, ECGREG, and ECLASSO, on the other hand, use the same benchmark sample that is much smaller than the internet-based data, but shares many common variables with the analytical sample. Due to the small benchmark sample size, it is not practical to perform weighting adjustments within each state for PSCORE, ECGREG, and ECLASSO. Instead, we assume that at the national-level, people with similar characteristics have the same voting tendencies. The individual voting tendencies are modified by the type of state the individual lives in. Similar model assumptions were made in the multi-level regression model of the Xbox election analysis by *Wang et al.* (2014). The assumed relationship between voting tendencies with individual characteristics and state types formulate the working models behind PSCORE, ECGREG, and ECLASSO, in constructing weights. The following section describes the weight construction process for each method.

4.4 Weight construction

We assume that the benchmark sample s_B has design weights $\underset{n_B \times 1}{\mathbf{w}_B}$ based on probability sampling, where n_B is the benchmark sample size. Furthermore, the benchmark and analytical samples have a common set of variables. Since the analytical sample is a non-probability internet sample, the initial design weight in the analytical sample is $\mathbf{w}_0 = \underset{n_A \times 1}{\mathbf{1}}$, where n_A is the analytical sample size. For the rest of the chapter, let $\mathbf{X}_B$, $\mathbf{X}_A$ be the design matrices with the common set of variables in the benchmark and analytical samples, and $\mathbf{W}_B = diag(\mathbf{w}_B)$, $\mathbf{W}_0 = diag(\mathbf{w}_0)$ be the corresponding weight matrices. The design matrix is explicitly defined for STATEWT in this section, while variables of $\mathbf{X}_A$ and $\mathbf{X}_B$ in PSCORE, ECGREG, and ECLASSO are described in detail in section 4.7.1.

4.4.1 STATEWT

For **STATEWT**, the benchmark sample has state-level demographic totals. We assume that state-level totals are estimated from a large-scale survey much bigger than the analytical sample. The analytical sample is controlled to demographic totals by Age, Gender, Race, and Education within each state. We define the design matrix for the analytical sample A in state r:

$$\mathbf{X}_{A(r)} = \left[\underset{n_{A(r)} \times 1}{\mathbf{1}}, \mathbf{X}_{A(r)}^{Age}, \mathbf{X}_{A(r)}^{Gendr}, \mathbf{X}_{A(r)}^{Race}, \mathbf{X}_{A(r)}^{Educ} \right]$$

The sub-matrix $\mathbf{X}_{A(r)}^{Age}$ is a dummy matrix for different *Age* categories. For instance, for *Age* with categories "18-24", "25-29", "30-39", "40-49", "50-59", "60-75", "75+", $\mathbf{X}_{A(r)}^{Age}$ has 6 columns, each column is a vector of $(0,1)$ values indicating age group membership of the respondents, with group "18-24" held out as the reference group. We assume the working model:

$$E\left[y_{i(r)}\middle|\mathbf{x}_{Ai(r)}, \boldsymbol{\beta}_{(r)}\right] = \mathbf{x}_{Ai(r)}^T\boldsymbol{\beta}_{(r)}, \quad V\left(y_i\right) = 1$$

Estimates of $\boldsymbol{\beta}_{(r)}$ can be obtained through weighted least-square regression:

$$\hat{\boldsymbol{\beta}}_{(r)} = \left(\mathbf{X}_{A(r)}^T\mathbf{W}_{0(r)}\mathbf{X}_{A(r)}\right)^{-1}\mathbf{X}_{A(r)}^T\mathbf{W}_{0(r)}\mathbf{y}_{(r)}$$

From the sample, estimate population totals of the covariates in the working model:

$$\begin{aligned}\hat{\mathbf{T}}^{X_{A(r)}} &= \left[\mathbf{w}_0^T\underset{n_{A(r)}\times 1}{\mathbf{1}}, \mathbf{w}_0^T\mathbf{X}_{A(r)}^{Age}, \mathbf{w}_0^T\mathbf{X}_{A(r)}^{Gender}, \mathbf{w}_0^T\mathbf{X}_{A(r)}^{Race}, \mathbf{w}_0^T\mathbf{X}_{A(r)}^{Educ}\right]\\ &= \left[n_{A(r)}, \hat{\mathbf{T}}_{A(r)}^{Age}, \hat{\mathbf{T}}_{A(r)}^{Gender}, \hat{\mathbf{T}}_{A(r)}^{Race}, \hat{\mathbf{T}}_{A(r)}^{Educ}\right]\end{aligned}$$

Given the state-level totals in benchmark:

$$\begin{aligned}\mathbf{T}^{X_{B(r)}} &= \left[\mathbf{w}_B^T\underset{n_{B(r)}\times 1}{\mathbf{1}}, \mathbf{w}_B^T\mathbf{X}_{B(r)}^{Age}, \mathbf{w}_B^T\mathbf{X}_{B(r)}^{Gender}, \mathbf{w}_B^T\mathbf{X}_{B(r)}^{Race}, \mathbf{w}_B^T\mathbf{X}_{B(r)}^{Educ}\right]\\ &= \left[N_{B(r)}, \mathbf{T}_{B(r)}^{Age}, \mathbf{T}_{B(r)}^{Gender}, \mathbf{T}_{B(r)}^{Race}, \mathbf{T}_{B(r)}^{Educ}\right]\end{aligned}$$

the calibrated weights for state r:

$$\begin{aligned}\mathbf{w}_{(r)}^{STATEWT} &= \mathbf{W}_{0(r)}\left(\underset{n_{A(r)}\times 1}{\mathbf{1}} + \mathbf{X}_{A(r)}\left(\mathbf{X}_{A(r)}^T\mathbf{W}_{0(r)}\mathbf{X}_{A(r)}\right)^{-1}\left(\mathbf{T}^{X_{B(r)}} - \hat{\mathbf{T}}^{X_{A(r)}}\right)^T\right)\\ &= \mathbf{w}_{0(r)} + \mathbf{W}_{0(r)}\mathbf{X}_{A(r)}\left(\mathbf{X}_{A(r)}^T\mathbf{W}_{0(r)}\mathbf{X}_{A(r)}\right)^{-1}\left(\mathbf{T}^{X_{B(r)}} - \hat{\mathbf{T}}^{X_{A(r)}}\right)^T\end{aligned} \tag{4.4.1.1}$$

The model parameters $\hat{\boldsymbol{\beta}}$ relate to the calibrated weights through estimation:

$$\begin{aligned}\hat{T}^{y}_{(r)} &= \left(\mathbf{w}^{STATEWT}_{(r)}\right)^{T}\mathbf{y}_{(r)} \\ &= \mathbf{w}^{T}_{0(r)}\mathbf{y}_{(r)} + \left(\mathbf{T}^{X_{B(r)}} - \hat{\mathbf{T}}^{X_{A(r)}}\right)\left(\mathbf{X}^{T}_{A(r)}\mathbf{W}_{0(r)}\mathbf{X}_{A(r)}\right)^{-1}\mathbf{X}^{T}_{A(r)}\mathbf{W}_{0(r)}\mathbf{y}_{(r)} \\ &= \hat{T}^{y.HT}_{(r)} + \left(\mathbf{T}^{X_{B(r)}} - \hat{\mathbf{T}}^{X_{A(r)}}\right)\hat{\boldsymbol{\beta}}_{(r)}\end{aligned}$$

where $\hat{T}^{y.HT}_{(r)}$ is the design unbiased Horvitz-Thompson estimator of total in state r. In STATEWT weighting adjustment method, we assume the voting tendencies are linearly related to Age, Gender, Race, and Education. The regression slopes are different across different states. If the linear relationship holds, weighted estimates based on $\mathbf{w}^{STATEWT}_{(r)}$ can have smaller variance relative to the variance of design-based estimator $\hat{T}^{y.HT}_{(r)}$.

4.4.2 PSCORE

For **PSCORE**, we combine the analytical and benchmark samples and estimate the probability of a respondent being in the analytical sample to generate pseudo-selection weights. The combined sample has design matrix: $\begin{pmatrix}\mathbf{X}_B \\ \mathbf{X}_A\end{pmatrix}$, and weights: $\begin{pmatrix}\mathbf{W}_B & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_0\end{pmatrix}$. Define the outcome variable:

$$z_i = \begin{cases} 1 & \text{if respondent } i \text{ is in the analytical sample} \\ 0 & \text{if respondent } i \text{ is in the benchmark sample} \end{cases}$$

We estimate the probability of a respondent being in the non-probability sample through logistic regression:

$$\begin{aligned} E\left[z_i \middle| \mathbf{x}_i, \boldsymbol{\beta}\right] &= \pi_i \\ log\left(\pi_i/(1-\pi_i)\right) &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

We compute estimates of π_i by solving the weighted logistic regression score equations:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{argmin}\left(\mathbf{X}^T \mathbf{W}\left(\mathbf{z} - expit\left(\mathbf{X}\boldsymbol{\beta}\right)\right)\right) \\ \hat{\pi}_i &= expit\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\right) \\ expit(u) &= 1/\left(1+e^{-u}\right) \end{aligned}$$

The adjusted weights based on PSCORE:

$$\mathbf{w}^{PSCORE} = \left(1/\hat{\pi}_i\right) = \left(1/expit\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\right)\right), i \in s_A \tag{4.4.2.1}$$

In PSCORE weighting adjustment method, we assume the probability of observing a respondent in the analytical sample is fully explained by the covariates in the design matrix $\mathbf{X}$. The inverse of the estimated probabilities serve as a non-response adjustment to correct the analytical sample to the target population. Because the initial design weights of the analytical sample are $\underset{n_A \times 1}{\mathbf{1}}$, the individuals in the analytical sample with the same values in the design matrix, i.e. $\mathbf{x}_{Ai} = \mathbf{x}_{Aj}$, receive the same weights.

4.4.3 ECGREG

For **ECGREG**, assume the working model:

$$E\left[y_i \middle| \mathbf{x}_{Ai}, \boldsymbol{\beta}\right] = \mathbf{x}_{Ai}^T \boldsymbol{\beta}, \quad V\left(y_i\right) = 1$$

The calibrated weights based on ECGREG:

$$\mathbf{w}^{ECGREG} = \mathbf{w}_0 + \mathbf{W}_0 \mathbf{X}_A \left(\mathbf{X}_A^T \mathbf{W}_0 \mathbf{X}_A\right)^{-1} \left(\hat{\mathbf{T}}^{X_B} - \hat{\mathbf{T}}^{X_A}\right)^T \tag{4.4.3.1}$$

where

$$\begin{aligned} \hat{\mathbf{T}}^{X_A} &= \left[\mathbf{w}_0^T \underset{n_A \times 1}{\mathbf{1}}, \mathbf{w}_0^T \mathbf{X}_A\right] \\ \hat{\mathbf{T}}^{X_B} &= \left[\mathbf{w}_B^T \underset{n_B \times 1}{\mathbf{1}}, \mathbf{w}_B^T \mathbf{X}_B\right] \end{aligned}$$

ECGREG is conceptually equivalent to calibrated STATEWT, except we replace known population totals $\mathbf{T}^{X_B}$ by estimates based on benchmark sample, $\hat{\mathbf{T}}^{X_B}$. In ECGREG weighting adjustments, we assume the outcome variable y_i is linearly related to the covariates in $\mathbf{X}_A$. Because the initial design weights of the analytical sample are $\underset{n_A \times 1}{\mathbf{1}}$, individuals with the same covariate values have the same weights.

4.4.4 ECLASSO

For **ECLASSO**, assume the working model:

$$E\left[y_i \middle| \mathbf{x}_{Ai}, \boldsymbol{\beta}\right] = \mu_i, \quad V\left(y_i\right) = \mu_i(1 - \mu_i)$$

We estimate μ_i by finding $\hat{\boldsymbol{\beta}}$ that minimizes the weighted penalized log-likelihood:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{argmin} \left(\sum_{i \in s_A} w_{0i} \left[-y_i \left(\mathbf{x}_i^T \beta \right) + log \left(1 + exp \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) \right) \right] + \lambda \sum_{j=1}^{p} \alpha_j^{\gamma} \left| \beta_j \right| \right) \\
\hat{\mu}_i &= expit \left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) \\
expit(u) &= 1 / \left(1 + e^{-u} \right)
\end{aligned}$$

where p is the number of parameters (excluding intercept) in design matrix $\mathbf{X}_A$, λ is the tuning parameter, and α_j is the weight parameter for a regression coefficient, and γ is a constant greater than 0. In this analysis, $\alpha_j = 1 / \left| \hat{\beta}_j^{glm} \right|$, where $\hat{\beta}_j^{glm}$ is the traditional logistic regression slope. We obtain λ and γ through cross-validations that minimize the area under curve of $\hat{\mu}_i$ in s_A. Initial values of λ can be obtained by fitting each regressor separately through generalized linear models, then find a grid of values (usually 100) between the minimum and maximum of the inverse of bi-variate regression coefficients. For γ, we cross-validate between 4 different values: 0.1, 0.5, 1, and 2. Small value of γ "flattens" α_j to 1, which is likely when all $\hat{\beta}_j^{glm}$ have similar magnitudes. For each λ and γ, $\hat{\boldsymbol{\beta}}$ is obtained through R *glmnet* package (*Friedman et al.*, 2010), version 1.9-8.

Once $\hat{\boldsymbol{\beta}}$ is obtained, we calculate μ_i for both the analytical and benchmark samples. Define:

$$\begin{aligned}
\mathbf{M}_A &= \left[\underset{n_A \times 1}{\mathbf{1}}, (\hat{\mu}_i)_{i \in s_A} \right] \\
\mathbf{M}_B &= \left[\underset{n_B \times 1}{\mathbf{1}}, (\hat{\mu}_i)_{i \in s_B} \right]
\end{aligned}$$

The calibrated weights based on ECLASSO:

$$\begin{aligned}\mathbf{w}^{ECLASSO} &= \mathbf{W}_0\left(\underset{n_A\times 1}{\mathbf{1}} + \mathbf{M}_A\left(\mathbf{M}_A^T\mathbf{W}_0\mathbf{M}_A\right)^{-1}\left(\mathbf{w}_B^T\mathbf{M}_B - \mathbf{w}_0^T\mathbf{M}_A\right)^T\right)\\ &= \mathbf{w}_0 + \mathbf{W}_0\mathbf{M}_A\left(\mathbf{M}_A^T\mathbf{W}_0\mathbf{M}_A\right)^{-1}\left(\mathbf{w}_B^T\mathbf{M}_B - \mathbf{w}_0^T\mathbf{M}_A\right)^T \qquad (4.4.4.1)\end{aligned}$$

In ECLASSO weighting adjustments, we assume the expected value of the outcome variable y_i is closely approximated by $\hat{\mu}_i$, given the data $\mathbf{X}_A$. Because the initial design weights of the analytical sample is $\underset{n_A\times 1}{\mathbf{1}}$, individuals in the analytical sample with the same $\hat{\mu}_i$ receive the same weights.

4.5 Variance estimates

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

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

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, method = PSCORE, EC-GREG, and ECLASSO, we use bootstrap variance estimate 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 $\hat{S}^{method}_{D-R(r)}$.

4.6 Data description

4.6.1 Election polling data

The online polling data is a random pull 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 end-page invitation to complete the online poll. Approximately 2-3% of respondents receiving the invitation completed the poll each day (roughly 6,000 per day). Although the sample was randomly selected among the survey takers, the pool of respondents who completed an initial SurveyMonkey survey is non-probability-based. The election polling data is thus considered a non-probability internet sample. The data was collected between October 3^{rd} and November 4^{th}, 2014 (the election day). A total of 168,924 responses were collected from 50 states and the District of Columbia, with 153,783 responses from states holding either a governor or a senator race.

4.6.2 Analytical sample

Experienced pollsters would agree that accurate election forecasts are based on responses of those who turn out to vote (*Perry*, 1973, 1979). Conditional on likely voters, there is a better chance to predict correct election winners (*Bolstein*, 1991; *Delavande and Manski*, 2010; *Gutsche et al.*, 2014). As part of the SurveyMonkey 2014 midterm election polling survey, respondents assess their voting chance in 7 categories: (1) already voted, (2) absolutely certain will vote, (3) very likely will vote, (4) 50-50 chance will vote, (5) less than 50-50, (6) do not plan to vote, and (7) other. We identify the likely voters as those who are in categories (1), (2), or (3). The analytical sample consists of likely voters, with a total of 85,668 respondents for 36 governor races, and 85,352 respondents for 36 senate races.

4.6.3 Benchmark sample

During September and October of 2014, Pew Research Center (`http://www.pewresearch.org`) selected probability samples of telephone and cellphone users 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 midterm elections. The survey also includes religion and political party identification along with other demographic variables that are also collected in the SurveyMonkey sample.

An attractive feature of PEW political data is the availability of likely voter weights for the 2014 midterm election. About half of the sample are identified as likely voters based on a 10-point scale voting interest variable. Since we are calibrating likely voters of the analytical sample, September/October 2014 PEW political survey is an ideal candidate benchmark sample that offers many possibilities for constructing assisting models used in estimation. There are totals of 3,047 PEW September/October likely voters for the governor race, and 2,244 likely voters for the senate race across all states. The PEW sample is used as benchmark sample for PSCORE, ECGREG, and ECLASSO weighting adjustments.

4.6.4 Final sample

Some states participating in the 2014 midterm election have small benchmark sample sizes, e.g. Hawaii $n = 9$, Wyoming $n = 12$. Calibration estimates of state-level voting spreads may be unreliable for states with few representatives in the benchmark (*Särndal*, 2007; *Dever*, 2008). Thus we narrow our analysis to states with sufficient benchmark sizes. For governor elections, we estimate voting spreads for 11 states that have at least 60 likely voters in the benchmark sample. For senate elections, we narrow the analysis to 8 states that have at least 55 likely voters in the benchmark sample. **The final benchmark sample sizes are 1,094 for governor race and 656 for senate race**. Since this chapter focuses on binary outcomes, we further narrow the analytical sample of the 11 governor states and 8 senate states to the likely voters who indicated a vote for either a Democratic or Republican candidate. **The final analytical sample sizes are 33,199 for the governor race and 28,686 for the senate race**.

4.7 Variables and working models

4.7.1 Variables

We seek variables in both analytical and benchmark samples on basic demographics, religious beliefs, political attitudes, and party identification. We assume that voting tendencies are similar for individuals with similar characteristics on these variables. In the U.S. elections, there are clear electoral divisions across the states. For example, California and New York historically favor the Democratic Party, while Arizona and Texas favor the Republican Party. There can be strong regional effects that modify individual vote intentions. Thus we group states by their electoral results of the past 4 presidential elections (2000-2012). State type: (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. This definition of state type is consistent with blue/purple/red state definition that has been associated with electoral divisions (*Wing and Walker*, 2010; *Levendusky and Pope*, 2011). For working models in the weighting adjustments, we assume that individual vote intentions are modified by state-type regional effects.

Table 4.1 lists the distributions of demographic and political variables by state and state type for the working models of the governor race. Table 4.2 lists the distributions of demographic and political variables by state and state type for the working models of the senate race. The analytical sample distributions are unweighted, while the benchmark sample distributions are weighted by the likely voter weights. The senate race has one more variable than the governor race - support for House. Since both House of Representatives and Senate are part of Congress, the variable is more relevant for senate elections. Variables with "Don't Know" (DK), "Refused" (RF), or missing values have a separate category to indicate unobserved measurement. Keeping the cases with unobserved values can maintain the sample size as well as capture the differences in missing patterns between analytical and benchmark samples. The internet-based analytical sample tends to be younger, more educated, none-minority, 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. The Democrat and Republican outcome in the analytical sample account for over 90% of the support, except for Florida and New York in the governor race, which have 90% and 86% majority party support respectively. As both Florida and New York are grouped with other states in the working models, the impact of removing non-major party outcome from the analytical sample should be small in the analysis.

4.7.2 Working models

In this section, we detail the variables for the design matrices of section 4.4. The design matrices consist of main effects based on state-type, the variables in tables 4.1 and 4.2, and a set of interactions between the main effects. Note that all variables in the analysis are categorical. Define:

$$\begin{aligned}
\mathbf{X}_{main} &= \left[\mathbf{X}^{Age}, \mathbf{X}^{Gender}, \mathbf{X}^{Race}, \mathbf{X}^{Educ}, \mathbf{X}^{Relig}, \mathbf{X}^{Attend}, \mathbf{X}^{Born}, \mathbf{X}^{Approval}, \mathbf{X}^{Party}, \mathbf{X}^{StateType}\right] \\
\mathbf{X}_{house} &= \left[\mathbf{X}^{House}\right] \\
\mathbf{X}_{interaction} &= \left[\mathbf{X}^{Gender:Age}, \mathbf{X}^{Gender:Race}, \mathbf{X}^{Race:Age}, \mathbf{X}^{Party:Approval}, \mathbf{X}^{StateType:Party}, \mathbf{X}^{StateType:Approval}\right]
\end{aligned}$$

The operator ':' denotes interaction between two variables. Each variable in the design matrix is a dummy matrix with columns of value (0,1) indicating a respondent's membership to one of the categories of the variable. One reference group of each variable corresponding to the first category listed in tables 4.1 and 4.2 is held out of the design matrix for that variable. To distinguish analytical and benchmark design

Table 4.1: Governor election covariates and outcome variables, by sample type

State Type 1			**Age**						**Gender**		**Race**				**Education**				**Religion**					**Born again Evangelical Christian**			**Attend religion**				**Approve Obama**			**Party lean**			**Outcome (whole state)**	
State	Sample	n	18-29	30-39	40-49	50-59	60-75	75+	Male	Female	non-Hispanic white	non-Hispanic black	Hispanic	Other	High school or less	Some college	College graduate	Post-graduate	Protestant	Catholic	Other Christian	Other	DK/RF/Missing	Yes	No	DK/RF/Missing	More than once a week	Once a week	A few times a month	A few times a year/none	Approve	Disapprove	DK/RF/Missing	Republican	Democrat	None/Other	Democrat	Republican
AZ	Analytical	974	10%	11%	17%	26%	30%	6%	47%	53%	82%	2%	9%	7%	6%	31%	28%	35%	23%	19%	15%	17%	26%	16%	22%	62%	8%	23%	10%	59%	34%	58%	8%	27%	37%	36%	41%	54%
	Benchmark	64	2%	13%	22%	25%	30%	7%	53%	47%	65%	2%	25%	7%	23%	30%	26%	22%	20%	31%	16%	20%	12%	24%	48%	28%	15%	27%	9%	49%	38%	62%	0%	42%	24%	35%		
GA	Analytical	2,306	10%	13%	23%	27%	24%	3%	50%	50%	75%	16%	2%	8%	8%	23%	33%	37%	39%	10%	22%	10%	18%	35%	25%	40%	15%	26%	15%	44%	36%	58%	6%	31%	39%	30%	40%	53%
	Benchmark	67	11%	13%	29%	22%	20%	6%	55%	45%	80%	11%	0%	9%	23%	37%	21%	19%	64%	12%	11%	7%	6%	51%	37%	11%	23%	34%	18%	25%	30%	68%	2%	42%	22%	37%		
TX	Analytical	2,575	12%	13%	21%	28%	22%	3%	49%	51%	75%	8%	10%	7%	7%	29%	34%	30%	33%	19%	20%	11%	17%	31%	22%	47%	13%	29%	13%	46%	30%	61%	8%	25%	42%	33%	36%	61%
	Benchmark	150	11%	15%	23%	17%	25%	9%	51%	49%	65%	16%	13%	7%	27%	38%	21%	14%	53%	13%	12%	18%	5%	47%	32%	21%	23%	25%	15%	38%	34%	59%	7%	42%	29%	29%		
Overall	Analytical	5,855	11%	13%	21%	27%	24%	3%	49%	51%	76%	10%	7%	7%	7%	27%	33%	33%	34%	16%	20%	12%	19%	30%	23%	47%	13%	27%	13%	47%	33%	59%	7%	28%	40%	32%	38%	57%
	Benchmark	281	9%	14%	24%	20%	25%	8%	52%	48%	68%	12%	12%	7%	25%	36%	22%	17%	49%	17%	12%	16%	7%	43%	37%	20%	21%	27%	14%	37%	34%	62%	4%	42%	26%	32%		

State Type 3			**Age**						**Gender**		**Race**				**Education**				**Religion**					**Born again Evangelical Christian**			**Attend religion**				**Approve Obama**			**Party lean**			**Outcome (whole state)**	
FL	Analytical	2,566	10%	11%	18%	27%	29%	5%	51%	49%	79%	8%	8%	5%	8%	27%	31%	33%	26%	22%	19%	14%	19%	23%	22%	55%	11%	22%	11%	56%	37%	56%	7%	32%	39%	29%	42%	48%
	Benchmark	134	10%	11%	16%	20%	36%	8%	55%	45%	76%	10%	10%	4%	28%	43%	16%	14%	47%	19%	11%	17%	7%	37%	38%	26%	13%	21%	18%	49%	44%	55%	1%	42%	41%	18%		
OH	Analytical	2,299	14%	12%	20%	28%	24%	3%	50%	50%	90%	5%	1%	4%	11%	26%	33%	30%	30%	24%	18%	12%	16%	24%	24%	52%	10%	26%	12%	51%	29%	64%	7%	28%	40%	31%	31%	65%
	Benchmark	87	12%	14%	18%	24%	24%	8%	59%	41%	78%	13%	2%	6%	42%	25%	22%	11%	33%	27%	18%	22%	0%	32%	44%	24%	13%	30%	15%	42%	33%	63%	4%	34%	39%	27%		
Overall	Analytical	4,865	12%	11%	19%	28%	27%	4%	50%	50%	84%	7%	5%	5%	9%	27%	32%	32%	28%	23%	19%	13%	17%	24%	23%	54%	11%	24%	12%	53%	33%	60%	7%	30%	39%	30%	37%	56%
	Benchmark	221	11%	12%	17%	22%	31%	8%	57%	43%	77%	11%	7%	5%	33%	35%	18%	13%	41%	22%	14%	19%	4%	35%	40%	25%	13%	25%	17%	46%	40%	58%	2%	39%	40%	21%		

State Type 5			**Age**						**Gender**		**Race**				**Education**				**Religion**					**Born again Evangelical Christian**			**Attend religion**				**Approve Obama**			**Party lean**			**Outcome (whole state)**	
CA	Analytical	2,354	13%	11%	20%	27%	26%	4%	47%	53%	72%	5%	11%	13%	6%	25%	31%	37%	19%	21%	16%	21%	23%	16%	19%	65%	8%	19%	10%	63%	44%	48%	8%	40%	30%	31%	55%	42%
	Benchmark	166	10%	12%	13%	27%	25%	13%	50%	50%	64%	4%	20%	12%	21%	37%	23%	19%	25%	21%	14%	30%	10%	24%	39%	36%	8%	20%	13%	59%	50%	47%	2%	25%	39%	36%		
IL	Analytical	2,955	12%	12%	20%	27%	25%	3%	48%	52%	84%	6%	4%	6%	8%	26%	32%	34%	24%	31%	15%	14%	16%	16%	23%	61%	8%	24%	13%	55%	40%	56%	3%	34%	30%	35%	41%	55%
	Benchmark	78	3%	14%	21%	25%	19%	18%	53%	47%	86%	8%	4%	2%	27%	37%	21%	16%	36%	21%	10%	27%	5%	36%	31%	33%	11%	19%	14%	57%	53%	46%	1%	24%	47%	29%		
MI	Analytical	6,025	13%	12%	19%	26%	26%	3%	50%	50%	88%	5%	1%	6%	7%	26%	31%	35%	29%	25%	17%	15%	14%	21%	25%	55%	10%	25%	12%	53%	41%	56%	4%	30%	31%	38%	42%	56%
	Benchmark	75	6%	16%	18%	21%	29%	11%	51%	49%	77%	21%	1%	1%	25%	44%	12%	19%	28%	25%	19%	25%	4%	32%	38%	31%	13%	24%	13%	49%	48%	44%	8%	19%	42%	39%		
NY	Analytical	1,962	12%	11%	19%	28%	27%	3%	49%	51%	80%	7%	5%	7%	9%	23%	30%	38%	15%	36%	10%	15%	23%	9%	16%	74%	6%	21%	11%	62%	41%	51%	7%	41%	28%	31%	49%	37%
	Benchmark	106	16%	15%	12%	22%	25%	9%	43%	57%	69%	13%	9%	10%	25%	33%	23%	19%	28%	26%	7%	26%	12%	20%	43%	37%	10%	25%	13%	53%	53%	45%	2%	22%	46%	32%		
PA	Analytical	2,318	12%	11%	19%	29%	26%	3%	48%	52%	90%	4%	1%	4%	11%	25%	30%	35%	27%	27%	12%	13%	21%	15%	24%	61%	7%	25%	13%	55%	36%	53%	10%	37%	35%	28%	54%	43%
	Benchmark	107	9%	10%	14%	28%	27%	12%	48%	52%	84%	12%	2%	2%	43%	29%	12%	15%	40%	22%	13%	19%	7%	30%	43%	26%	13%	26%	15%	46%	38%	53%	9%	41%	38%	21%		
WI	Analytical	6,865	13%	14%	22%	28%	21%	2%	52%	48%	93%	2%	1%	4%	9%	28%	33%	30%	23%	30%	17%	17%	14%	15%	25%	60%	6%	25%	15%	54%	41%	54%	5%	31%	36%	33%	46%	52%
	Benchmark	60	8%	11%	28%	18%	27%	9%	54%	46%	92%	3%	0%	5%	31%	35%	20%	14%	46%	30%	9%	13%	1%	21%	64%	15%	2%	40%	26%	32%	43%	57%	1%	28%	38%	33%		
Overall	Analytical	22,479	13%	12%	20%	27%	24%	3%	50%	50%	87%	5%	3%	6%	8%	26%	32%	34%	24%	28%	15%	16%	17%	16%	23%	61%	8%	24%	13%	55%	41%	54%	5%	34%	32%	34%	46%	50%
	Benchmark	592	9%	13%	16%	25%	25%	12%	49%	51%	76%	10%	8%	6%	28%	35%	19%	18%	32%	24%	12%	24%	7%	27%	42%	31%	10%	24%	15%	51%	48%	48%	4%	27%	41%	32%		

Table 4.2: Senate election covariates and outcome variables, by sample type

State Type 1			Age						Gender		Race				Education				Religion					Born again Evangelical Christian			Attend religion				Approve Obama			Party lean			House of Repre-sentative support			Outcome (whole state)	
state	sample	n	18-29	30-39	40-49	50-59	60-75	75+	Male	Female	non-Hispanic white	non-Hispanic black	Hispanic	Other	High school or less	Some college	College graduate	Post-graduate	Protestant	Catholic	Other Christian	Other	DK/RF/Missing	Yes	No	DK/RF/Missing	More than once a week	Once a week	A few times a month	A few times a year/none	Approve	Disapprove	DK/RF/Missing	Republican	Democrat	None/Other	None/Other	Republican	Democrat	Democrat	Republican
GA	Analytical	3,020	12%	13%	23%	26%	23%	2%	49%	51%	74%	17%	2%	7%	8%	23%	33%	36%	31%	8%	18%	9%	33%	28%	20%	51%	14%	25%	14%	47%	35%	54%	11%	32%	39%	29%	6%	54%	40%	41%	53%
	Benchmark	67	11%	13%	29%	22%	20%	6%	55%	45%	80%	11%	0%	9%	23%	37%	21%	19%	64%	12%	11%	7%	6%	51%	37%	11%	23%	34%	18%	25%	30%	68%	2%	42%	22%	37%	10%	60%	30%		
MN	Analytical	3,774	14%	14%	20%	28%	22%	2%	52%	48%	92%	2%	1%	5%	7%	27%	35%	31%	24%	18%	15%	12%	31%	15%	23%	62%	7%	25%	15%	54%	40%	45%	15%	38%	30%	32%	8%	44%	48%	49%	48%
	Benchmark	57	6%	12%	18%	25%	21%	19%	50%	50%	83%	1%	4%	11%	34%	32%	18%	17%	43%	28%	5%	19%	5%	18%	54%	28%	9%	37%	11%	43%	34%	59%	7%	26%	25%	49%	23%	37%	40%		
TX	Analytical	3,273	14%	13%	21%	27%	22%	3%	49%	51%	74%	8%	11%	7%	8%	30%	33%	29%	26%	17%	17%	9%	31%	25%	18%	57%	12%	26%	12%	51%	28%	58%	14%	26%	43%	31%	6%	60%	34%	51%	46%
	Benchmark	150	11%	15%	23%	17%	25%	9%	51%	49%	65%	16%	13%	7%	27%	38%	21%	14%	53%	13%	12%	18%	5%	47%	32%	21%	23%	25%	15%	38%	34%	59%	7%	42%	29%	29%	13%	47%	40%		
Overall	Analytical	10,067	13%	14%	21%	27%	22%	3%	50%	50%	81%	8%	5%	6%	8%	27%	34%	32%	27%	15%	16%	10%	32%	22%	21%	57%	11%	25%	13%	51%	35%	52%	13%	32%	37%	31%	7%	52%	41%	47%	49%
	Benchmark	274	10%	14%	23%	20%	23%	10%	52%	48%	72%	12%	8%	8%	27%	37%	20%	16%	54%	15%	10%	15%	5%	43%	37%	20%	20%	29%	15%	36%	33%	61%	6%	39%	26%	35%	14%	48%	38%		

State Type 2			Age						Gender		Race				Education				Religion					Born again Evangelical Christian			Attend religion				Approve Obama			Party lean			House of Repre-sentative support			Outcome (whole state)	
NC	Analytical	8,010	12%	12%	20%	26%	27%	4%	53%	47%	83%	9%	2%	6%	8%	25%	33%	34%	32%	9%	17%	11%	31%	27%	22%	51%	14%	25%	12%	48%	39%	54%	7%	34%	35%	31%	5%	51%	44%	55%	42%
	Benchmark	90	6%	11%	19%	25%	29%	10%	49%	51%	74%	17%	5%	5%	27%	32%	21%	21%	60%	11%	13%	12%	4%	50%	36%	13%	16%	28%	10%	46%	29%	64%	7%	39%	26%	35%	12%	49%	39%		

State Type 3			Age						Gender		Race				Education				Religion					Born again Evangelical Christian			Attend religion				Approve Obama			Party lean			House of Repre-sentative support			Outcome (whole state)	
VA	Analytical	6,875	15%	12%	19%	26%	25%	3%	49%	51%	82%	8%	2%	8%	8%	22%	30%	41%	24%	11%	12%	11%	41%	16%	21%	64%	8%	21%	10%	61%	31%	43%	26%	33%	32%	35%	7%	48%	45%	48%	44%
	Benchmark	81	18%	12%	19%	18%	20%	13%	54%	46%	77%	16%	3%	3%	30%	26%	21%	24%	43%	12%	15%	25%	6%	32%	39%	29%	12%	29%	15%	43%	44%	51%	5%	33%	38%	29%	15%	40%	45%		

State Type 5			Age						Gender		Race				Education				Religion					Born again Evangelical Christian			Attend religion				Approve Obama			Party lean			House of Repre-sentative support			Outcome (whole state)	
IL	Analytical	4,061	14%	13%	20%	26%	24%	3%	49%	51%	83%	7%	4%	6%	9%	26%	31%	34%	18%	25%	12%	12%	33%	13%	18%	70%	8%	23%	12%	57%	42%	52%	6%	36%	29%	34%	7%	46%	46%	49%	48%
	Benchmark	78	3%	14%	21%	25%	19%	18%	53%	47%	86%	8%	4%	2%	27%	37%	21%	16%	36%	21%	10%	27%	5%	36%	31%	33%	11%	19%	14%	57%	53%	46%	1%	24%	47%	29%	4%	44%	51%		
MI	Analytical	7,820	15%	12%	19%	26%	25%	3%	51%	49%	87%	6%	1%	6%	8%	27%	31%	34%	22%	20%	14%	13%	31%	17%	20%	64%	9%	24%	12%	54%	41%	54%	5%	33%	32%	36%	7%	47%	46%	51%	46%
	Benchmark	75	6%	16%	18%	21%	29%	11%	51%	49%	77%	21%	1%	1%	25%	44%	12%	19%	28%	25%	19%	25%	4%	32%	38%	31%	13%	24%	13%	49%	48%	44%	8%	19%	42%	39%	16%	30%	54%		
NJ	Analytical	1,120	18%	10%	18%	27%	24%	3%	45%	55%	78%	8%	5%	9%	13%	22%	34%	31%	11%	32%	9%	11%	38%	7%	13%	80%	7%	19%	11%	64%	39%	48%	13%	38%	26%	35%	7%	44%	49%	55%	41%
	Benchmark	58	7%	12%	20%	35%	20%	6%	53%	47%	78%	11%	6%	5%	15%	31%	22%	33%	13%	48%	13%	15%	11%	17%	56%	27%	7%	31%	22%	39%	47%	50%	4%	34%	32%	34%	23%	40%	38%		
Overall	Analytical	13,001	15%	12%	19%	26%	25%	3%	50%	50%	85%	6%	3%	6%	9%	26%	31%	34%	20%	22%	13%	12%	32%	15%	18%	67%	9%	23%	12%	56%	41%	53%	6%	34%	31%	35%	32%	37%	31%	51%	46%
	Benchmark	211	5%	14%	20%	26%	23%	12%	52%	48%	80%	14%	4%	2%	23%	38%	18%	22%	27%	30%	14%	23%	6%	29%	40%	31%	11%	24%	16%	49%	49%	46%	4%	25%	41%	34%	39%	26%	35%		

matrices, we write:

$$\begin{aligned}
\mathbf{X}_{A.main} &= \text{design matrix of main effects in analytical sample} \\
\mathbf{X}_{A.house} &= \text{design matrix of House support variable in analytical sample} \\
\mathbf{X}_{A.interation} &= \text{design matrix of interaction effects in analytical sample}
\end{aligned}$$

$$\begin{aligned}
\mathbf{X}_{B.main} &= \text{design matrix of main effects in benchmark sample} \\
\mathbf{X}_{B.house} &= \text{design matrix of House support variable in benchmark sample} \\
\mathbf{X}_{B.interation} &= \text{design matrix of interaction effects in benchmark sample}
\end{aligned}$$

The full design matrices for governor election working models:

$$\begin{aligned}
\mathbf{X}_{A.governor} &= \left[\underset{n_A\times 1}{\mathbf{1}}, \mathbf{X}_{A.main}, \mathbf{X}_{A.interaction}\right] \\
\mathbf{X}_{B.governor} &= \left[\underset{n_B\times 1}{\mathbf{1}}, \mathbf{X}_{B.main}, \mathbf{X}_{B.interaction}\right]
\end{aligned}$$

The full design matrices for senate election working models:

$$\begin{aligned}
\mathbf{X}_{A.senate} &= \left[\underset{n_A\times 1}{\mathbf{1}}, \mathbf{X}_{A.main}, \mathbf{X}_{A.house}, \mathbf{X}_{A.interaction}\right] \\
\mathbf{X}_{B.senate} &= \left[\underset{n_B\times 1}{\mathbf{1}}, \mathbf{X}_{B.main}, \mathbf{X}_{B.house}, \mathbf{X}_{B.interaction}\right]
\end{aligned}$$

4.7.3 ECGREG adjusted weights

As an illustration, we revisit the weight construction for ECGREG and outline the steps for obtaining weights in equation 4.4.3.1. Similar process can be carried out for PSCORE and ECLASSO to obtain the adjusted weights for those methods. Recall that s_A is the analytical sample, with sample size n_A and initial design weights $\mathbf{w}_0 = \underset{n_A\times 1}{\mathbf{1}}$, and s_B is the benchmark sample, with sample size n_B and probability-based weights $\mathbf{w}_B$. For ECGREG, we assume for governor election:

$$
\begin{aligned}
E\left[y_i^{governor}|\mathbf{x}_i, \boldsymbol{\beta}\right] = \beta_0 &+ \beta_{k[i]}^{Age} + \beta_{k[i]}^{Gender} + \beta_{k[i]}^{Race} + \beta_{k[i]}^{Educ} + \\
&\beta_{k[i]}^{Relig} + \beta_{k[i]}^{Attend} + \beta_{k[i]}^{Born} + \beta_{k[i]}^{Approval} + \beta_{k[i]}^{Party} + \beta_{k[i]}^{StateType} + \\
&\beta_{k[i]}^{Gender:Age} + \beta_{k[i]}^{Gender:Race} + \beta_{k[i]}^{Race:Age} + \\
&\beta_{k[i]}^{Party:Approval} + \beta_{k[i]}^{StateType:Party} + \beta_{k[i]}^{StateType:Approval}
\end{aligned}
$$

where $k[i]$ is the category respondent i belongs to for a given variable. For example, model coefficient for respondents age 30-39 is denoted by $\beta_{(30-39)[i]}^{Age}$. Similarly, for senate election model:

$$
\begin{aligned}
E\left[y_i^{senate}|\mathbf{x}_i, \boldsymbol{\beta}\right] = \beta_0 &+ \beta_{k[i]}^{Age} + \beta_{k[i]}^{Gender} + \beta_{k[i]}^{Race} + \beta_{k[i]}^{Educ} + \\
&\beta_{k[i]}^{Relig} + \beta_{k[i]}^{Attend} + \beta_{k[i]}^{Born} + \beta_{k[i]}^{Approval} + \beta_{k[i]}^{Party} + \beta_{k[i]}^{StateType} + \beta_{k[i]}^{House} \\
&\beta_{k[i]}^{Gender:Age} + \beta_{k[i]}^{Gender:Race} + \beta_{k[i]}^{Race:Age} + \\
&\beta_{k[i]}^{Party:Approval} + \beta_{k[i]}^{StateType:Party} + \beta_{k[i]}^{StateType:Approval}
\end{aligned}
$$

The only difference between senate and governor working models is the inclusion of main effects $\beta_{k[i]}^{House}$ in the senate model. To construct weights under ECGREG, we first obtain estimates of population totals from the analytical and benchmark samples:

$$
\begin{aligned}
\hat{\mathbf{T}}^{X_{A.main}} &= \Big[\mathbf{w}_0^T\mathbf{X}_A^{Age}, \mathbf{w}_0^T\mathbf{X}_A^{Gender}, \mathbf{w}_0^T\mathbf{X}_A^{Educ}, \\
&\qquad \mathbf{w}_0^T\mathbf{X}_A^{Relig}, \mathbf{x}_0^T\mathbf{X}_A^{Born}, \mathbf{w}_0^T\mathbf{X}_A^{Attend}, \mathbf{w}_0^T\mathbf{X}_A^{Party}, \mathbf{w}_0^T\mathbf{X}_A^{StateType}\Big] \\
\hat{\mathbf{T}}^{X_{A.house}} &= \left[\mathbf{w}_0^T\mathbf{X}_A^{House}\right] \\
\hat{\mathbf{T}}^{X_{A.interaction}} &= \Big[\mathbf{w}_0^T\mathbf{X}_A^{Gender:Age}, \mathbf{w}_0^T\mathbf{X}_A^{Gender:Race}, \mathbf{w}_0^T\mathbf{X}_A^{Race:Age}, \\
&\qquad \mathbf{w}_0^T\mathbf{X}_A^{Party:Approval}, \mathbf{w}_0^T\mathbf{X}_A^{StateType:Party}, \mathbf{w}_0^T\mathbf{X}_A^{StateType:Approval}\Big]
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathbf{T}}^{X_{B.main}} &= \Big[\mathbf{w}_B^T \mathbf{X}_B^{Age}, \mathbf{w}_B^T \mathbf{X}_B^{Gender}, \mathbf{w}_B^T \mathbf{X}_B^{Educ}, \\
&\qquad \mathbf{w}_B^T \mathbf{X}_B^{Relig}, \mathbf{x}_B^T \mathbf{X}_B^{Born}, \mathbf{w}_B^T \mathbf{X}_B^{Attend}, \mathbf{w}_B^T \mathbf{X}_B^{Party}, \mathbf{w}_B^T \mathbf{X}_B^{StateType}\Big] \\
\hat{\mathbf{T}}^{X_{B.house}} &= \left[\mathbf{w}_B^T \mathbf{X}_B^{House}\right] \\
\hat{\mathbf{T}}^{X_{B.interaction}} &= \Big[\mathbf{w}_B^T \mathbf{X}_B^{Gender:Age}, \mathbf{w}_B^T \mathbf{X}_B^{Gender:Race}, \mathbf{w}_B^T \mathbf{X}_B^{Race:Age}, \\
&\qquad \mathbf{w}_B^T \mathbf{X}_B^{Party:Approval}, \mathbf{w}_B^T \mathbf{X}_B^{StateType:Party}, \mathbf{w}_B^T \mathbf{X}_B^{StateType:Approval}\Big]
\end{aligned}
$$

For governor race:

$$
\begin{aligned}
\mathbf{X}_{A.governor} &= \left[\underset{n_A \times 1}{\mathbf{1}}, \mathbf{X}_{A.main}, \mathbf{X}_{A.interaction}\right] \\
\hat{\mathbf{T}}^{X_{A.governor}} &= \left[\mathbf{w}_0^T \underset{n_A \times 1}{\mathbf{1}}, \hat{\mathbf{T}}^{X_{A.main}}, \hat{\mathbf{T}}^{X_{A.interaction}}\right] \\
\hat{\mathbf{T}}^{X_{B.governor}} &= \left[\mathbf{w}_B^T \underset{n_B \times 1}{\mathbf{1}}, \hat{\mathbf{T}}^{X_{B.main}}, \hat{\mathbf{T}}^{X_{B.interaction}}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{w}^{ECGREG.governor} &= \mathbf{w}_0 + \mathbf{W}_0 \mathbf{X}_{A.governor} \left(\mathbf{X}_{A.governor}^T \mathbf{W}_0 \mathbf{X}_{A.governor}\right)^{-1} \times \\
&\left(\hat{\mathbf{T}}^{X_{B.governor}} - \hat{\mathbf{T}}^{X_{A.governor}}\right)^T
\end{aligned}
$$

For senate race:

$$
\begin{aligned}
\mathbf{X}_{A.senate} &= \left[\underset{n_A \times 1}{\mathbf{1}}, \mathbf{X}_{A.main}, \mathbf{X}_{A.house}, \mathbf{X}_{A.interaction}\right] \\
\hat{\mathbf{T}}^{X_{A.senate}} &= \left[\mathbf{w}_0^T \underset{n_A \times 1}{\mathbf{1}}, \hat{\mathbf{T}}^{X_{A.main}}, \hat{\mathbf{T}}^{X_{A.house}}, \hat{\mathbf{T}}^{X_{A.interaction}}\right] \\
\hat{\mathbf{T}}^{X_{B.senate}} &= \left[\mathbf{w}_B^T \underset{n_B \times 1}{\mathbf{1}}, \hat{\mathbf{T}}^{X_{B.main}}, \hat{\mathbf{T}}^{X_{B.house}}, \hat{\mathbf{T}}^{X_{B.interaction}}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{w}^{ECGREG.senate} &= \mathbf{w}_0 + \mathbf{W}_0 \mathbf{X}_{A.senate} \left(\mathbf{X}_{A.senate}^T \mathbf{W}_0 \mathbf{X}_{A.senate}\right)^{-1} \times \\
&\left(\hat{\mathbf{T}}^{X_{B.senate}} - \hat{\mathbf{T}}^{X_{A.senate}}\right)^T
\end{aligned}
$$

4.8 Results

Following the convention of published voting spreads, we write $+\%R$ for a predicted spread indicating higher proportion of votes for the Republican Party, and $+\%D$ for a predicted spread indicating higher proportion of votes for the Democratic Party.

4.8.1 Direction and error

Table 4.3 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. Without weighting adjustments, the sample has higher Republican turn-out than Democratic turn-out, with 10 out of 11 states biasing toward Republican candidates. STATEWT reduced the bias for most states, while PSCORE and ECGREG may have over-adjusted toward Democratic direction. ECLASSO is the only estimator that reduced unadjusted sample bias to within 6% of true values. On average, ECLASSO also has the smallest relative error across the states.

Table 4.4 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 most states, while PSCORE, and ECGREG over-adjusted toward Democratic direction. ECLASSO is the only estimator that reduced unadjusted sample bias to within 8% of true values. On average, ECLASSO again has the smallest relative error across states.

Table 4.3: U.S. 2014 midterm election governor voting spread estimates and direction

				D-R estimates				
State	**analytical n**	**benchmark n**	**True D-R**	**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						

	Direction Correct					Relative Error				
State	**UNWT**	**STATEWT**	**PSCORE**	**ECGREG**	**ECLASSO**	**UNWT**	**STATEWT**	**PSCORE**	**ECGREG**	**ECLASSO**
Arizona	R - YES	R - YES	R - YES	D - NO	R - YES	+1.29%R	+1.63%D	+8.65%D	+23.51%D	+3.74%D
California	D - YES	D - YES	D - YES	D - YES	D - YES	+4.98%R	+0.50%D	+1.44%D	+17.44%D	+0.42%R
Florida	R - YES	R - YES	R - YES	D - NO	R - YES	+4.69%R	+0.98%R	+0.50%R	+8.08%D	+0.02%D
Georgia	R - YES	R - YES	R - YES	R - YES	R - YES	+5.84%R	+0.69%R	+1.77%R	+5.51%D	+0.38%R
Illinois	R - YES	R - YES	R - YES	R - YES	R - YES	+9.62%R	+3.86%R	+9.37%R	+12.89%R	+5.11%R
Michigan	R - YES	R - YES	R - YES	R - YES	R - YES	+10.00%R	+7.87%R	+7.69%R	+14.31%R	+5.71%R
New York	D - YES	D - YES	D - YES	D - YES	D - YES	+0.11%R	+4.83%D	+4.56%D	+25.16%D	+4.04%D
Ohio	R - YES	R - YES	R - YES	R - YES	R - YES	+4.49%R	+3.66%R	+0.39%R	+4.47%R	+0.45%R
Pennsylvania	D - YES	D - YES	D - YES	D - YES	D - YES	+1.53%D	+1.97%R	+12.93%D	+23.78%D	+5.78%D
Texas	R - YES	R - YES	R - YES	R - YES	R - YES	+5.32%R	+1.72%D	+0.05%R	+0.36%D	+0.29%R
Wisconsin	R - YES	R - YES	R - YES	R - YES	R - YES	+0.73%R	+10.79%R	+3.49%D	+4.60%D	+4.36%D
					AVERAGE	+4.14%R	+1.92%R	+1.03%D	+6.98%D	+0.51%D

Table 4.4: U.S. 2014 midterm election senate voting spread estimates and direction

				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						

	Direction Correct					Relative Error				
State	**UNWT**	**STATEWT**	**PSCORE**	**ECGREG**	**ECLASSO**	**UNWT**	**STATEWT**	**PSCORE**	**ECGREG**	**ECLASSO**
Georgia	R - YES	R - YES	R - YES	D - NO	R - YES	+5.63%R	+0.31%D	+4.12%D	+9.75%D	+2.83%R
Illinois	D - YES	D - YES	D - YES	D - YES	D - YES	+9.08%R	+5.51%R	+4.46%D	+2.43%D	+4.25%R
Michigan	D - YES	D - YES	D - YES	D - YES	D - YES	+8.15%R	+10.05%R	+7.61%D	+2.60%D	+5.19%R
Minnesota	D - YES	D - YES	D - YES	D - YES	D - YES	+4.04%R	+9.23%R	+1.98%D	+3.98%R	+0.42%R
New Jersey	D - YES	D - YES	D - YES	D - YES	D - YES	+2.03%D	+5.99%D	+17.55%D	+20.70%D	+3.11%D
North Carolina	R - YES	R - YES	D - NO	D - NO	R - YES	+3.00%R	+5.28%R	+2.46%D	+17.11%D	+1.10%R
Texas	R - YES	R - YES	R - YES	R - YES	R - YES	+7.76%R	+0.03%D	+0.52%R	+0.54%D	+4.51%R
Virginia	D - YES	D - YES	D - YES	D - YES	D - YES	+4.36%D	+4.73%D	+17.54%D	+23.06%D	+7.54%D
					AVERAGE	+3.91%R	+2.38%R	+6.90%D	+9.03%D	+0.96%R

4.8.2 Root-mean-square-error

Table 4.5 lists bias, standard error, and root-mean-square error of each estimator in predicting governor election 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. Thus 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 has the lowest average RMSE across the states.

Table **??** lists bias, standard error, and root-mean-square error of each estimator in predicting senate election spreads. Similar to the results for governor election, on average, the bias-reduction of STATEWT was not sufficient to offset the increased variance in the estimates. The average RMSE of STATEWT is larger than UNWT's. Both PSCORE and ECGREG performed poorly with larger average bias and standard error than unadjusted estimates. With the same benchmark sample, working model, and variance estimator as PSCORE and ECGREG, ECLASSO again is able to produce standard errors that are comparable to STATEWT's, and is the only estimator with improved average RMSE over unweighted estimates.

Table 4.5: U.S. 2014 governor race root-mean-square-error

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

Table 4.6: U.S. 2014 senate race root-mean-square-error

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

4.8.3 Coverage

Table 4.7 lists 90% confidence intervals of each estimator in governor elections. The UNWT confidence intervals are too narrow, covering true spreads in only 4 out of 11 states. ECLASSO and STATEWT confidence intervals both covered 9 out of 11 true spreads (82%), close to the expected 90% coverage rate. PSCORE covered 8, and ECGREG covered only 6. Figure 4.1 displays the boxplots of 90% confidence intervals of each estimator across 11 states, as well as the true values in solid red horizontal lines. ECLASSO confidence intervals are consistently around the true values. Among weighted estimators, ECLASSO also has comparable interval width as STATEWT's, if not narrower.

Table 4.8 lists 90% confidence intervals of each estimator in senate elections. The UNWT confidence intervals performed even worse than governor forecasts, covering only 1 out of 8 true spreads. 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, ECGREG covered 4, while PSCORE covered 5. Figure 4.2 displays the boxplots of 90% confidence intervals of each estimator across 8 states, as well as the true values in solid red horizontal lines. Even with wide confidence intervals, PSCORE and ECGREG still do not cover the true spread in estimates for Virginia and New Jersey. ECGREG estimates are also largely biased for Georgia and North Carolina. Besides estimates for Virginia, which no estimator performed well, ECLASSO confidence intervals are consistently around the true values.

Table 4.7: U.S. 2014 senate race 90% CI coverage

		UNWT			STATEWT			PSCORE			ECGREG			ECLASSO		
State	**D-R Truth**	**90% CI**		**Covered**	**90% CI**		**Covered**	**90% CI**		**Covered**	**90% CI**		**Covered**	**90% CI**		**Covered**
Arizona	+12%R	*(+8%R*	*+18%R)*	Yes	*(+2%R*	*+19%R)*	Yes	*(+9%D*	*+15%R)*	Yes	*(+1%R*	*+25%D)*	No	*(+1%R*	*+15%R)*	Yes
California	+19%D	*(+10%D*	*+17%D)*	No	*(+14%D*	*+24%D)*	Yes	*(+12%D*	*+28%D)*	Yes	*(+20%D*	*+53%D)*	No	*(+13%D*	*+23%D)*	Yes
Florida	+1%R	*(+3%R*	*+9%R)*	No	*(+3%D*	*+7%R)*	Yes	*(+8%D*	*+12%R)*	Yes	*(+2%R*	*+16%D)*	Yes	*(+4%D*	*+7%R)*	Yes
Georgia	+8%R	*(+10%R*	*+17%R)*	No	*(+3%R*	*+14%R)*	Yes	*(+0%R*	*+19%R)*	Yes	*(+8%D*	*+12%R)*	Yes	*(+2%R*	*+14%R)*	Yes
Illinois	+5%R	*(+11%R*	*+17%R)*	No	*(+4%R*	*+13%R)*	Yes	*(+7%R*	*+21%R)*	No	*(+3%R*	*+32%R)*	Yes	*(+5%R*	*+15%R)*	No
Michigan	+4%R	*(+12%R*	*+16%R)*	No	*(+9%R*	*+15%R)*	No	*(+6%R*	*+17%R)*	No	*(+10%R*	*+27%R)*	No	*(+5%R*	*+14%R)*	No
New York	+13%D	*(+9%D*	*+17%D)*	Yes	*(+13%D*	*+24%D)*	Yes	*(+9%D*	*+26%D)*	Yes	*(+24%D*	*+52%D)*	No	*(+12%D*	*+22%D)*	Yes
Ohio	+31%R	*(+32%R*	*+39%R)*	No	*(+30%R*	*+40%R)*	Yes	*(+22%R*	*+40%R)*	Yes	*(+26%R*	*+44%R)*	Yes	*(+27%R*	*+36%R)*	Yes
Pennsylvania	+10%D	*(+8%D*	*+15%D)*	Yes	*(+2%D*	*+13%D)*	Yes	*(+15%D*	*+29%D)*	No	*(+20%D*	*+46%D)*	No	*(+10%D*	*+20%D)*	Yes
Texas	+20%R	*(+23%R*	*+29%R)*	No	*(+13%R*	*+24%R)*	Yes	*(+11%R*	*+29%R)*	Yes	*(+12%R*	*+28%R)*	Yes	*(+15%R*	*+26%R)*	Yes
Wisconsin	+6%R	*(+4%R*	*+8%R)*	Yes	*(+13%R*	*+20%R)*	No	*(+4%D*	*+8%R)*	Yes	*(+8%D*	*+11%R)*	Yes	*(+3%D*	*+6%R)*	Yes

Table 4.8: U.S. 2014 senate race 90% CI coverage

		UNWT			STATEWT			PSCORE			ECGREG			ECLASSO		
State	**D-R Truth**	**90% CI**		**Covered**	**90% CI**		**Covered**	**90% CI**		**Covered**	**90% CI**		**Covered**	**90% CI**		**Covered**
Georgia	+8%R	*(+11%R*	*+17%R)*	No	*(+2%R*	*+13%R)*	Yes	*(+5%D*	*+12%R)*	Yes	*(+6%R*	*+10%D)*	No	*(+4%R*	*+17%R)*	Yes
Illinois	+10%D	*(+1%R*	*+4%D)*	No	*(+0%D*	*+9%D)*	No	*(+7%D*	*+22%D)*	Yes	*(+1%D*	*+25%D)*	Yes	*(+1%D*	*+11%D)*	Yes
Michigan	+13%D	*(+3%D*	*+7%D)*	No	*(+0%R*	*+7%D)*	No	*(+14%D*	*+29%D)*	No	*(+9%D*	*+23%D)*	Yes	*(+3%D*	*+13%D)*	No
Minnesota	+10%D	*(+4%D*	*+9%D)*	No	*(+4%R*	*+6%D)*	No	*(+5%D*	*+19%D)*	Yes	*(+0%R*	*+13%D)*	Yes	*(+5%D*	*+15%D)*	Yes
New Jersey	+13%D	*(+11%D*	*+21%D)*	Yes	*(+11%D*	*+27%D)*	Yes	*(+19%D*	*+41%D)*	No	*(+18%D*	*+49%D)*	No	*(+10%D*	*+22%D)*	Yes
North Carolina	+2%R	*(+3%R*	*+7%R)*	No	*(+4%R*	*+10%R)*	No	*(+7%R*	*+9%D)*	Yes	*(+5%D*	*+26%D)*	No	*(+2%D*	*+8%R)*	Yes
Texas	+27%R	*(+33%R*	*+38%R)*	No	*(+22%R*	*+32%R)*	Yes	*(+21%R*	*+35%R)*	Yes	*(+20%R*	*+33%R)*	Yes	*(+26%R*	*+37%R)*	Yes
Virginia	+1%D	*(+3%D*	*+7%D)*	No	*(+2%D*	*+9%D)*	No	*(+11%D*	*+25%D)*	No	*(+16%D*	*+32%D)*	No	*(+3%D*	*+13%D)*	No

Figure 4.1: Voting spread for governor race and 90% CI

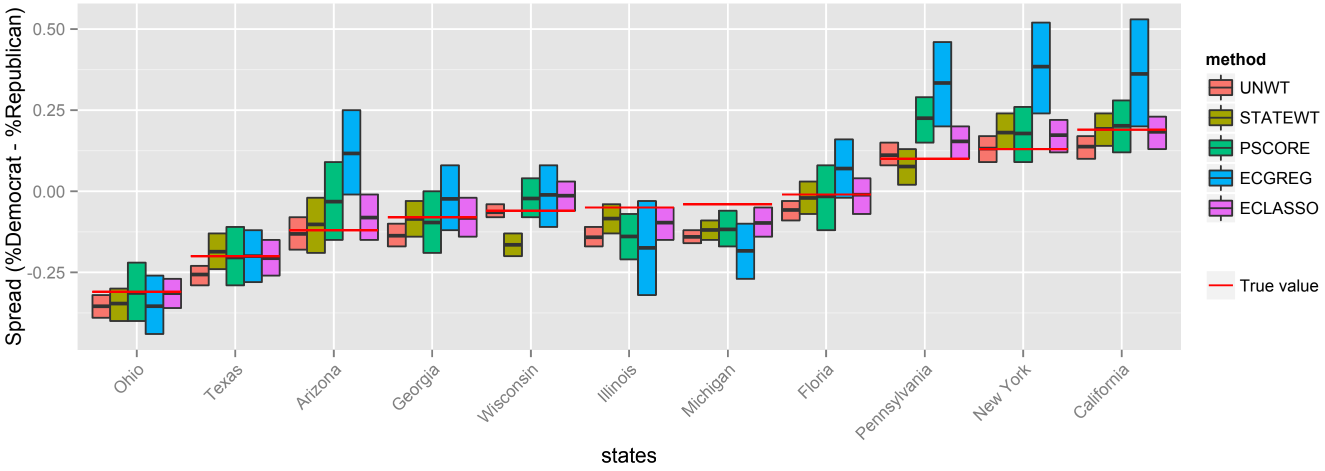
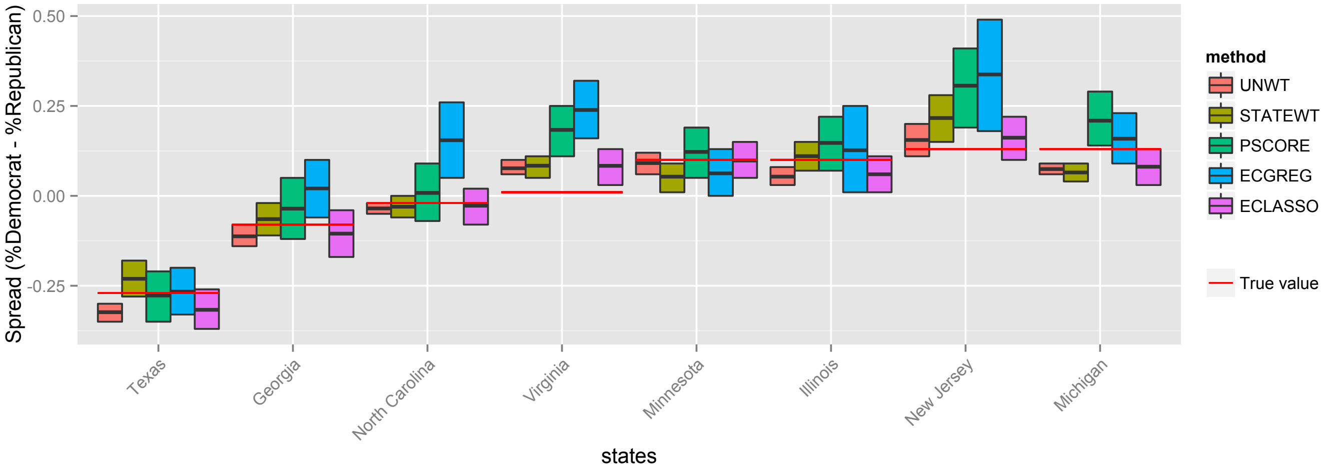


Figure 4.2: Voting spread for senate race and 90% CI



4.9 Discussion

Among the weighting adjustments performed in this analysis, ECLASSO is the most successful in reducing the unweighted bias in predicting voting spreads. For both governor and senate elections, ECLASSO reduced the overall bias from roughly 4% to under 1%. Unlike PSCORE and ECGREG, there is little evidence of ECLASSO over-adjusting the bias from Republican to Democrat. **The result answers the first research question: Yes, with ECLASSO, a small probability benchmark sample can correct the bias in a large non-probability internet-based election polls.**

We anticipate larger variances for PSCORE, ECGREG, and ECLASSO relative to the variances of STATEWT due to the small benchmark sample size. This is evident for PSCORE and ECGREG in both governor and senate election forecasts. For ECLASSO, the standard errors are comparable to STATEWT's in both races. In election data analysis, this chapter 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 benchmark. **Thus we have also addressed the second research question: Yes, with ECLASSO, we can make precise estimates of population quantities from a non-probability internet-based data with a small probability-based benchmark sample.**

In terms of bias, 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. However, there are several limitations with the analysis in this chapter. First, although ECLASSO can be extended to multinomial setting, we stayed within binary outcome framework and removed non-major party supporters from the analytical sample. Florida and New York are two states with non-trivial proportions of non-major party support (10% and 14%). However, ECLASSO estimates for both states have relatively small biases (0.05% and 5.72% in governor race). Discarding non-major party supporters may not have significant impact in our analysis. 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 together. The predictions are less sensitive to state benchmark sample sizes. Thus we do not observe a clear pattern of bias and variance as functions of state-level benchmark sample sizes. Larger state benchmark samples would allow for calibration to be done at the state level.

4.10 Conclusion

To date, there are well over 300 independent polling organizations aiming to predict winners of various U.S. elections (*Silver*, 2015). Most of the polling methodology is based on probability samples. With low response rates, however, it has become increasingly difficult and costly to collect probability samples given the short time-frame of election polling. In this chapter, we have demonstrated that post-survey adjustment with non-probability sample is possible to generate accurate election forecasts. ECLASSO is effective in leveraging information from benchmark sample, even if the benchmark sample is small. The results support the growing literature that suggests inference based on non-probability samples is possible with appropriate statistical modeling.

CHAPTER V

Conclusion

5.1 Summary

More than a decade ago, the survey research field anxiously anticipated the breakthrough of a new data collection medium – internet. In one of the first in-depth research into post-survey adjustment of web-based samples, *Lee* (2004) wrote:

> With the advance in communication technology and the accompanying societal and cultural changes, Web surveys are here to stay.

Indeed, web-based non-probability samples have not only proliferated the market research and public opinion sectors, they have become a data collection method in clinical and psychological studies, and among other research disciplines (*Liu et al.*, 2010; *Casini and Scozzafava*, 2014; *Hamama-Raz et al.*, 2014). With probability-based sampling cost continue to rise, we will see a steady, if not rapid growth in the use of non-probability samples. Current methods in post-survey adjustments of non-probability samples are met with multiple challenges, mainly because the effort to date focuses on constructing weights to correct all errors in a non-probability sample. The resulting weights that are meant to enable population inference of all variables, however, have limited success in providing unbiased inference to the population (*Schonlau et al.*, 2004, 2009; *Bethlehem*, 2010). This dissertation addresses the growing demand for making unbiased inference based on non-probability samples. Under the model-assisted calibration framework, we focus on constructing a single set of weights that specifically allow for estimates of population totals with small root-mean-square-error. We employ a modern statistical model, Least Angle Shrinkage and Selection Operator (LASSO), as the assisting model to perform model selection and parameter estimation simultaneously. LASSO reduces the risk of sample over-fitting, which is a key feature required for successful model-assisted calibration.

In chapter II, we developed LASSO calibration and derived a theoretically unbiased estimator for population totals from non-probability samples. Simulation demonstrated that LASSO calibration has a significant advantage over traditional calibration for estimating totals of binary outcome variables in terms of having smaller bias and variance. The improvement over the traditional generalized regression estimator, GREG, is even more pronounced when the calibration variables exhibit high correlations, and under non-ignorable sampling. Although the asymptotic linearized variance estimates did poorly in terms of coverage and bias, a more robust variance estimate of LASSO calibration estimator with bootstrap re-sampling is shown to be a viable alternative. We also applied LASSO calibration to National Health Interview Survey (NHIS) for estimating the population total of adults diagnosed with cancer. Without correct design weights, LASSO calibration was able to adjust unweighted estimates to produce an estimate that is close to the correctly-weighted estimate. In short, in chapter II, we studied and understood generally how LASSO performs under different sampling schemes, outcome types, and calibration covariance structures. We then extended the theoretical framework of LASSO calibration to estimated-control LASSO calibration in chapter III, where the benchmark samples can be small. We derived ECLASSO calibration estimator of total and its asymptotic linearized variance estimates incorporating benchmark sampling errors. Through a simulation study with an actual dataset, the National Health Interview Survey 2013, we showed that

ECLASSO estimator can achieve better root-mean-square-error than traditional post-survey weighting adjustment methods, even when the benchmark sample is small. We compared the asymptotic linearized variance estimates with boot-strap variance estimates, and demonstrated again that the bootstrap variance estimate is more robust both in terms of coverage and bias. In Chapter IV, we demonstrated the potential of the LASSO calibration estimator in reducing the bias of an actual internet-based non-probability election polling data. Using a phone-based probability benchmark sample that is less than one-thirtieth of the internet-based data in size, ECLASSO calibration was able to make accurate predictions of voting spreads for the 2014 U.S. mid-term elections. In this application, ECLASSO uniformly outperformed traditional weighting adjustment methods in average bias and root-mean-square-error. The method and framework developed in research provide a valuable tool for the growing number of researchers interested in analyzing non-probability samples.

5.2 Limitations

This dissertation has four key limitations. First, for the theoretical framework, we assumed that the benchmark samples are drawn from a single-stage probability-based-sampling design. This assumption is used to facilitate the development of linearized variance estimates of predicted benchmark totals. In practice, different types of probability-based benchmark samples can be used, many of which are multi-stage area-probability samples. The current theoretical development does not include such sampling designs. One can, however, easily incorporate finite-population re-sampling methods of the benchmark sample to calculate a variance estimate for ECLASSO calibration estimators.

Secondly, the theoretical framework assumes that the true superpopulation regression parameters are a subset of the LASSO regression estimator. In practice, no one can be certain of what constitutes true superpopulation parameters. Thus it is not feasible to check whether the assumption is met. Although the applications in this dissertation demonstrate that without knowing the true underlying model, LASSO calibration can still produce approximately unbiased estimates, an extension of the theoretical framework to include miss-specified models will make LASSO calibration even more practical.

Thirdly, both linearized and bootstrap variance estimates of LASSO calibration estimators are prone to biases. Asymptotic linearized variance estimates have the undesirable negative bias property, while bootstrap variance estimates tend to be positively biased. One approach to develop more robust variance estimates is to incorporate model variances, i.e., variances of model parameters, to develop a variance estimate not based on asymptotic properties. For LASSO regression, this can be challenging, since there is no known theoretical variance formula for LASSO parameter estimates. *Tibshirani* (1996) suggested using ridge-regression parameter variance estimates for non-zero parameters of LASSO. This may under-estimate LASSO estimator variance, since it does not account for variance due to variable selection. Another possibility is to explore bias-corrected bootstrap to reduce the positive bias observed in the simulation studies (*DiCiccio and Efron*, 1996).

Lastly, the research uses unweighted LASSO regression estimates, because the focus is on non-probability samples where we do not have design weights. Survey-weighted LASSO has been developed for linear LASSO regression (*McConville*, 2011). The convergence of survey-weighted LASSO parameter estimates to true superpopulation parameters remain true. The covariance component needs to be adjusted in the oracle property to incorporate survey weights. The theoretical framework of model-assisted calibration, however, does not rely on the covariance of LASSO parameter estimates, only on the convergence to the true superpopulation parameters values. Thus survey-weighted LASSO is still fully applicable under probability-based framework. In fact, R *glmnet* function for obtaining LASSO regression parameter estimates can take in survey weights (*Friedman et al.*, 2010). Thus the simulations in this dissertation can be repeated with survey-weighted LASSO calibration.

5.3 Future research

This dissertation has a very specific goal to establish a method capable of producing unbiased estimated totals from non-probability samples. To accomplish the goal, we made strong assumptions to facilitate the theoretical developments. There are certainly many more research paths along LASSO calibration which can lead to fruitful results. I state some here that may be of interests to other researchers.

(1) This research has shown that LASSO calibration estimator is asymptotically model-unbiased even without using the correct design weights, regardless of how the samples are generated when: (i) The correct model is within the full LASSO regression model, and (ii) the full range of values of $\mathbf{X}$ is observed. There can still be persistent sample bias if LASSO fails to satisfy the oracle property. Further research on the relationship between sample generation and the rate of convergence of LASSO penalty parameter, λ_n, can provide more insights to when the sample bias persists if the correct design weights are not used in estimation.

(2) In calibration, a distance measure is used between calibrated weights and original design weights. When the original design weights do not guarantee unbiased estimates, certain distance measures may work better than others in constructing the calibrated weights. For non-probability samples, it will be an area of interest to see how distance measures can impact the reduction in sample bias.

(3) Many internet-based data have categorical outcomes of interest. For example, opinion polls often attempt to find the public's support on certain agendas: Very likely/Likely/Unlikely, Against/Neutral/Support, etc. Extension of the theoreti-

cal framework to include multinomial LASSO assisting models can enable LASSO calibration on a wide range of outcome types in non-probability samples.

(4) Some researchers may wish the LASSO calibrated weights be applicable to more than one variable. This can be impractical for inference from non-probability samples, but is relevant for probability samples. From an operational perspective, this can be easily achieved by predicting the expected values of multiple outcome variables, then calibrate on a vector of predicted outcome totals. If a function of the multiple outcome totals is of interest, an additional theoretical step to derive the covariances between predicted outcome totals will be required.

(5) The asymptotic linearized variance estimates developed in this dissertation all are prone to negative biases, while naive bootstraps typically have positive biases. For ECLASSO estimator, one potential approach is to bootstrap only the analytical sample to estimate the variance component given the benchmark sample, and use linearized variance estimate on benchmark estimated totals. This approach combines both linearization and re-sampling, which can possibly average out the observed biases in the simulations.

(6) Finally, the potential of LASSO calibration in this research is demonstrated through an internet-based non-probability sample that is not based on a web-volunteer panel. The most complex type of error in non-probability samples is self-selection-bias, which is more likely to be observed in web-volunteer samples. It will be a great addition to calibration literature with an analysis of LASSO calibration on web-volunteer samples.

APPENDICES

APPENDIX A

Appendix

A.1 Proofs

A.1.1 Lemma II.2

Lemma II.2. *Assume the superpopulation model:*

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

Let $\mathbf{B}$ *be the finite-population quasilikelihood estimate of* $\boldsymbol{\beta}$*,* $\mathbf{B} \to \boldsymbol{\beta}$*, under conditions (2.4.2.i)-(2.4.2.v), the model-assisted asymptotic estimator for a population total is:,*

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

where

$$\begin{aligned} \mu_i &= \mu(\mathbf{x}_i, \mathbf{B}) \\ B^{MC} &= \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2} \end{aligned}$$

Proof. The proof is adopted and expanded from the proof of Theorem 1 in (*Wu and Sitter*, 2001), with slight modifications in notations to be consistent with the dissertation. We begin by deriving the asymptotic model-assisted estimator for a population mean, $\hat{\bar{y}}^{MC} = N^{-1}\hat{T}_y^{MC}$ (see equation (2.2.2.4)). By conditions (2.4.2.ii) and (2.4.2.iii), the second order Taylor series expansion of $\mu(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ around $\mathbf{B}$ is:

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

for $\mathbf{B}^* \in (\hat{\mathbf{B}}, \mathbf{B})$ or $(\mathbf{B}, \hat{\mathbf{B}})$. Let

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

Note that $\mathbf{h}$ is a vector of length m and $\mathbf{k}$ is a matrix of size $m \times m$, where m is the number of parameters in $\boldsymbol{\beta}$. By conditions (2.4.2.ii) and (2.4.2.iii),

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

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

The population mean of (A.2) based on sample s_A:

$$\begin{aligned}
N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\hat{\mathbf{B}}) &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i,\mathbf{B})\right)^T (\hat{\mathbf{B}}-\mathbf{B}) + \\
&\qquad (\hat{\mathbf{B}}-\mathbf{B})^T N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{k}(\mathbf{x}_i,\mathbf{B}^*)\right)(\hat{\mathbf{B}}-\mathbf{B}) \\
&\text{(by conditions (2.4.2.i) and (2.4.2.iii))} \\
&= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i,\mathbf{B})\right)^T (\hat{\mathbf{B}}-\mathbf{B}) + \\
&\qquad O_p(1/\sqrt{n})O_p(1/\sqrt{n}) \\
&= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i,\mathbf{B})\right)^T (\hat{\mathbf{B}}-\mathbf{B}) + O_p\left(\frac{1}{n}\right)
\end{aligned} \tag{A.5}$$

By conditions (2.4.2.i), (2.4.2.iv), and equation (A.5):

$$\begin{aligned}
&N^{-1}\sum_{k=1}^{N} \mu(\mathbf{x}_k,\hat{\mathbf{B}}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\hat{\mathbf{B}}) \\
&\quad = N^{-1}\sum_{k=1}^{N} \mu(\mathbf{x}_i,\mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\mathbf{B}) + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{n}\right) \\
&\quad = N^{-1}\sum_{k=1}^{N} \mu(\mathbf{x}_i,\mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i,\mathbf{B}) + O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{A.6}$$

Note that,

$$
\begin{aligned}
\bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \hat{\mathbf{B}}) \Big/ \sum_{i\in s_A} d_i^A \\
&= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)(\hat{\mathbf{B}} - \mathbf{B})\right) \\
&\text{(by conditions (2.4.2.i) and (2.4.2.iii))} \\
&= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B})\right) + O_p(1/n) \\
&= \bar{\mu} + \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + O_p(1/n) \\
&\text{(by condition (2.4.2.i) and (A.3))} \\
&= \bar{\mu} + O_p(1/\sqrt{n}) + O_p(1/n) \\
&= \bar{\mu} + O_p(1/\sqrt{n}) \qquad \text{(A.7)}
\end{aligned}
$$

Then from (A.2) and (A.7),

$$N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})$$

$$= N^{-1}\sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T\mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)(\hat{\mathbf{B}} - \mathbf{B}) - \hat{\mu}\right)$$

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + N^{-1}\sum_{i\in s_A}\mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) +$$

$$N^{-1}\sum_{i\in s_A}(\hat{\mathbf{B}} - \mathbf{B})^T\mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)(\hat{\mathbf{B}} - \mathbf{B}) - O_p(1/\sqrt{n})$$

(by conditions (2.4.2.i) and (2.4.2).iii)

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + N^{-1}\sum_{i\in s_A}\mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + O_p(1/n) - O_p(1/\sqrt{n})$$

(by condition (2.4.2.i) and (A.3))

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + O_p(1/\sqrt{n}) + O_p(1/n) - O_p(1/\sqrt{n})$$

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + O_p(1/\sqrt{n}) \tag{A.8}$$

$$N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})^2 = N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + (O_p(1/\sqrt{n}))^2$$

$$= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + O_p(1/n) \tag{A.9}$$

From (A.8) and (A.9) we have:

$$\begin{aligned}\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})^2} = \frac{N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})(y_i - \bar{y})}{N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \hat{\bar{\mu}})^2} \\ &= \frac{\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})(y_i - \bar{y}) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + O_p\left(\frac{1}{n}\right)} \\ &\to B^{MC} \quad \text{as } n \to \infty \end{aligned} \tag{A.10}$$

Thus $\hat{B}^{MC} = B^{MC} + o_p(1)$, and we have:

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

Since $N = O_p(N)$, we have $N \cdot o_P(1/\sqrt{n}) = O_p(N) o_p(1/\sqrt{n}) = o_p(N/\sqrt{n})$. Thus,

$$
\begin{aligned}
\hat{t}_y^{MC} = N\hat{\bar{y}}^{MC} &= N\left(N^{-1}\mathbf{d}^A\mathbf{y} + \left(N^{-1}\sum_{k=1}^{N}\mu(\mathbf{x}_k,\mathbf{B}) - N^{-1}\sum_{i\in s_A}\mu(\mathbf{x}_i,\mathbf{B})\right) B^{MC} + o_p\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \mathbf{d}^A\mathbf{y} + \left(\sum_{k=1}^{N}\mu(\mathbf{x}_k,\mathbf{B}) - \sum_{i\in s_A}\mu(\mathbf{x}_i,\mathbf{B})\right) B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right) \qquad \text{(A.11)}
\end{aligned}
$$

□

A.2 R Code

```
# function       : mycv.glmnet
# description    : performs cross-validation for lasso regression
#
# args
# x              : candidate variables
# y              : dependent variable
# standardize    : whether to standardize (subtract mean and divide by sd,
#                  FALSE if all variables are categorical)
# intercept      : whether to include intercept in model
# penalty.factor : penalty factor associated with each candidate variable
#                  (default to 1)
# seed           : seed used to generate cross-validation samples
# nfolds         : number of cross-validation groups
# alpha          : penalty coefficient, 0=bridge regression, 1=lasso
# family         : distribution family (binomial or gaussian supported)
# fun            : function to determine the best metric
#                  (max for AUC, min for rmse)
# type           : predicted value type (response = predicted mean)
# nlambda        : number of lambda grids
# weights        : observation weights
# cores          : number of cpu cores for parallel processing
# adaptive       : use adaptive Lasso (default=FALSE)
#
# return value : coef =lasso coefficient corresponding to
#                  optimal penalty parameter obtained through cross-validation
#                metrics = matrix of metrics, each row is a lambda,
#                  each column is measure on a test set (# of column = # of CV)
#                lambda = vector of lambda grids
#                best.metric.index = the index
#                  (row number in metrics) the has the best average measure
#                method = the method with best weight power for adaptive
#
# assume x does not contain intercept column, and all variables are
# dummy vectors (glmnet does not dummify variables internally)
mycv.glmnet ←
function (x, y, standardize = FALSE, intercept = TRUE, penalty.factor = NULL,
    seed = NULL, nfolds = 5, family = ”binomial”, alpha = 1,
    measure = ”auc”, fun = ”max”, type = ”response”, weights = NULL,
    nlambda = 100, cores = 1, adaptive = FALSE) {

    predictors ← colnames(x);
    glm.formula ← update(formula(formula.str(predictors)),”y∼.”);

    if (family == ”binomial”) {
      y ← as.factor(y);
    }
    if (length(penalty.factor) == 0) {
      penalty.factor = rep(1, ncol(x));
    }
    if (is.null(weights)) {
      weights = rep(1, nrow(x));
    }
    if (!is.null(seed)) set.seed(seed);

    N = length(y);
```

```
foldsize = round(N/nfolds);
fold.index ← list();
fold.index[[nfolds]] ← NULL;
index ← c(1:N);
for (i in 1:(nfolds - 1)) {
  if (i == 1) {
    samp.index ← sample(index, size = foldsize, replace = FALSE);
  } else {
    samp.index ← sample(index[-unlist(fold.index)], size = foldsize,
                                        replace = FALSE);
  }
  fold.index[[i]] ← samp.index;
}
fold.index[[nfolds]] = index[-unlist(fold.index)];

glmnet.obj0 ← NULL;
if (adaptive == FALSE) {
  glmnet.obj0 ← glmnet(x, y, penalty.factor = penalty.factor,
      family = family, alpha = alpha, standardize = standardize,
      intercept = intercept, weights = weights, nlambda = nlambda);

  metrics0 ← matrix(rep(NA, nfolds * length(glmnet.obj0$lambda)),
                            ncol = nfolds);

  for (i in 1:nfolds) {
    train.index ← unlist(fold.index[-i]);
    test.index ← fold.index[[i]];
    train.glmnet ← glmnet(x[train.index, ], y[train.index],
            penalty.factor = penalty.factor, family = family,
            lambda = glmnet.obj0$lambda, standardize = standardize,
            intercept = intercept, alpha = alpha,
            weights = weights[train.index], nlambda = nlambda);
    predict.y ← predict(train.glmnet, x[test.index, ], type = type);
    metric.fold ← rep(NA, length(glmnet.obj0$lambda));

    for (j in 1:length(glmnet.obj0$lambda)) {
      if (family == "binomial") {
        metric.fold[j] ← glmnet::auc(y[test.index],predict.y[,j],
                                                       weights[test.index]);
      } else {
        metric.fold[j] ←sum(abs(predict.y[, j] - y[test.index]) *
            (weights[test.index]))/sum(weights[test.index]);
      }
    }
    metrics0[, i] ← metric.fold
  }

  metric0.mean ← apply(metrics0, 1, mean);
  best.metric0 = eval(call(fun, metric0.mean));
  best.metric0.index = which(metric0.mean == best.metric0)[1];
  return(list(coef = coef(glmnet.obj0)[, best.metric0.index],
      metrics = metrics0, lambda = glmnet.obj0$lambda,
      best.metric.index = best.metric0.index, method = 0))
}

# adaptive
glmnet.obj1 ← NULL
glmnet.obj2 ← NULL
glmnet.obj3 ← NULL
glmnet.obj4 ← NULL

taus ← c(0.1,0.5,1,2);
tau1 = taus[1];
tau2 = taus[2];
tau3 = taus[3];
tau4 = taus[4];

# initial weights, whole data
data.ls.fit ← data.frame(y=y, x);
ls.fit ← NULL;
if (family == "binomial") {
  dsgn ← svydesign(ids=~1,weights=~weights,data=data.ls.fit);
  ls.fit ← svyglm(glm.formula, design=dsgn, family=quasibinomial());
} else {
  ls.fit ← lm(glm.formula, data = data.ls.fit, weights = weights);
}
beta.init.whole ← (coef(ls.fit))[-1];

# initial weights, each fold
beta.list ← list(); beta.list[[nfolds]] ← NULL;
for (i in 1:nfolds) {
  train.index ← unlist(fold.index[-i]);
  test.index ← fold.index[[i]];
  data.ls.fit ← data.frame(y=y[train.index], x[train.index,]);
  ls.fit ← NULL
  if (family == "binomial") {
    dsgn ← svydesign(ids=~1,weights=~weights[train.index],data=data.ls.fit);
    ls.fit ← svyglm(glm.formula, design=dsgn, family=quasibinomial());
  } else {
    ls.fit ← lm(glm.formula, data = data.ls.fit, weights = weights[train.index]);
  }
  beta.list[[i]] ← coef(ls.fit)[-1];
}

glmnet.objs ← list();
for (i in 1:4) {
  glmnet.objs[[i]] ← glmnet(x, y, penalty.factor = (1/abs(beta.init.whole))^taus[i],
          family = family, alpha = alpha, standardize = standardize,
          intercept = intercept, weights = weights, nlambda = nlambda);
}
glmnet.obj1 ← glmnet.objs[[1]];
glmnet.obj2 ← glmnet.objs[[2]];
glmnet.obj3 ← glmnet.objs[[3]];
glmnet.obj4 ← glmnet.objs[[4]];

metrics ← list();
metrics[[4]] ← NULL;
for (k in 1:4) {

  metric.folds ← matrix(rep(NA, nfolds*length(glmnet.objs[[k]]$lambda)),
                             ncol=nfolds);
  for (i in 1:nfolds) {
    train.index ← unlist(fold.index[-i]);
    test.index ← fold.index[[i]];
    beta.init ← beta.list[[i]];
    train.glmnet ← glmnet(x[train.index, ], y[train.index],
            penalty.factor = (1/abs(beta.init))^taus[k], family = family,
            lambda = glmnet.objs[[k]]$lambda, standardize = standardize,
            intercept = intercept, alpha = alpha,
            weights = weights[train.index], nlambda = nlambda);
    predict.y ← predict(train.glmnet, x[test.index, ], type = type);
    metric.fold ← rep(NA, length(glmnet.objs[[k]]$lambda));
    grid.n ← min(length(train.glmnet$lambda),
                          length(glmnet.objs[[k]]$lambda));
    for (j in 1:grid.n) {
      if (family == ”binomial”) {
        metric.fold[j] ← glmnet::auc(y[test.index],predict.y[,j],
                                                  weights[test.index]);
      } else {
        metric.fold[j] ←sum(abs(predict.y[, j] - y[test.index]) *
            (weights[test.index]))/sum(weights[test.index]);
      }
    }
    metric.folds[, i] ← metric.fold
  }
  metrics[[k]] ← metric.folds;
}
metrics1 ← metrics[[1]];
metrics2 ← metrics[[2]];
metrics3 ← metrics[[3]];
metrics4 ← metrics[[4]];

metric1.mean ← apply(metrics1, 1, mean, na.rm = T);
metric2.mean ← apply(metrics2, 1, mean, na.rm = T);
metric3.mean ← apply(metrics3, 1, mean, na.rm = T);
metric4.mean ← apply(metrics4, 1, mean, na.rm = T);
best.metric1 = eval(call(fun, metric1.mean));
best.metric2 = eval(call(fun, metric2.mean));
best.metric3 = eval(call(fun, metric3.mean));
best.metric4 = eval(call(fun, metric4.mean));
best.metric = eval(call(fun, best.metric1, best.metric2,
                            best.metric3, best.metric4));
best.metric.method = which(c(best.metric1, best.metric2,
                                 best.metric3, best.metric4) == best.metric)[1];
if (best.metric.method == 1) {
  best.metric1.index = which(metric1.mean == best.metric1)[1];
  return(list(coef = coef(glmnet.obj1)[, best.metric1.index],
            metrics = metrics1, lambda = glmnet.obj1$lambda,
            best.metric.index = best.metric1.index, method = 1));
}
if (best.metric.method == 2) {
  best.metric2.index = which(metric2.mean == best.metric2)[1]
  return(list(coef = coef(glmnet.obj2)[, best.metric2.index],
            metrics = metrics2, lambda = glmnet.obj2$lambda,
                best.metric.index = best.metric2.index, method = 2));
    }
    if (best.metric.method == 3) {
      best.metric3.index = which(metric3.mean == best.metric3)[1]
      return(list(coef = coef(glmnet.obj3)[, best.metric3.index],
                metrics = metrics3, lambda = glmnet.obj3$lambda,
                best.metric.index = best.metric3.index, method = 3));
    }
    best.metric4.index = which(metric4.mean == best.metric4)[1]
    return(list(coef = coef(glmnet.obj4)[, best.metric4.index],
                metrics = metrics4, lambda = glmnet.obj4$lambda,
                best.metric.index = best.metric4.index, method = 4));
}
```

BIBLIOGRAPHY

BIBLIOGRAPHY

Abramovich, F., and Y. Benjamini (1996), Adaptive thresholding of wavelet coefficients, *Computational Statistics and Data Analysis*, *22*, 351–361.

Abramowitz, A. I. (2008), Forecasting the 2008 presidential election with the time-for-change model, *PS: Political Science & Politics*, *41*(04), 691–695.

Bethlehem, J. G. (2010), Selection bias in web surveys, *International Statistical Review*, *78*(2), 161–188.

Bickel, P. J., and B. Li (2008), Regularization in statistics, *Test*, *15*, 271–344.

Biernacki, P., and D. Waldorf (1981), Snowball sampling: Problems and techniques of chain referral sampling, *Sociological Methods & Research*, *10*(2), 141–163.

Boboth, Z., A. Zimmer, M. Scott, E. Hilarides, D. Gromet, and C. Massey (2007), Weighting for coverage bias in internet surveys, `http://sci.sdsu.edu/math-reu/2007-6.pdf`.

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

Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber (2013), Data resource profile: The survey of health aging and retirement in europe (share), *International Journal of Epidemiology*, *42*(5), 1–10.

Brick, J. M., and D. Williams (2013), Explaining rising nonresponse rates in cross-sectional surveys, *The ANNALS of the American academy of political and social science*, *645*(1), 36–59.

Butler, N. A., and M. C. Denham (2000), The peculiar shrinkage properties of partial least squares regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(3), 585–593.

Butt, Z., J. Peipert, K. Webster, C. Chen, and D. Cella (2013), General population norms for the functional assessment of cancer therapy–kidney symptom index (fksi), *Cancer*, *119*(2), 429–437.

Candès, E. J., J. Romberg, and T. Tao (2006), Stable signal recovery from incomplete and inaccurate measurements, *Communications on Pure and Applied Mathematics*, *59*, 1207–1223.

Casini, L., and G. Scozzafava (2014), Market strategies in the beef sector: a discrete choice experiment for the segmentation of consumer preferences, in *Sustainability of the Agri-food System: Strategies and Performances: Proceedings of the 50th SIDEA Conference. Lecce, Chiostro dei Domenicani, 26-28 September 2013*, p. 113, Universitas Studiorum.

Coleman, J. S. (1958), Relational analysis: the study of social organizations with survey methods, *Human Organization*, *17*(4), 28–36.

Couper, M. P., J. A. Dever, and K. J. Gile (2013), Report of the aapor task force on non-probability sampling.

Curtin, R., S. Presser, and E. Singer (2005), Changes in telephone survey nonresponse over the past quarter century, *Public opinion quarterly*, *69*(1), 87–98.

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

Declercq, E. R., C. Sakala, M. P. Corry, and S. Applebaum (2007), Listening to mothers ii: Report of the second national us survey of women's childbearing experiences: Conducted january–february 2006 for childbirth connection by harris interactive® in partnership with lamaze international*, *The Journal of perinatal education*, *16*(4), 9.

Delavande, A., and C. F. Manski (2010), Probabilistic polling and voting in the 2008 presidential election evidence from the american life panel, *Public opinion quarterly*, *74*(3), 433–459.

Dever, J., and R. Valliant (2016), Greg estimation with undercoverage and estimated controls, forthcoming.

Dever, J. A. (2008), Sampling weight with estimated control totals, Ph.D. thesis, University of Maryland, College Park.

Dever, J. A., and R. Valliant (2010), A comparison of variance estimators for post-stratification to estimated control totals, *Survey Methodology*, *36*(1), 45–56.

DiCiccio, T. J., and B. Efron (1996), Bootstrap confidence intervals, *Statistical science*, pp. 189–212.

Donoho, D. L., and I. M. Johnstone (1994a), Ideal denoising in an orthonormal basis chosen from a library of bases, *C. R. Acad. Sci. Paris Ser. I Math*, *319*, 1317–1322.

Donoho, D. L., and I. M. Johnstone (1994b), Ideal spatial adaption by wavelet shrinkage, *Biometrika*, *81*, 425–455.

Esmer, Y., and T. Pettersson (2007), The effects of religion and religiosity on voting behavior.

Fan, J., and R. Li (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association*, *96*(456), 1348–1360.

Foucart, S., and H. Rauhut (2013), *A mathematical introduction to compressive sensing*, Birkhäuser, Boston.

Frank, O., and T. Snijders (1994), Estimating the size of hidden populations using snowball sampling, *Journal of Official Statistics*, *10*, 53–53.

Frankel, M. R., and L. R. Frankel (1987), Fifty years of survey sampling in the united states, *Public Opinion Quarterly*, pp. S127–S138.

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

Fu, W. J. (1998), Penalized regressions: the bridge versus the lasso, *Journal of computational and graphical statistics*, *7*(3), 397–416.

Fuchs, J. J. (1998), Detection and estimation of superimposed signals, *in Proceedings of IEEE International Conference of Acoustics, Speech, and Signal*, pp. 1649–1652.

Goldstein, T., and S. Osher (2009), The split bregman method for l1-regularized problems, *SIAM Journal on Imaging Sciences*, *2*(2), 323–343.

Goodman, L. A. (1961), Snowball sampling, *The annals of mathematical statistics*, pp. 148–170.

Groves, R. M. (2006), Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, R. M. (2011), Three eras of survey research, *Public Opinion Quarterly*, *75*(5), 861–871.

Gutsche, T. L., A. Kapteyn, E. Meijer, and B. Weerman (2014), The rand continuous 2012 presidential election poll, *Public Opinion Quarterly*, *78*(S1), 233–254.

Hamama-Raz, Y., Y. Palgi, A. Shrira, R. Goodwin, K. Kaniasty, and M. Ben-Ezra (2014), Gender differences in psychological reactions to hurricane sandy among new york metropolitan area residents, *Psychiatric quarterly*, pp. 1–12.

Healy, A. J., N. Malhotra, and C. H. Mo (2010), Irrelevant events affect voters' evaluations of government performance, *Proceedings of the National Academy of Sciences*, *107*(29), 12,804–12,809.

Heckathorn, D. D. (1997), Respondent-driven sampling: a new approach to the study of hidden populations, *Social problems*, *44*(2), 174–199.

Hill, B. M., and A. Shaw (2013), The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation, *PloS ONE*, *8*(6), e65,782.

Holbrook, A., J. Krosnick, A. Pfent, et al. (2007), The causes and consequences of response rates in surveys by the news media and government contractor survey research firms, *Advances in telephone survey methodology*, pp. 499–528.

Jagannathan, R., and T. Ma (2003), Risk reduction in large portfolios: Why imposing the wrong constraints helps, *The Journal of Finance*, *58*(4), 1651–1684.

Jayasuriya, B., and R. Valiant (1996), An application of restricted regression estimation in a household survey, *Survey Methodology*, *22*, 127–138.

Kamarianakis, Y., W. Shen, and L. Wynter (2012), Real-time road traffic forecasting using regime-switching space-time models and adaptive lasso, *Applied Stochastic Models in Business and Industry*, *28*(4), 297–315.

Kato, T., and M. Uemura (2012), Period analysis using the least absolute shrinkage and selection operator (lasso), *Publications of the Astronomical Society of Japan*, *64*(6), 122.

Kennel, T. L. (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, *Pew Research Center, Washington, DC*.

Kott, P. S. (2006), Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology*, *32*(2), 133–142.

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*(2), 196.

Lee, S. (2004), Statistical estimation methods in volunteer panel web surveys, Ph.D. thesis, University of Maryland, College Park.

Levendusky, M. S., and J. C. Pope (2011), Red states vs. blue states going beyond the mean, *Public Opinion Quarterly*, p. nfr002.

Li, Y., W.-K. Sung, and J. J. Liu (2007), Association mapping via regularized regression anaylsis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows, *The American Journal of Human Genetics*, *80*(4), 705–715.

Lindesmith, A. R. (1947), Opiate addiction.

Liu, H., D. Cella, R. Gershon, J. Shen, L. S. Morales, W. Riley, and R. D. Hays (2010), Representativeness of the patient-reported outcomes measurement information system internet panel, *Journal of clinical epidemiology*, *63*(11), 1169–1178.

Liu, Y. K., K. Henry, and M. Struddler (2012), Practical issues when calibrating weights for multiple skewed variables, *Proceedings of Survey Research Section, American Statistical Association*, pp. 4619–4632.

McConville, K. (2011), Improved estimation for complex surveys using modern regression techniques, Ph.D. thesis, Colorado State University.

McConville, K. S., F. J. Breidt, and T. C. M. Lee (2015), Model-assisted survey regression estimation with the lasso, submitted.

Meier, L., S. van de Geer, and P. Bühlmann (2008), The group lasso for logistic regression, *Journal of the Royal Statistics Society, B*, *70*, 53–71.

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

Nowell, C., and L. R. Stanley (1991), Length-biased sampling in mall intercept surveys, *Journal of Marketing Research*, pp. 475–479.

Perry, P. (1973), A comparison of the voting preferences of likely voters and likely nonvoters, *Public Opinion Quarterly*, pp. 99–109.

Perry, P. (1979), Certain problems in election survey methodology, *Public Opinion Quarterly*, *43*(3), 312–325.

pewinternet.org (2015), Internet user demographics, `http://www.pewinternet.org/data-trend/internet-use/latest-stats/`, accessed March 25, 2015.

Popova, L., and P. M. Ling (2014), Nonsmokers' responses to new warning labels on smokeless tobacco and electronic cigarettes: an experimental study, *BMC public health*, *14*(1), 997.

Rubin, D. B. (1976), Inference and missing data, *Biometrika*, *63*(3), 581–592.

Särndal, C.-E. (2007), The calibration approach in survey theory and practice, *Survey Methodology*, *33*(2), 99–119.

Särndal, C.-E., and J.-C. Deville (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, *87*(418), 376–382.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1989), The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, *Biometrika*, *76*(3), 527–537.

Schonlau, M., A. Van Soest, A. Kapteyn, and M. Couper (2009), Selection bias in web surveys and the use of propensity scores, *Sociological Methods & Research*, *37*(3), 291–318.

Schonlau, M., et al. (2004), A comparison between responses from a propensity-weighted web survey and an identical rdd survey, *Social Science Computer Review*, *22*(1), 128–138.

Silver, N. (2015), How fivethirtyeight calculates pollster ratings, `http://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/`, accessed October 10, 2015.

Singh, A. C., B. Kennedy, and S. Wu (2001), Regression composite estimation for the canadian labour force survey with a rotating panel design, *Survey Methodology*, *27*(1), 33–4W.

Sirken, M. G. (1970), Household surveys with multiplicity, *Journal of the American statistical Association*, *65*(329), 257–266.

Squire, P. (1988), Why the 1936 literary digest poll failed, *Public Opinion Quarterly*, *52*(1), 125–133.

Stephan, F. F. (1948), History of the uses of modern sampling procedures, *Journal of the American Statistical Association*, *43*(241), 12–39.

Taylor, H. (2000), Does internet research work? comparing online survey results with telephone survey, *International journal of market research*, *42*(1), 51–63.

Terhanian, G., and J. Bremer (2012), A smarter way to select respondents for surveys?, *International Journal of Market Research*, *54*(6), 751–780.

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

Tikhonov, A. N. (1943), On the stability of inverse problems, *Doklady Akademii Nauk SSSR*, *39*(5), 195–198.

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

Valliant, R., and J. A. Dever (2011), Estimating propensity adjustments for volunteer web surveys, *Sociological Methods & Research*, *40*(1), 105–137.

Wang, H., and C. Leng (2008), A note on adaptive group lasso, *Computational Statistics and Data Analysis*, *52*, 5277–5286.

Wang, L., and J. Zhu (2010), Financial market forecasting using a two-step kernel learning method for the support vector regression, *Annals of Operations Research*, *174*(1), 103–120.

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2014), Forecasting elections with non-representative polls, *International Journal of Forecasting*.

Wing, I. S., and J. L. Walker (2010), The geographic dimensions of electoral polarization in the 2004 us presidential vote, in *Progress in Spatial Analysis*, pp. 253–285, Springer.

Witten, D. M., and R. Tibshirani (2009), Covariance-regularized regression and classification for high dimensional problems, *Journal of the Royal Statistical Society: Series B*, *71*(3), 615–636.

Wu, C., and R. R. Sitter (2001), A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, *96*(453), 185–193.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009), Genome-wide association analysis by lasso penalized logistic regression, *Bioinformatics*, *25*(6), 714–721.

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *The Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006), The adapptive lasso and its oracle properties, *Journal of the American Statistical Association*, *101*(476), 1418–1429.