Target Selection as Variable Selection: Using the Lasso to Select Auxiliary Vectors for the Construction of Survey Weights*

Devin Caughey
MIT

Erin Hartman
UCLA

June 30, 2017

Abstract

Survey nonresponse is a ubiquitous problem in modern survey research. As individuals have become less likely to respond to surveys there has been a simultaneous rise in highly granular data sources that can be used to help ameliorate the nonresponse problem. While much research has been done on post-hoc weighting methods, which provide a flexible and general solution for unit nonresponse, there is an open question of how to select the optimal auxiliary vector to include in the weighting method. We formulate this as a methodological question of variable and interaction selection where the goal is, assuming an individual level stochastic response probability, to construct an optimal set of weights for each individual respondent to account for an observed pattern of nonresponse. We use recent literature on hierarchical group-lasso regularization to determine the best auxiliary vector for weighting. We show the advantages of this method in simulations that are derived from real survey data sampled off of an individual level voter file in recent elections. We also apply the method to historic quota sampled survey data from the 1930s and 1940s to show the advantages of this method even where the sampling design is unknown.

*For comments and assistance, we thank Adam Berinsky, Sara Chatfield, Liz Gerber, Eric Schickler, Jasjeet Sekhon, Mallory Wang, Yiqing Xu, and Teppei Yamamoto. We are thankful for the comments from the participants of the 2016 MPSA conference, 2017 Pacific Coast American Politics Conference, as well as the Washington University at St. Louis American Politics and Stanford American Politics workshops. We also gratefully acknowledge the willingness of BlueLabs to share their data with us.

Contents

1 Introduction

Perhaps the most important recent trend in survey research is the declining probability that sampled individuals will respond to surveys. Response rates have declined in both face-to-face and telephone surveys, in those conducted by academic institutions, non-profits, and government agencies as well as by private firms (Curtin, Presser, and Singer 2005; Bethlehem, Cobben, and Schouten 2011). Even high-quality telephone surveys like those conducted by the Pew Foundation have seen their response rates drop below 10% (Pew Research Center 2012), and the rates for less-rigorous surveys are assuredly even lower.

The collapse in survey response rates raises the corresponding risk of nonresponse bias. Nonresponse bias arises when there are systematic differences between the target population and the individuals who provide valid responses. If willingness or availability to participate in surveys is correlated with personal characteristics, such as socioeconomic status (SES) or political engagement, then the scope for nonresponse bias increases as response rates fall. Moreover, when 9 out of 10 sampled individuals refuse to participate, traditional statistical techniques predicated only on design-based sampling probabilities lose much of their credibility. Ironically, low response rates have led some survey organizations to revisit techniques from the early days of survey research, such as the quota-sampling techniques American surveys abandoned decades ago in favor of probability samples (Berinsky 2006). Additionally, there has been a rise in the use of high quality convenience panels. It is hard to say, *a priori*, if these convenience and quota-based samples are more or less representative than probability samples with high non-response, but the techniques used

to address the issue of representativeness are similar.

A variety of techniques have been developed to address nonresponse bias, but by far the most common is survey weighting. Weighting respondents to "look like" the target population in specified respects can substantially reduce, if not completely eliminate, nonresponse bias and other discrepancies from the population (Särndal and Lundstrom 2005). Weighting is appealing because the weights can be calculated once and passed on to users to incorporate into their preferred analysis method. Weights are also nonparametric in the sense that they do not require an explicit model of the outcome variable.[1] Nevertheless, weighting is effective at reducing nonresponse bias only to the extent that individuals with the same weight are homogenous with respect to the outcome variable, the response probability, or both. This is likely to hold only if the auxiliary variables used to create the weights are powerful predictors of outcome variables and/or response probabilities.[2]

The need for powerful auxiliary variables dovetails with another recent trend, the increasing availability of auxiliary information on target populations (Smith 2011). Governments in some European countries, for example, keep detailed administrative data on their citizens, which government statistical agencies can link with the individuals sampled in official surveys.[3] Polling firms and political campaigns can now purchase access to commercial databases with hundreds of variables on citizens' political, economic, and demographic characteristics (West et al. 2015). The cov-

1. This stands in contrast to model-based alternatives, such as (multiple) imputation and full-information maximum likelihood (Schafer and Graham 2002, 157).

2. An auxiliary variable is one whose value is known for each respondent and whose population distribution is at least partially known.

3. See, for example, Särndal and Lundstrom (2005, 22) on Sweden and Finland and Bethlehem, Cobben, and Schouten (2011, 258) on the Netherlands.

erage and specificity (if not necessarily the accuracy) of this auxiliary information continues to increase. The question is, what is the best way to make use of this information?

Perhaps the best-known method for calculating weights is poststratification, which involves classifying units into cells based on a complete cross-classification of categorical auxiliary variable and weighting each cell to match its proportion in the population. Poststratification is appealing due to its transparency, non-parametric nature, and the fact that poststratified samples will, by construction, exactly match the auxiliary variables' joint distribution in the population. But as the number of auxiliary variables increases, so does the probability that at least one cell will be empty in the sample, rendering poststratification infeasible. Faced with empty cells, survey analysts must either create larger cells by dropping auxiliary variables or else turn to an alternative weighting technique.

The most common alternative to poststratification is raking, which generates weights that match the marginal (as opposed to joint) distributions of the auxiliary variables (Deming and Stephan 1940). Unlike poststratification, which requires only that data be missing at random within cells, raking weights are valid only under a parametric model of nonresponse, specifically that the interior cell proportions are a log-linear function of the marginals (Little and Wu 1991). Raking and poststratification, however, are only two extreme alternatives, between which lie a multitude of other weighting options that can be subsumed under the general framework of calibration estimation (Deville and Särndal 1992).

Calibration entails finding the set of weights (if any) that ensure that the weighted

sample matches a set of population targets while also differing as little as possible from design-based prior weights. The targets may include only the marginal distributions of the auxiliary variables, as in raking, or they may consist of their complete joint distribution, as in poststratification. But they may also be any arbitrary combination of marginal and joint distributions—say, the marginal distribution of variable V, the two-way interaction of W and X, and the three-way interaction of X, Y, and Z. The number of possible targets increases exponentially in the number of auxiliary variables, and is astronomically vast for administrative databases and other auxiliary datasets that may contain hundreds of variables.

Our focus in this paper is on how to select from among the myriad possible target benchmarks the subset that lead to calibration weights that most effectively ameliorate the problem of nonresponse. We formulate this as a problem of variable subset selection, where the goal is to select the subset of auxiliary variables and their interactions that best predicts response probabilities and outcome variables. Because the selected targets can simply be used as inputs to standard calibration software (e.g., Lumley 2004), this approach is easy to implement, and it is also highly interpretable, since the resulting weights can be described to others in terms of population targets matched in the sample.

Formulating target selection as variable selection naturally suggests the use of the lasso as a computationally efficient method of variable selection (Tibshirani 1996). Building on this insight, we propose a two-stage procedure for selecting population tagets with the lasso. The first stage is to employ an appropriate version of the lasso to select the optimal variable subsets for ascending levels of model complexity.

The second is to identify the most complex (least regularized) variable subset that leads to feasible weights.[4] We illustrate the performance of this procedure with two examples, one based on contemporary voter-file data and the other on quota-sampled opinion polls from the 1930s–50s.

2 Calibration Weighting

Calibration constructs weights that are "calibrated" to known population benchmarks, such that the weighted probability sample, s, exactly matches the target population U in specified respects.[5] The goal of calibration is to find the set of respondent weights $\boldsymbol{w} = (w_1, \ldots, w_i, \ldots, w_n)$ that differ as little as possible from a set of prior weights $\boldsymbol{b}$ while ensuring that the weighted sample matches a set of the target benchmarks $\boldsymbol{t}$.[6] Formally, calibration minimizes the distance

$$\sum_i D(w_i, b_i) \tag{1}$$

subject to the moment constraints

$$t_g = \sum_i w_i x_{ig}, \forall g \in 1 \ldots G, \tag{2}$$

4. We define feasibility as the set of weights, given the calibration distance metric, that can be calculated. For example, for raking this implies convergence and for post-stratification this implies no empty cells among the set of interactions.

5. Calibration for survey weights is closely related to the method of entropy balance for causal inference described by Hainmueller (2012).

6. In general $\boldsymbol{b}$ could consist of any prior weights (e.g., design-based inverse-probability weights), but since in our application we lack prior information about the sampling process we set $b_i = 1 \; \forall i$.

where the auxiliary vector $\boldsymbol{x}_i = (x_{i1}, \ldots x_{ig}, \ldots x_{iG})'$ encodes the sample information corresponding to the target benchmarks $\boldsymbol{t} = (t_1, \ldots t_g, \ldots t_G)$. When the auxiliary variables are all categorical, as is the case with the applications we consider, $\boldsymbol{x}_i$ consists of a series of dummy variables indicating whether respondent i belongs to a set of G (possibly overlapping) population groups, and $\boldsymbol{t}$ consists of the population totals for the groups. Suppose, for example, there are three groups: women, people under 40, and women under 40. Then the auxiliary vector for a man under 40 would be $\boldsymbol{x}_i = (0, 1, 0)'$, and $\boldsymbol{t}$ would be the three groups' sizes in the population.

Since there is no unique way to measure the distance of $\boldsymbol{w}$ from $\boldsymbol{b}$, calibration weights depend on the choice of the distance metric $D(w_i, b_i)$. The two most weighting methods–raking and linear weighting–correspond to two different distance metrics. Raking minimizes the cross-entropy, $w_i \log(w_i/b_i)$. Linear weights, for which post-stratification is a special case in which the calibration vector is composed of dummy variables corresponding to mutually exclusive strata–minimize the Euclidean distance, $(w_i - b_i)^2$. Different distance metrics have distinct advantages—raking, for instance, guarantees that all weights are strictly positive whereas linear weights can result in negative weights—and each is exactly unbiased under a slightly different model of the sampling process (see Little and Wu 1991). For the purposes of reducing nonresponse bias, however, the choice of distance metric is much less important than the choice of auxiliary vector and corresponding population targets (Lumley 2010, 166).

Nonresponse bias in the sample mean of responding units equals

$$\text{Bias}(\bar{y}) = \sigma_\rho \sigma_y R_{\rho y} / \bar{\rho}, \tag{3}$$

where $\bar{\rho}$ is the average response probability ρ in the population, σ_ρ and σ_y are the population standard deviations of ρ and y, and $R_{\rho y}$ is the population correlation between ρ and y (Bethlehem, Cobben, and Schouten 2011, 249). Weights reduce nonresponse bias to the extent that they break the correlation between response probabilities and the outcome variable of interest. As an example, take the case of poststratification, which is equivalent to calibration using an auxiliary vector composed of dummy variables for a set of exhaustive mutually exclusive population categories, such that one element of $\boldsymbol{x}_i$ is equal to 1 and the rest are 0. In this case, the optimal weights are $w_i^{\text{PS}} = \pi_{g[i]}/p_{g[i]}$, where $\pi_{g[i]}$ and $p_{g[i]}$ are respectively the population and sample proportions of individual i's group, $g[i]$. The poststratification estimator $\bar{y}_{\text{PS}} = \sum_g w_g \bar{y}_g / \sum w_g$ (i.e., the weighted average of the group-specific averages) is consistent and nearly bias-free if, within groups, response probabilities ρ_i are uncorrelated with y_i.[7] The correlations will be zero if either ρ_i or y_i is constant within groups—that is, if either is perfectly predicted by the group identifiers in $\boldsymbol{x}_i$ (Särndal and Lundstrom 2005, 93).

7. For the remainder of the paper we assume that the prior weights $b_i = 1 \forall i$, as would be true in a simple random sample or in sample for which the sampling probabilities are not known. If design weights are not equal within groups, $\bar{y}_g$ should be a weighted average. We say "nearly bias-free" because an exact formula for the bias is not obtainable for an arbitrary auxiliary vector. It is, however, possible to derive a close approximation, called the "nearbias" by Särndal and Lundstrom, which converges to the exact bias as the sample size increases. The approximation stems from the finite-sample discrepancy between the sample and population coefficients for the regression of y_i on $\boldsymbol{x}_i$, which is of order in probability $O_p(m^{-1/2})$ and converges to 0 as the number of respondents m tends to ∞ (Särndal and Lundstrom 2005, 106–8).

This result can be generalized to calibration weights based on any auxiliary vector $\boldsymbol{x}_i$. Calibration is often conducted using the generalized regression estimator (GREG) (Deville and Särndal 1992)

$$\bar{y}_{\mathrm{W}} = \sum_i w_i \bar{y}_i / \sum w_i = \sum_s d_i y_i + (\sum_s w_i \mathbf{x}_i - \sum_s b_i \mathbf{x}_i) * \frac{\sum_s b_i \mathbf{x}_i \mathbf{x}_i'}{\sum_s b_i \mathbf{x}_i \mathbf{y}_i} \tag{4}$$

While the assumption that the outcome is linear in $\mathbf{x_i}$ is strong, this formulation provides a clear interpretation for the conditions under which calibration reduces bias in estimates of the outcome. The approximate bias of this calibration estimator $\bar{y}_{\mathrm{W}} = \sum_i w_i \bar{y}_i / \sum w_i$ is

$$\text{nearbias}(\bar{y}_{\mathrm{W}}) = -N^{-1} \sum_{i \in U} (1 - \rho_i) e_{\rho,i} \tag{5}$$

where N is the population size and $e_{\rho,i} = (y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_\rho)$ is the residual from a ρ_i-weighted least-squares regression of y_i on $\boldsymbol{x}_i$ (Särndal and Lundstrom 2005, 99). If every unit has the same response probability, then the sample average is unbiased for the population mean and (5) reduces to

$$\text{nearbias}(\bar{y}_{\mathrm{W}}) = -N^{-1}(1 - \rho) \sum_{i \in U} e_i, \tag{6}$$

which equals 0 since the equally weighted residuals e_i sum to 0.

More relevant to survey weighting are two other sufficient conditions for zero nearbias. The first is if the "response influence" $\omega_i = 1/\rho_i$ has a perfect linear

relationship with the auxiliary vector:

$$\omega_i = \boldsymbol{\gamma}'\boldsymbol{x}_i \ \forall i \in U. \tag{7}$$

This condition is a generalization of Little and Wu's (1991) demonstration that raking weights are MLEs for a sampling model in which the interior cell proportions are a log-linear function of the marginal proportions.[8] It is also closely related to the "missing-at-random" condition invoked in the literature on missing data (e.g., Rubin 1976). The second condition is if the outcome variable has a perfect linear relationship with the auxiliary vector:

$$y_i = \boldsymbol{\beta}'\boldsymbol{x}_i \ \forall i \in U. \tag{8}$$

If this relationship holds, then $e_{\rho,i} = 0 \ \forall i \in U$ and thus $-N^{-1}\sum_{i\in U}(1-\rho_i)e_{\rho,i} = 0$ (Särndal and Lundstrom 2005, 101–2). In addition, an auxiliary vector that is highly predictive of y_i tends to reduce variance of the estimator, whereas, conversely, an auxiliary vector that predicts ω_i but not y_i often increases variance without reducing bias (Little and Vartivarian 2005).

For these reasons, methodological texts typically recommend that survey researchers select an auxiliary vector that strongly predicts both response probabilities and outcome variables of interest. Beyond this, however, much of the specific ad-

8. Specifically, Little and Wu (1991) show that four methods of adjusting cell counts in an $I \times J$ table to known margins correspond to the MLEs of four variations on a common model, $g(\pi_{ij}/p_{ij}) = \mu + \alpha_i + \beta_j$, where α_i and β_j are row and column effects related to the marginal proportions. Raking, for example, corresponds to $g(\pi_{ij}/p_{ij}) = \log(\pi_{ij}/p_{ij})$, and least-squares weighting to $g(\pi_{ij}/p_{ij}) = \pi_{ij}/p_{ij}$. All four methods are asymptotically equivalent (see also Deville and Särndal 1992).

vice on choosing auxiliary variables is essentially heuristic. Bethlehem, Cobben, and Schouten (2011, chapter 9), for instance, suggests pre-selecting certain auxiliary variables based on substantive and theoretical knowledge, then selecting the rest based on an ad hoc assessment of the strength of their relationships with nonresponse and the main survey variables.

Only a few formal selection procedures have been proposed. Särndal and Lundstrom (2008), for example, recommend an R^2-like statistic that captures variability in predicted response probabilities, and propose that it be used in a stepwise selection procedure for the "best possible" auxiliary vector. Due to the computational burden, however, Särndal and Lundstrom restrict their attention to "main effects" only—that is, to the marginal rather than joint distributions of the auxiliary variables. Wagner (2012) surveys a wider array of nonresponse indicators, including some that take into account auxiliary vectors' relationship with outcome variables, but does not propose a procedure for using them to select an auxiliary vector. Others, such as Andridge and Little (2011), do propose such procedures but focus on maximum-likelihood or other model-based estimators. In short, little attention has been paid to developing computationally feasible procedures for auxiliary-vector selection that take into account both response probabilities and outcome variables.

3 Target Selection as Variable Selection

Our proposed procedure for selecting target benchmarks and corresponding auxiliary vectors formulates the task as a problem of variable selection for linear regression.

This follows naturally from that fact that nonresponse bias is eliminated if either ω_i or y_i is linearly related to the auxiliary vector $\boldsymbol{x}_i$—that is, if either $\boldsymbol{\omega} = \boldsymbol{\gamma}'\boldsymbol{x}_i$ or $\boldsymbol{y} = \boldsymbol{\beta}'\boldsymbol{x}_i$. Thus, from a bias-reduction perspective, we should aim to select the $\boldsymbol{x}_i$ that best predicts ω_i, y_i, or preferably both. The relevant regression relationships are those in the population U, not in the set of responding units R. Later we discuss how best to proxy for the population relations, but for now we assume that an appropriate dataset for doing so is available.

If the number of possible auxiliary vectors is small, then this problem can be solved by best subset selection, which searches over all possible auxiliary vectors and identifies the one with the highest R^2 (or other fit criterion). Enumerating all possible vectors, is computationally impossible for $p > 40$ (James et al. 2013, 207), a limit quickly reached when interactions are considered and there are more than a few auxiliary variables. Suppose, for example, that there are 6 auxiliary variables, each with 2 levels, and that all possible 2-, 3-, 4-, 5-, and 6-way interactions are considered. Including interactions, the total number of parameters is $p = 2^6 - 1 = 63$, and the total number of possible specifications is $2^{63} \approx 9.2 \times 10^{18}$.

Since best subset selection is infeasible with more than a few auxiliary variables, some shortcut procedure must be used instead. Among the various options, the most attractive is the lasso, which is computationally efficient, can handle $p \gg n$, and provides a good approximation to best subset selection.[9] As its name suggests, the lasso ("least absolute shrinkage and selection operator") is a shrinkage estimator that also acts as a variable selector (Tibshirani 1996). The lasso regression estimator

9. Other options include forward- and backward-stepwise selection (Hastie, Tibshirani, and Friedman 2009, 58–60).

can be written as

$$\hat{\beta}^{\text{lasso}} = \text{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} |\beta_j| \right\}, \tag{9}$$

where λ is a tuning parameter regulating how much the coefficients are shrunk away from the least-squares estimates and towards zero. In other words, λ determines the simplicity of the model specification. Like ridge regression and other shrinkage estimators, the lasso has lower variance and is less prone to overfitting than least squares. Unlike ridge regression, the lasso's L_1 penalty $\sum_{j=1}^{J} |\beta_j|$ causes some coefficients to be shrunk all the way to zero—that is, to be dropped from the regression specification (Hastie, Tibshirani, and Friedman 2009, 68–9). The coefficients that are left in the model are collectively the most predictive of the outcome for a given level of model complexity.

Another advantage of the lasso is that it is easy to extend and modify it in useful ways. For example, the variable-specific penalties can also be modified so as to require the inclusion of certain variables in every variable subset. This is useful for target selection because it is often the case that for the analyst wishes to ensure that some benchmarks are matched—for example, because they define population domains of interest (Särndal and Lundstrom 2005, 22) or because they are known *a priori* to be important predictors of nonresponse (Bethlehem, Cobben, and Schouten 2011, chapter 9). The lasso can also be modified so that the coefficients are grouped, which is helpful in (at least) two respects. First, it makes it possible to require that if any interaction involving an auxiliary variable is included, so is the variable's

main effect. Second, grouping provides a way to implement lasso for multivariate regression, such as one in which the multivariate regression is $\boldsymbol{y}_i = (\omega_i, y_{1i}, \ldots y_{Ji})$, where there may be $J > 1$ outcome variables of interest.

4 Target Selection with the Lasso

Recall that our goal is to generate weights that most reduce nonresponse bias. Practically, this entails selecting the target benchmarks $\boldsymbol{t}$ and corresponding auxiliary vector $\boldsymbol{x}_i$ that best predict ω_i and y_i, such that the moment constraints in (2) can actually be satisfied (i.e., such that the weights can actually be created). Analysts may also wish to impose additional standards on the weights beyond mere feasibility. They may, for example, constrain the weights to lie within a specified interval so as to avoid extreme weights (e.g., Bethlehem, Cobben, and Schouten 2011, 237–8).[10] The goal, then, is to find the most predictive auxiliary vector that satisfies these constraints.

The most complex auxiliary vector that can be derived from a given set of L auxiliary variables is one composed of the variables' main effects plus all possible 2- through L-way interactions among them. This is equivalent to poststratification on all auxiliary variables. The simplest auxiliary vector, which matches the population total only, contains only an intercept: $x_i = 1 \; \forall i$. Between these extremes of most and least complex lie a very large number of intermediate cases. In the lasso, the regularization parameter λ determines where a specification falls on this spectrum.

10. These restrictions can also be imposed through the distance metric (e.g., by using logit calibration; Deville, Särndal, and Sautory 1993).

A large value of λ heavily penalizes coefficients, leading to a relatively simple model (and, in the extreme, one with only an intercept). A small value of λ will yield a relatively complex model—that is, one with many terms—and converges to the unpenalized regression model as $\lambda \to 0$.[11] Thus, by running the lasso using a sequence of λ values, one can select an optimal specification for a given level of model complexity. This suggests the following procedure for identifying the optimal auxiliary vector $\boldsymbol{x}_i^*$:

1. Select a value of λ_0, an initial candidate value for λ that corresponds to the least penalized/most complex weighting model.

2. Using the (multivariate) lasso with $\lambda = \lambda_0$, estimate the model $\boldsymbol{y}_i = \boldsymbol{z}_i\boldsymbol{B} + \boldsymbol{e}_i$, where $\boldsymbol{y}_i$ is a possibly multivariate response, $\boldsymbol{B}$ is a matrix of coefficients, and $\boldsymbol{z}_i$ consists of the entire pool of auxiliary variables and all possible interactions.

3. Construct an auxiliary vector $\boldsymbol{x}_{0i}$ composed of those elements of $\boldsymbol{z}_i$ with nonzero values of $\hat{\boldsymbol{B}}_0$, the coefficient estimates from the lasso model with $\lambda = \lambda_0$.

4. Attempt to calibrate the sample of interest subject to the target benchmarks $t_g = \sum_i w_i x_{0ig}$.

   (a) If calibration is feasible and the weights satisfy other user-specified standards, identify $\boldsymbol{x}_{0i}$ as the optimal auxiliary vector $\boldsymbol{x}_i^*$ and end the procedure.

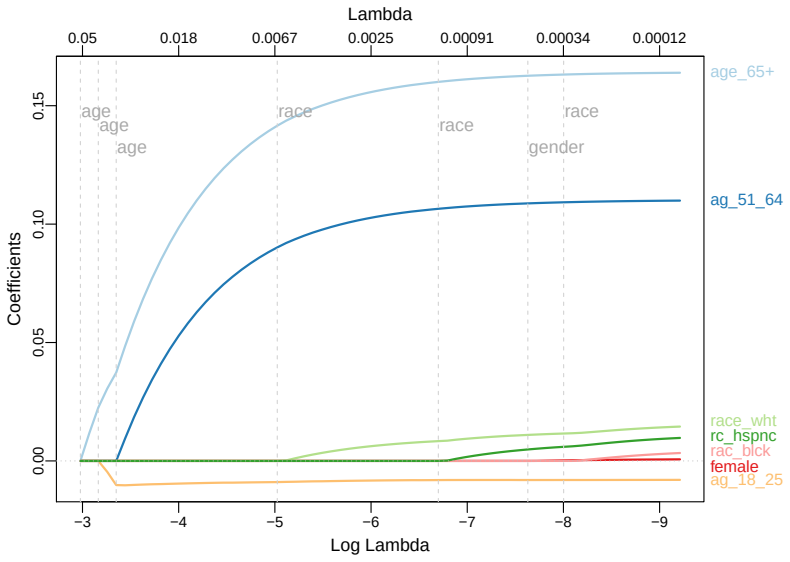
11. The unpenalized model with include all possible terms only if X is full rank; otherwise collinear terms will be dropped.

(b) If calibration is not feasible, repeat steps 1–4 with $\lambda_1 = \lambda_0 + \delta$ and continue until a feasible auxiliary vector is identified.

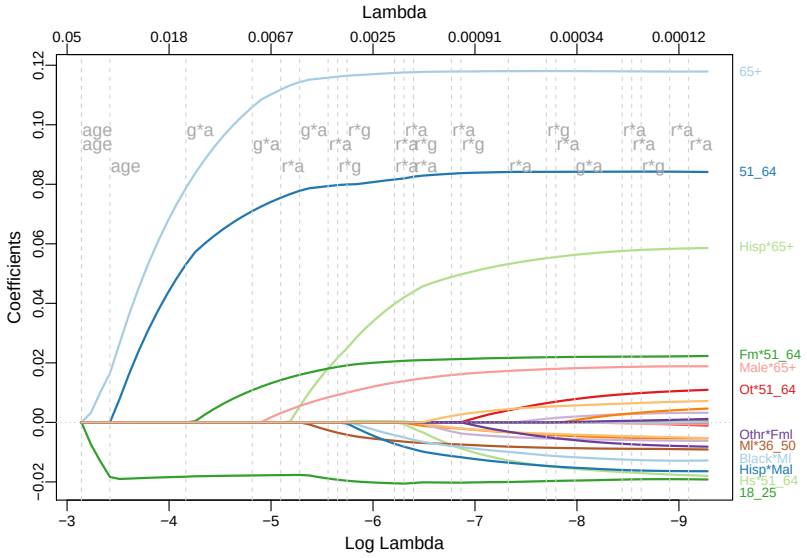
An example of how the Lasso can be used as target selection is presented in Figure 1, which target selection using simulated individual level response data. In Figure 1(a), only main margins of three categorical variables: age (5 classifications), gender, and race (3 classifications). In Figure 1, all main margins and two-way interactions of the aforementioned variables were included. The far right of the graphs shows the least penalized/most complex model, with increasingly penalized models described in the generalized procedure above shown moving left. The figure shows the output of the lasso procedure, and indicates two main ways that the researcher can interpret the result. The researcher could include all cells that are non-zero in a given penalized regression, such as only adjusting for $race =$ '$white$' or $age =$ '65+'. Alternatively, the researcher could include all cells for a variable who has any non-zero cells. For example, weighting to all age margins once any individual age cell is included. The categories are labeled by the gray vertical lines.

The stylized procedure above glosses over the question of what data to use to estimate the lasso. This issue is somewhat problem-specific. One common scenario is one where the set of respondents R is a subset of a population U for which auxiliary data are available every member, responding or not. Suppose that one wishes to select the auxiliary vector based only on its ability to predict nonresponse. In that case, one simply follows the lasso-based target-selection procedure described above, setting $\boldsymbol{y}_i = r_i = 1_{i \in R}$.

Alternatively, one can estimate an aggregated version of the same model. This



(a) Lasso target selection with only main margins



(b) Lasso target selection with main margins and two-way interactions

Figure 1: Examples of Lasso Target Selection
Colored lines indicate non-penalized regressors to include corresponding to an individual cell to adjust. Gray vertical bars indicate the associated variables.

entails classifying all population units $i \in U$ into K cells defined by the cross-classification of all auxiliary variables and calculating each cell's proportion in the response set (p_k) and in the population (π_k). One then sets $\boldsymbol{y}_k = p_k/\pi_k = \hat{\rho}_k$ (the estimated average response probability in cell k) and estimates a weighted lasso, where the weights are proportional to the cell population frequencies.[12] The aggregated approach is also useful when one does not have individual-level data on the population, but only information at the cell level.[13]

Optimizing with respect to outcome variables as well as response probabilities presents several complications. The first is that survey outcomes are, of course, observed for the response set only. One solution to this problem is to run the target selection procedure using only the response set (or, in the aggregated model, cells with at least one respondent). The downside of such an approach is that cells where nonresponse bias is greater are more likely to be empty in the sample and thus dropped from the data used to estimate the lasso. An alternative is to proxy for outcomes of interest using auxiliary variables. For example, if a political survey contains data on party registration available for both respondents and nonrespondents, then party could be predicted in the multivariate lasso along with the response probability. The obvious downside to this is that if party is a good proxy for other outcomes of interest, then it should be included in the auxiliary vector itself. A potential way

12. This model corresponds closely to the "maximum likelihood under random sampling" model described by Little and Wu (1991), under which the response probabilities in each cell are a linear combination of marginal effects. It is also an inverted form of (7), which is expressed in terms of $\omega_k = \rho_k^{-1}$. This inversion avoids the problem of empty cells in the denominator of $\hat{\omega}_k = \pi_k/p_k$.

13. A disadvantage of the aggregated model is that proper inference requires that the degrees-of-freedom be adjusted manually because most software does treats the weights as precision rather than frequency weights.

to avoid these tradeoffs is to combine the lasso with multiple imputation (Sabbe, Thas, and Ottoy 2013), but at present software for doing so is not easily available.

In addition to which outcome variables and training data to use, the lasso-based target selection requires at least two other choices on the part of the analyst. The first is the starting value λ_0. Here, the variance-reducing properties of the lasso come into play along with its role as a subset selector. It might be wise to determine the value of λ with the lowest cross-validation error for the lasso model itself (see James et al. 2013, 227–8) and set λ_0 equal to that value. Even if a more complex set of weights is technically feasible, weights based on the cross-validated λ will likely have lower mean-square-error (see Little and Vartivarian 2005, on bias-variance trade-offs in the selection of weights).

A second issue is involves the penalties for coefficients in the lasso model. Using a grouped lasso, for example, it is possible to link the penalties for multiple coefficients together so that they are selected or not selected together (Hastie, Tibshirani, and Friedman 2009, 90–1). An example of a situation where this might be useful is with a set of dummy variables that represent different levels of the same categorical variable. A more elaborate version of the grouped lasso is the lasso for hierarchical interactions (Bien, Taylor, and Tibshirani 2013). This variant of the lasso enforces that requirement that if an interaction term is chosen, then so must its main effect. This requirement, known as the "marginality principle," leads to more interpretable models. Finally, for reasons of interpretability, domain estimation, or prior theoretical knowledge, it is often useful to force the inclusion of some coefficients in the lasso model, which can be done by setting their penalty term to 0.

5 Simulations of Lasso-Based Target Selection

To demonstrate the effectiveness of using the Lasso to select the auxiliary vector, we conducted simulations comparing post-stratification, raking on main margins, and the lasso procedure using both linear and raking distance functions. The simulations select across an expected final sample of $n = \{200, 500, 2000\}$, where the response propensity is modeled in a sample of $10 \times n$ is drawn from a universe of $N = 1,000,000$, and the average response propensity is 10%. The number of non-zero coefficients in the response propensity is held constant at eight, which are used to determine the response propensity using an inverse logit, but the number of parameters includes all main, 2-, and 3-way interactions of $P = \{10, 15, 20, 30\}$ independently drawn binomial ($p = 0.5$) variables. This means there are $\{175, 575, 1350, 4525\}$ parameters in the corresponding analyses from which the methods are trying to select an appropriate auxiliary vector. The individual response propensity, p_i, and an indicator for if an individual response, r_i, are calculated as

$$\begin{aligned} p_i = logit^{-1}(&-3.95 - 1.06\mathbf{X}_1 + 1.05X_6 - 1.2\mathbf{X}_1 * \mathbf{X}_7 + 1.25X_3 * X_6 \\ &+ 1.17\mathbf{X}_1 * \mathbf{X}_2 * \mathbf{X}_3 + 1.25X_1 * X_2 * X_8 + 1.5X_2 * X_6 * X_{10} \\ &+ 1.14X_3 * X_6 * X_7) \end{aligned} \tag{10}$$

$$r_i = Bern(p_i) \tag{11}$$

We consider four outcomes: the weighted probability of responding, $p_i = Pr(respond)$, the linear estimator for in the probability of response, Y(high corr), a linear outcome with 20 non-zero coefficients, with three of the same non-zero coefficients used to de-

termine the probability of response, $Y(\text{very low corr})$, and a convex combination of $0.2 * Y(\text{high corr}) + 0.8 * Y(\text{very low corr})$, which we call $Y(\text{low corr})$. The outcomes are calculated as

$$\begin{aligned} Y(\text{very low corr}) = &-1.51\mathbf{X}_1 - 1.66\mathbf{X}_1 * \mathbf{X}_7 + 1.62X_1 * X_8 \\ &- 1.62X_1 * X_{10} + 1.53X_2 * X_4 - 1.62X_4 * X_5 + 2.38X_6 * X_7 \\ &+ 2.85X_9 * X_{10} - 2.1\mathbf{X}_1 * \mathbf{X}_2 * \mathbf{X}_3 - 1.61X_1 * X_7 * X_9 \\ &- 1.6X_1 * X_7 * X_{10} - 1.75X_2 * X_4 * X_5 + 2.02X_2 * X_4 * X_{10} \\ &- 2.62X_2 * X_6 * X_{10} + 2.59X_3 * X_4 * X_6 - 1.89X_3 * X_4 * X_{10} \\ &- 1.53X_3 * X_9 * X_{10} + 1.64X_4 * X_5 * X_8 + 2.31X_5 * X_6 * X_8 \\ &+ 2.27X_5 * X_8 * X_{10} \end{aligned} \tag{12}$$

$$Y(\text{low corr}) = 0.2 * logit(p_i) + 0.8 * Y(\text{very low corr}) \tag{13}$$

$$Y(\text{high corr}) = logit(p_i) \tag{14}$$

with the bold variables corresponding to the non-zero coefficients that overlap between p_i and $Y(\text{very low corr})$. We conduct 504 simulations in for each parameterization.

For the Lasso selection, we use a 5-way cross-validation on a binomial outcome. All weighting is done on those individuals for whom $r_i = 1$, weighted to the margins as defined in the full population. For the "lassorake" method, we conduct raking to the selected targets–where we use the most complex feasible model starting at the λ that minimizes the cross-validation error. For "lassolinear", we use a similar

method, but conduct linear weighting. It is not necessary for the raking and linear weighting algorithms to select the same λ–in practice, linear weighting will select a more complex model. For comparison, we include "rake", in which we conduct a logit regression of whether an individual responded on the non-interacted variables, and rake on the main margins of any variables with $p < 0.05$. For "ps" method, we order, in increasing order, the predictors from the logit regression of whether an individual responded on the non-interacted variables by their p-values. We then conduct post-stratification of the most-significant variables, adding interactions until post-stratification is not feasible due to small cells. Finally, the "naive" method is just the unweighted mean value of the responders, and the "oraclerake" is raking conducted on all the true non-zero predictors of response propensity.

Figure 2 presents results for the different weighting methods on the estimation of p_i across different number of parameters and n. We see that the "lassorake" exhibits lower bias and MSE relative to all methods, except for the "oraclerake" method, indicating that the lasso method does a good job of finding the optimal set of auxiliary targets. Importantly, the "lassorake" even outperform post-stratification. Given that all of the main variables are independent dichotomous variables, with $p = 0.5$, this is the best case scenario for post-stratification because empty cells should be randomly distributed across the population, and yet the lasso method outperforms post-stratification. We see that raking on only the non-interacted variables actually performs worse as power increases, most likely because the we're more likely to weigh on main margins as power increases, but are no more likely to find the important interactions. Finally, we see that linear weighting is not efficient for non-linear

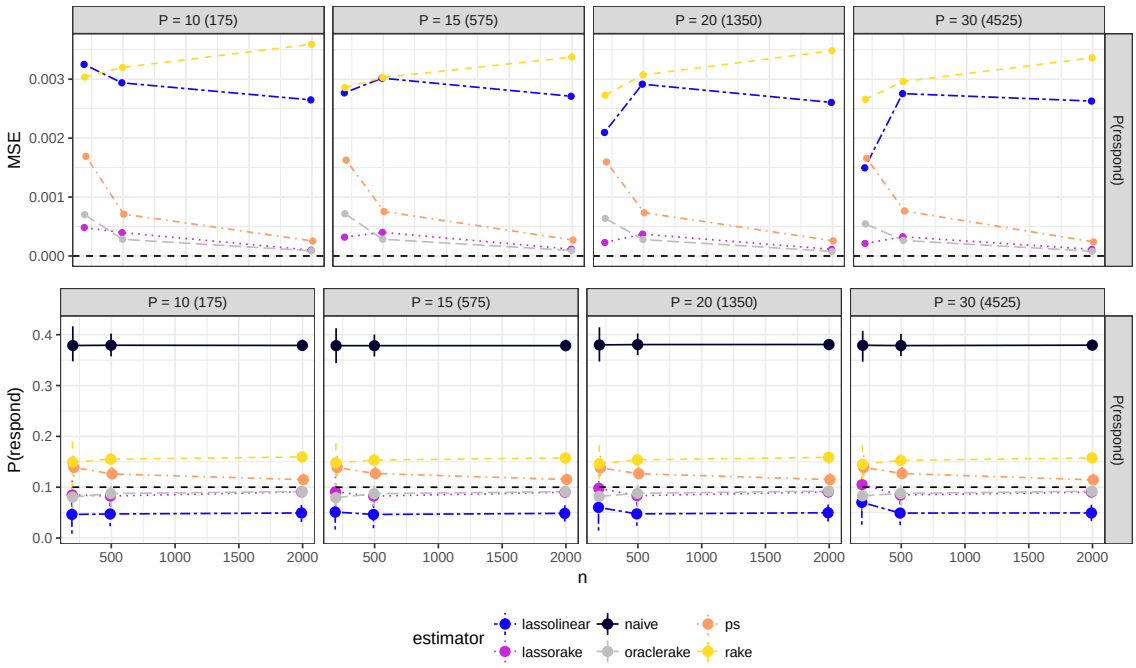


Figure 2: MSE and Bias in estimation of p_i

outcomes, and raking out performs linear weighting.

Turning to the linear outcomes, we present the bias analysis in Figure 3 and the MSE analysis in Figure 4. In this case, where all the outcomes are linear outcomes, the "lassolinear" weighting dominates the "lassorake" in all three outcomes. In the case where the outcome has a very low correlation with the response propensity, we see that raking on the significant main margins narrowly dominates the lasso methods. This is likely because the raking algorithm on significant main margins may correct for variables that are non-zero, or are correlated with non-zero, variables in the true outcome model. We may expect, for example, that using a multivariate lasso would improve the performance of the Lasso methods. As the correlation of the outcome and the response propensity increases, all methods reduce their MSE, and the Lasso methods dominate. Additionally, we see that post-stratification begins to outperform raking, showing the value of considering interactions when adjusting for non-response. We see that, relative to the naive estimate, any sort of principled adjustment reduces bias, often significantly–exemplifying the importance of calibration in survey results.

These simulations show that using the Lasso to select an auxiliary vector can perform as well, and often better, than traditional methods of auxiliary vector selection.

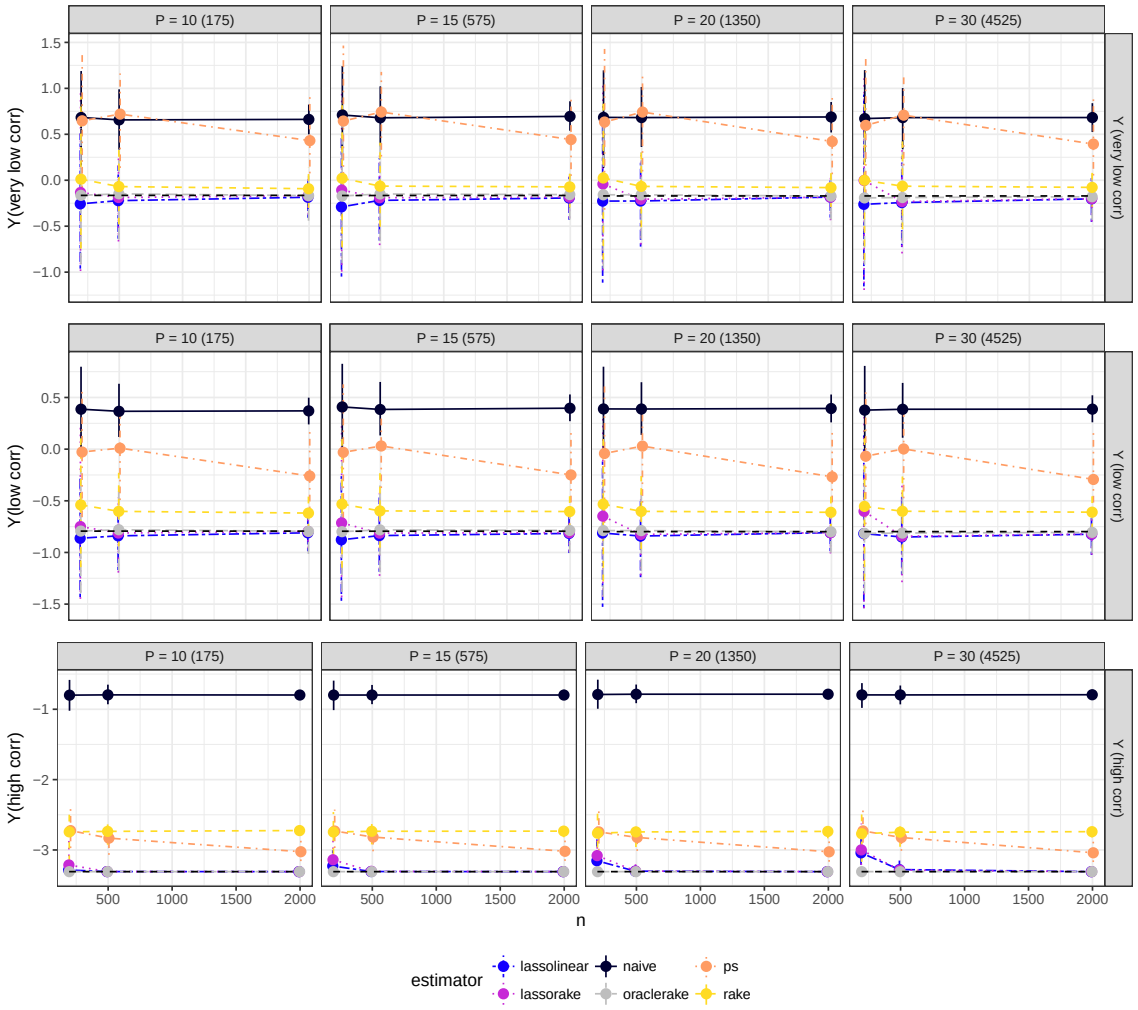


Figure 3: Bias in estimation of Y_i

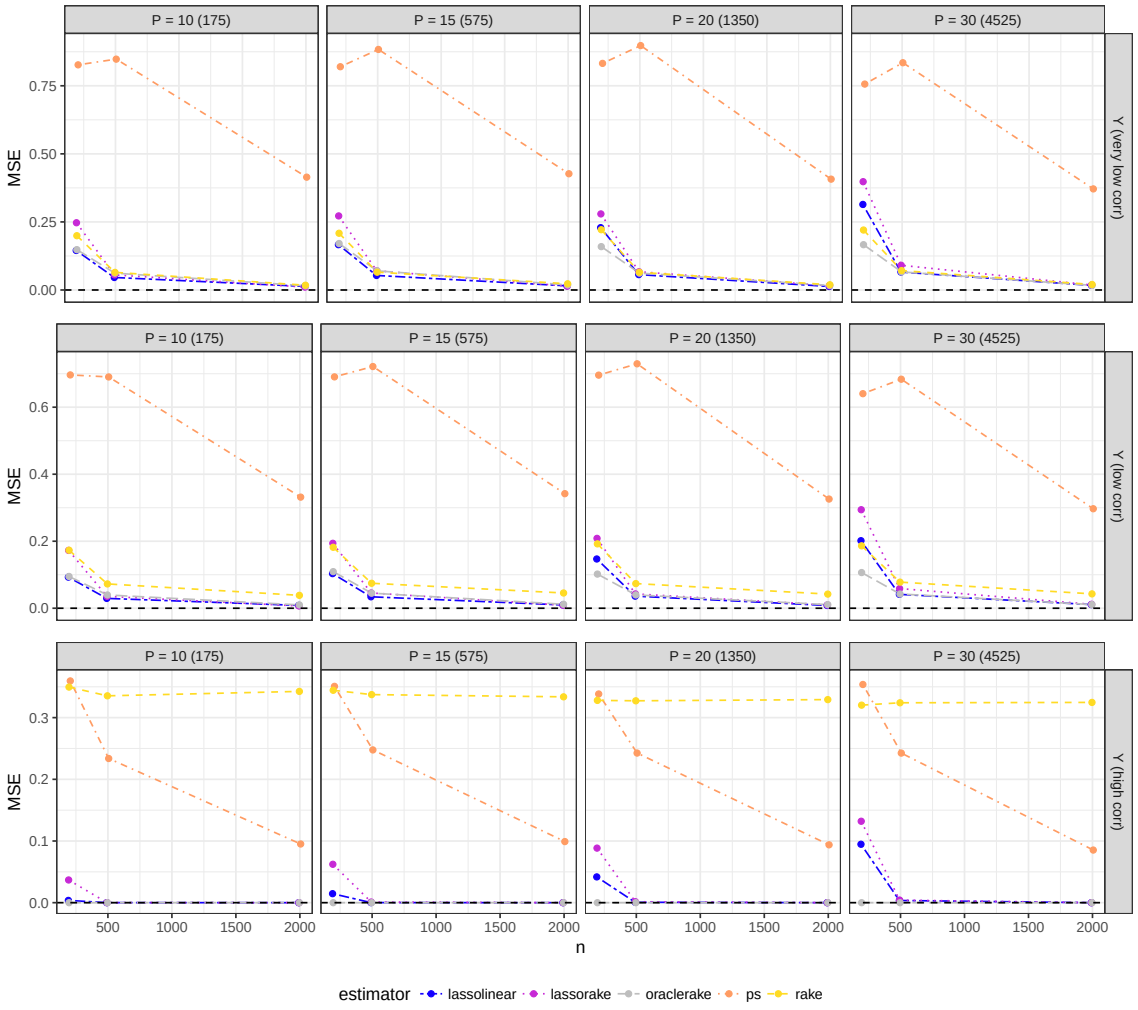


Figure 4: MSE in estimation of Y_i

6 Simulations Based on Voter File

As we have discussed, the decline in response rates has been concurrent with a meteoric rise in auxiliary variables available to researchers. For example, in US politics, companies and organizations have invested in maintaining national voter files that have been merged with many sources of consumer and behavioral data, leading to national databases with thousands of potential auxiliary variables. This has been of great value to researchers, because it means that we know more not only about those individuals who respond to surveys, but also about those individuals who do not. This has allowed for many methodological advancements in the approach to nonresponse, including model-based approaches to data imputation, more sophisticated nonresponse weighting, as well as list-based sampling.[14] The example described in this section uses data from a list-based sample in which the sampling weights are unknown, but for which we have individual level disposition data on if an attempted respondent completed the survey.[15]

Our voter file example comes from data from a survey conducted by a progressive analytics firm.[16] These surveys are short in length, typically comprising of two to five outcomes of interest. We have a random subset of 2,489 attempted contacts from a survey, including 169 individuals who answered and completed surveys. The nonresponse pattern, based on true nonresponse, is unknown. In this case, while the

14. Even internet panel surveys have been able to leverage the power of voter lists, such as the methodology used by organizations such as YouGov / Polimetrix (Rivers and Bailey 2009).

15. This disposition data is based off of short surveys with one primary question of interest, who the respondent intends on supporting on election day, and therefore we focus on completing a survey. However, one could consider modeling different types of nonresponse, such as pick-up rates, partial completion, or full completion of the survey.

16. We thank BlueLabs for providing this data for analysis.

survey is quite small, we also have limited variables available to us for correcting for the nonresponse. In our potential auxiliary vector we have data available on the age (bucketed as: 18-29, 30-39, 40-49, 50-64, and 65+), party registration, three-way ethnicity (black, hispanic, white/other), vote history (midterm voter, presidential voter, new registrant, or sporadic/non-voter), and gender. For our outcome of interest, we use a predicted probability of supporting the democratic candidate, which was built by the analytics firm and which includes the above variables plus many more not available to the researchers.[17]

Population targets for calibration are constructed using a sample of 4,957 individuals from the voter file in the same geography, weighted by the probability of turning out in the forthcoming election.[18] Using the same predicted probability of supporting the Democratic candidate available in our individual response data, we can simulate the outcome of interest, assuming no measurement error. By defining the predictive model as our outcome of interest, we can determine how much error remains in the estimate post-weighting by comparing the adjusted estimate to the true population value. We know that this outcome model is constructed using variables not available to us, meaning that we know that we cannot fully predict the outcome of interest.

Table 1, which presents two outcomes of interest–the expected vote percentage for the Democrat and the expected turnout rate–shows that neither the initial sample nor the final data are representative of the target population. The sample, which is

17. These models have proven very accurate in aggregate in post-election validation.

18. Note that we are treating this population as the target, however in surveys employing such a methodology for creating population targets, the uncertainty in the estimates should be incorporated.

Table 1: Nonresponse Bias in Survey from Voter File

	Target Population	Sample	Respondents
Expected % Vote for Democrat	46.6	50.6	52.4
Expected % Turnout	78.2	63.6	76.4

not a simple random sample from the population,[19] over-represents low propensity voters and voters who are more likely to vote for the Democrat than the expected electorate. The final respondents show an even stronger skew towards the Democrat, although lower propensity voters were also less likely to respond to the survey.

As described in Section 3, researchers must decide both what outcome(s) to model and the most appropriate lasso procedure for the problem at hand. With the data in this example, analysts typically present results for overall estimates as well as across domains (i.e. cross-tabbed by levels of the key auxiliary variables available in our dataset) for multiple outcomes. Because of the way in which the data is analyzed, namely by domains of interest, we chose to use a lasso that ensures balance in the main margins of these domains, but which also ensures balance in interactions of these key domains. We use a hierarchical lasso, which ensures a type of strong hierarchy, namely that all main effects are included for variables in which an interaction is included. This ensures that, by balancing on some interactions, we do not lose balance on the main margins. This also preserves the interpretability of the weights–only interactions that are significant above and beyond the main effects are

19. These samples are typically stratified random samples that oversample proportionally to previously estimated response rates.

included.[20] This ensures that, when an analyst looks at a cross-tab of the outcome of interest by, say, race, that the proportion for black, hispanic, and white/other voters is matched to the population totals.

Additionally, a researcher must decide whether or not to run the lasso procedure to predict the response probability, the outcome of interest, or both. We know that reducing the impact of nonresponse correctly will decrease bias in all outcomes of interest, and so in a case where researchers cannot *a priori* decide which outcome(s) are most important, it may be best to focus on simply the nonresponse. This would be the case for a PI conducting a large survey, for instance. In our case, we simply let $\boldsymbol{y}_i = r_i = 1_{i \in R}$, where we predict a binary outcome of if the attempted respondent completed a survey. This leaves us flexible to analyzing many outcomes of interest.[21] We included all five categorical variables in the algorithm, meaning there are five main margins, and ten sets of interacted margins to choose from.

Simulation Results

Using five-fold cross-validation, the lasso selects four of the five variables for main margins, and three of the ten possible interacted margins. The selected variables are included in Table 2.

As outlined in the two-stage algorithm in Section 3, we begin by testing if the

20. There is some empirical evidence that the main margins of the variables studied here show significant, strong correlations with nonresponse in campaign polling. This substantive knowledge of researchers, even if not fully statistically justified, informs the final analyses of the data, and therefore should be encoded in the weighting method

21. Also, at this time, the `glinternet` package in `R`, which performs the grouped lasso with strict hierarchy, does not allow for multivariate outcomes, however we are currently looking in to reformulating the problem for `glmnet` with grouped constraints. Given the desire to prioritize the hierarchical structure, we proceed with simply modeling the response probability.

Table 2: Cross-validated Variable Selection Results

Main Margins
Party Registration
Age Bucket
Three-Way Ethnicity
Vote History
Interacted Margins
Party Registration × Vote History
Age Bucket × Three-Way Ethnicity
Age Bucket × Vote History

cross-validated $\lambda = \lambda_0$ is a feasible weighting set. The optimal λ results in a feasible set of weights that does not require changing the estimand of interest (i.e. there are no null strata).[22] We use raking on the selected main and interacted margins to construct weights, calibrated to the population defined by the turnout propensity in the target population. We use the predicted propensity to vote for the Democrat, our outcome of interest, to evaluate the degree to which we've recovered the truth.

Table 3 shows two measures of correction using the lasso+raking method. The first column shows all main and interacted margins, with bold indicating the margins selected for weighting by the hierarchical lasso method. The second and third columns show the mean absolute error in the population target, weighted by the true population total by level, both before and after weighting. The fourth and fifth columns show the mean squared error on the outcome of interest (propensity to vote for the Democrat), weighted by the true population total by level, both before and

22. Given the size of the respondent set, we limited ourselves to two-way interactions, however we will explore larger interaction sets in further research.

Table 3: Measures of Correction (Percentage Points)

Variable	Initial Mean Abs. Err. on Margin	Final Mean Abs. Err. on Margin	Initial MSE on Outcome	Final MSE on Outcome
Main Margins				
Party Registration	**6.05**	**0**	**3.21**	**1.40**
Age Bucket	**4.40**	**0**	**13.89**	**1.52**
Three-way Ethnicity	**1.98**	**0**	**15.91**	**2.42**
Gender	1.04	0.41	17.08	1.64
Vote History	**14.13**	**0**	**16.80**	**1.14**
Interacted Margins				
Party Registration × Age Bucket	2.19	0.35	4.62	3.88
Party Registration × Three-way Ethnicity	4.50	0.22	4.58	2.82
Party Registration × Gender	3.16	0.89	4.06	1.84
Party Registration × Vote History	**6.17**	**0.00**	**4.00**	**3.76**
Age Bucket × Three-way Ethnicity	**3.63**	**0.00**	**18.00**	**11.12**
Age Bucket × Gender	2.22	0.14	16.72	6.34
Age Bucket × Vote History	**3.85**	**0.00**	**27.35**	**11.85**
Three-way Ethnicity × Gender	1.00	0.24	18.81	5.23
Three-way Ethnicity × Vote History	11.09	0.34	20.15	8.11
Gender × Vote History	7.07	0.32	18.44	3.80

after weighting.

What we see from Table 3 is that the weighting meets the calibration targets quite closely, even for targets not included in the final weighting set. Where as some population targets were off by upwards of ten percentage points before weighting, no margins, either main or interacted effects, are off by an average of more than a point. The final data looks very representative of the target population on the auxiliary variables we have available. Looking at the outcome of interest, we see that mean squared error, average across cells, is reduced for all main and interacted margins. Overall, whereas the original survey of 169 respondents had a bias of 5.86 percentage points from the truth, the weighted survey is off by only 0.17 percentage

points. The normalized weights are also relatively well behaved, with a minimum weight of 0.19, a maximum weight of 16.16, and 95% of weights below 3.

7 Application to Quota-Sampled Opinion Polls

In this section, we apply lasso-based target selection to a very different setting: quota-sampled opinion polls from the 1930s–50s. Between the first Gallup poll in 1936 the first American National Election Study in 1952, hundreds of non-academic opinion surveys were fielded in the United States (Berinsky 2006; Berinsky et al. 2011). Because these polls were based on purposive quota-controlled samples rather than probability samples, they are biased in systematic ways relative to the general public. Some of these biases were an intentional part of the sampling design. For example, because Gallup surveys were designed to resemble the voting public, most contain few or no black Southerners, who were then largely disfranchised. Other sources of bias, such as the overrepresentation of upper-class and educated respondents, were unintentional.

Following Berinsky (2006) and Berinsky et al. (2011), we using weighting to ameliorate these biases. Fortunately, a great deal of auxiliary information is available to construct these weights. In this example, we have access to the following seven auxiliary variables, all binary unless otherwise indicated: *Black*, *Farm*, *Female*, *Phone Ownership*, *Professional*, *Urban*, and *Region* (Midwest, Northeast, South, and West). Using the data sources and interpolation model described by Caughey and Wang (2014), we derived estimates of the annual population proportions of each of the 256

cells defined by the cross-classification of these variables. We also have a very large database of polls on which to run target-selection procedure, containing over 625,000 respondents in all. As long as the nonresponse mechanism is reasonably stable, using this large training dataset is beneficial because it limits the risk of overfitting in the selection of optimal benchmarks.

The survey outcomes of interest in this example all relate to domestic politics, so it is useful to select auxiliary vectors that predict political outcomes as well as response probabilities. As noted earlier, the cost of doing so is that empty cells must be dropped, but the benefit is a potential decrease in variance as well as bias. To do so, we make use of the multivariate gaussian lasso, as implemented by the R package `glmnet` (Friedman, Hastie, and Tibshirani 2010). One of the few political variables common to all the component polls is retrospective presidential vote, which we recode to three dummy variables: *Voted Democratic*, *Voted Republican*, and *Did Not Vote*.[23] Together, these variables provide a good proxy for respondents' political engagement and orientation.

To proxy for the response probabilities, we constructed postratification weights for every respondent in the training dataset using the auxiliary variable set defined above.[24] Note that this is possible only because we pool all the polls, meaning that only about 12 the 256 cells are empty and must therefore be dropped. We treat each respondent's postratification weight as estimates of the response influence $\hat{\omega}_i$. The three vote variables and the logarithm of the poststratification weights form the

23. A small residual category, those who voted for minor-party candidates, was excluded. As in nearly all other surveys, voting was substantially over-reported in our data.

24. We poststratified the data separately by year and normalized the weights to have a mean of 1 in each year. We ignored population cells that did not appear in the sample.

multivariate response surface for the lasso.[25]

We convert the seven auxiliary variables to nine non-collinear dummy variables.[26] We include all nine in the lasso variable set along with their (non-collinear) two-way, three-way, and four-way interactions, for a total of 182 potential variables. Based on substantive knowledge of the sampling scheme (for details, see Berinsky 2006), we require that any variable subset selected include the main effects of all the auxiliary variables as well as an indicator for Southern blacks. We accomplish this by multiplying the corresponding coefficients in the penalty $\lambda \sum_j |\beta_j|$ by 0.[27]

We specify a grid of 50 values for the shrinkage parameter λ ranging from 10^{-4} to 10^{-1}, equally spaced on the $\log_{10}$ scale, and select a variable subset for each one.[28] The simplest (most regularized) variable subset selected by the lasso includes only the required variables. The number of selected variables increases log-linearly as λ decreases, reaching a maximum value of almost 150 variables at $\lambda = 10^{-4}$.

Based on this ranking of variable subsets, we generate weights for two Gallup polls respectively fielded in February and October 1940.[29] We choose the first because it asked respondents whether they own a car, providing a useful indicator of the class bias in the sample. The second poll is the last one Gallup conducted before the 1940 presidential election, enabling us to compare election predictions under various

25. We take the natural logarithm of the weights to make their distribution more symmetric. In addition, this has the additional benefit of matching the log-linear functional form of the model underlying raking weights.

26. *Region* was decomposed into the indicators *Northeast*, *South*, and *West*, with Midwest as the excluded category.

27. We also require the inclusion of fixed effects for presidential election cycle.

28. To implement the lasso itself, we used the function `glmnet` (Friedman, Hastie, and Tibshirani 2010) in the R computing environment.

29. The polls were AIPO #183 (February 2–7, 1940) and AIPO #219 (October 26–31, 1940).

weighting schemes to the actual result. Each of these poll samples contained close to 3,000 respondents. Since neither includes a single black Southerner, we drop them from the target population.

We attempted to generate raking weights using each lasso variable subset as a set of moment targets.[30] We started with the simplest subset and tried each successive one until weighting proved impossible. Weighting for the February poll failed on the eighth variable subset, and the October poll failed on the seventh.[31] Under the most complex weighting specifications, the largest weights were 4 times larger than the average weight, somewhat greater than is considered ideal (e.g., Deville, Särndal, and Sautory 1993, 1018). The results reported below varied little across lasso-selected weighting specifications, suggesting that extreme weights did not cause much of a problem.

In the February 1940 poll, weighting is quite effective at reducing class bias, at least for the top two-thirds of the SES spectrum, which drives the variation in car ownership. Based on a 1948 probability-sampled consumer survey and yearly numbers on automobile registrations and population (from 1930–50), we calculate a rough population target of 49–52% car ownership in 1940.[32] The unweighted percentage of car owners in the Gallup poll is much higher at 60%, with a standard error of 1%. But weighting the poll to match the most complex (feasible) set of targets yields an estimate of 51%, right in the range of our out-of-sample estimates.

30. We used the `R` function `rake` from the `survey` package (Lumley 2012).

31. The February poll failed because the eighth variable subset included *Farm* × *Professional*, but this cell was empty in the poll sample. The October poll similarly failed due to an empty *Black* × *Phone* × *Urban* cell.

32. Part of the uncertainty stems from having to extrapolate backwards from 1948, and part stems from the removal of Southern blacks from the target population.

The October poll, fielded immediately before the 1940 presidential election, contains a question on presidential vote preference. Given its high correlation with the retrospective vote variable used to select targets, we should expect weights to substantially improve estimates of prospective presidential vote. The results are consistent with this expectation. In the unweighted data, 51% of respondents who expressed a preference said they would vote for the Democratic Franklin Roosevelt—almost certainly an underestimate of the population proportion, given the sample's overrepresentation of Republican-leaning higher-SES respondents. Under any of the six lasso-selected weight sets, the estimate rises to about 56.6%, which is above FDR's ultimate share of 54.7%. However, since we are weighting to a voting-eligible population, which skews more Democratic than Election-Day electorates, we see that the bias is in a direction we would expect. [33]

8 Conclusion

This paper has proposed a method for addressing an increasingly salient issue in survey research: which population targets to match in the construction of calibration weights designed to ameliorate nonresponse bias. Building on the fact that nonresponse bias is eliminated if either inverse response probabilities or the survey outcome is a linear function of the auxiliary vector, we reformulate the problem of target selection as a problem of regression variable selection. We thus turn to the

33. The weighted estimates of FDR's share are higher in all regions, especially outside the South, where the base percentage is higher and the class bias in presidential partisanship less pronounced. Note that the exclusion of Southern blacks from the target population does little to bias the estimates because Southern blacks were excluded from the voting population as well.

lasso as a low-variance and computationally efficient means of selecting the auxiliary vector that best predicts nonresponse and (optionally) survey outcomes of interest. Through two applications, we showed that the basic lasso can be modified in several useful ways depending on the needs of the analyst, and that the weights it selects seem to perform well. In future research, we plan to explore the properties lasso-based target selection more deeply through simulations, with a particular interest how the variance-reducing properties of the lasso carry over into calibration estimators based on lasso-selected weights.

References

Andridge, Rebecca R., and Roderick J. A. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27 (2): 153–180.

Berinsky, Adam J. 2006. "American Public Opinion in the 1930s and 1940s: The Analysis of Quota-Controlled Sample Survey Data." *Public Opinion Quarterly* 70 (4): 499–529.

Berinsky, Adam J., Eleanor Neff Powell, Eric Schickler, and Ian Brett Yohai. 2011. "Revisiting Public Opinion in the 1930s and 1940s." *PS: Political Science & Politics* 44, no. 3 (June): 515–520.

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys.* Hoboken, NJ: Wiley.

Bien, Jacob, Jonathan Taylor, and Robert Tibshirani. 2013. "A Lasso for Hierarchical Interactions." *The Annals of Statistics* 41 (3): 1111–1141.

Caughey, Devin, and Mallory Wang. 2014. "Bayesian Population Interpolation and Lasso-Based Target Selection in Survey Weighting." Paper presented at the Annual Meeting of The Society for Political Methodology, University of Georgia, Athens, GA, July 24.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 69 (1): 87–98.

Deming, W. Edwards, and F. Frederick Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *Annals of Mathematical Statistics* 11 (4): 427–444.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87, no. 418 (June): 376–382.

Deville, Jean-Claude, Carl-Erik Särndal, and Olivier Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88 (423): 1013–1020.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R.* Springer. doi:10.1007/978-1-4614-7138-7.

Little, Roderick J. A., and Mei-Miau Wu. 1991. "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ." *Journal of the American Statistical Association* 86 (413): 87–95.

Little, Roderick J., and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31 (2): 161–168.

Lumley, Thomas S. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9 (1): 1–19.

———. 2010. *Complex Surveys: A Guide to Analysis Using R.* Hoboken, NJ: Wiley.

———. 2012. ***survey**: Analysis of Complex Survey Samples.* R package version 3.28-2.

Pew Research Center. 2012. *Assessing the Representativeness of Public Opinion Surveys.* Technical report. Washington, DC, May 15. `http://www.people-press.org/files/legacy-pdf/Assessing%5C%20the%5C%20Representativeness%5C%20of%5C%20Public%5C%20Opinion%5C%20Surveys.pdf`.

Rivers, Douglas, and Delia Bailey. 2009. "Inference from matched samples in the 2008 US national elections." In *Proceedings of the joint statistical meetings,* 628–639.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–592.

Sabbe, N., O. Thas, and J.-P. Ottoy. 2013. "EMLasso: logistic lasso with missing data." *Statistics in Medicine* 32 (18): 3143–3157.

Särndal, Carl-Erik, and Sixten Lundstrom. 2005. *Estimation in Surveys with Nonresponse.* Hoboken, NJ: Wiley.

———. 2008. "Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator." *Journal of Official Statistics* 24 (2): 167–191.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–177.

Smith, Tom W. 2011. "The Report of the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys." *International Journal of Public Opinion Research* 23 (3): 389–402.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

Wagner, James. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76 (3): 555–575.

West, Brady T., James Wagner, Frost Hubbard, and Haoyu Gu. 2015. "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth." *Journal of Survey Statistics and Methodology* 3 (2): 240–264.