Appendix

1 Determining tuning parameter values for adaptive LASSO

In practice, we do not observe the theoretical value of λ_n, which optimized model fit measures like AIC or BIC, 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 are less data-driven. Some modelers choose one of $\gamma = 0.1, 0.5, 1, 2$. Here we determine (λ_n, γ) through cross-validation as follows:

1. Obtain $\alpha_j = 1/\left|\hat{\beta}_j^{MLE}\right|$.
2. Determine 100 equally spaced values of λ_n based on R *glmnet*'s implementation.
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 $\hat{\boldsymbol{\beta}}$.
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).
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).

2 Derivations of Asymptotic Results in Section 3.3.2

We first derive an asymptotic ECLASSO estimator of total, $\hat{T}_y^{ECLASSO}$. We use this result to show asymptotic unbiasedness of $\hat{T}_y^{ECLASSO}$, and obtain its asymptotic variance.

2.1 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:

1. $\hat{\mathbf{B}} = \mathbf{B} + O_p(1/\sqrt{n})$, $\mathbf{B}$ is the finite-population regression slope estimate of $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$.
2. 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. 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 $N^{-1}\sum_{i\in U} k(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.
4. The Horvitz-Thompson estimators of certain population means computed using $\mathbf{d}^A$ are asymptotically normal.
5. The Horvitz-Thompson estimators of certain population means computed using $\mathbf{d}^B$ are asymptotically normal.
6. $\lambda_n / \left(\sqrt{n}/(\sqrt{n})^\gamma\right) \to \infty \quad \text{and} \quad \lambda_n/\sqrt{n} \to 0$.

The mean in conditions in (4) and (5) 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 in 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 2.1. *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 size. 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 $\boldsymbol{\beta}$, such that $\mathbf{B} \to \boldsymbol{\beta}$. Under conditions (1)-(5), the asymptotic estimated-control calibration estimator of population total is:

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

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

Proof. We begin by deriving the asymptotic model-assisted estimator for a population mean, $\hat{\bar{y}}^{ECMC} = N^{-1}\hat{T}_y^{ECMC}$. By conditions (2) and (3), the second order Taylor series expansion of $\mu(\mathbf{x}_i, \hat{\mathbf{B}})$ around $\mathbf{B}$ is:

$$\mu(\mathbf{x}_i, \hat{\mathbf{B}}) = \mu(\mathbf{x}_i, \mathbf{B}) + \left\{ \frac{\mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\mathbf{B}} \right\}^T (\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \left\{ \frac{\partial^2 \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^T}\Big|_{\mathbf{t}=\mathbf{B}^*} \right\} (\hat{\mathbf{B}} - \mathbf{B}) \tag{1}$$

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

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

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) and (3),

$$max_i \, |\mathbf{h}(\mathbf{x}_i, \mathbf{B})| \leq h(\mathbf{x}_i, \mathbf{B}) \tag{2}$$
$$max_{k,j} \, |\mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)| \leq k(\mathbf{x}_i, \mathbf{B}^*) \tag{3}$$

The population mean of (1) based on sample s_B is:

$$\begin{aligned} N^{-1}\sum_{i \in s_B} d_i^B \mu(\mathbf{x}_i, \hat{\mathbf{B}}) &= N^{-1}\sum_{i \in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i \in s_B} d_i^B \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T (\hat{\mathbf{B}} - \mathbf{B}) + \\ &\quad O_p\left(\frac{1}{\sqrt{n_B}}\right) O_p\left(\frac{1}{\sqrt{n_B}}\right) \\ &= N^{-1}\sum_{i \in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i \in s_B} d_i^B \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T (\hat{\mathbf{B}} - \mathbf{B}) + O_p\left(\frac{1}{n_B}\right) \end{aligned} \tag{4}$$

Similarly, the population mean of (1) based on sample s_A is:

$$N^{-1}\sum_{i \in s_A} d_i^A \mu(\mathbf{x}_i, \hat{\mathbf{B}}) = N^{-1}\sum_{i \in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i \in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T (\hat{\mathbf{B}} - \mathbf{B}) + O_p\left(\frac{1}{n_A}\right) \tag{5}$$

By conditions (1), (4), (5), and equations (4) and (5):

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

where $n^* = min(n_A, n_B)$. Using conditions (1) and (3), we have

$$\begin{aligned}
\bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \hat{\mathbf{B}}) \Big/ \sum_{i\in s_A} d_i^A \\
&= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \Big(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)(\hat{\mathbf{B}} - \mathbf{B})\Big) \\
&= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \Big(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B})\Big) + O_p(1/n_A) \\
&= \bar{\mu} + \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B}) + O_p(1/n_A) \\
&= \bar{\mu} + O_p(1/\sqrt{n_A}) + O_p(1/n_A) \\
&= \bar{\mu} + O_p(1/\sqrt{n_A}) \qquad (7)
\end{aligned}$$

where $\mu_b ar = \sum_{i\in s_A} d_i^A \mu_i / \sum_{i\in s_A} d_i^A$.

Then from (1) and (7), and conditions (1) and (3)

$$
\begin{aligned}
&N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}}) \\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu(\mathbf{x}_i,\mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i,\mathbf{B})(\hat{\mathbf{B}}-\mathbf{B}) + (\hat{\mathbf{B}}-\mathbf{B})^T\mathbf{k}(\mathbf{x}_i,\mathbf{B}^*)(\hat{\mathbf{B}}-\mathbf{B}) - \hat{\mu}\right) \\
&= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + N^{-1}\sum_{i\in s_A}\mathbf{h}^T(\mathbf{x}_i,\mathbf{B})(\hat{\mathbf{B}}-\mathbf{B}) + \\
&\quad N^{-1}\sum_{i\in s_A}(\hat{\mathbf{B}}-\mathbf{B})^T\mathbf{k}(\mathbf{x}_i,\mathbf{B}^*)(\hat{\mathbf{B}}-\mathbf{B}) - O_p(1/\sqrt{n_A}) \\
&= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + N^{-1}\sum_{i\in s_A}\mathbf{h}^T(\mathbf{x}_i,\mathbf{B})(\hat{\mathbf{B}}-\mathbf{B}) + O_p(1/n_A) - O_p(1/\sqrt{n_A}) \\
&= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + O_p(1/\sqrt{n_A}) + O_p(1/n_A) - O_p(1/\sqrt{n_A}) \\
&= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu}) + O_p(1/\sqrt{n_A}) \qquad (8)
\end{aligned}
$$

$$ \qquad (9) $$

Similarly,

$$
\begin{aligned}
N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})^2 &= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + (O_p(1/\sqrt{n_A}))^2 \\
&= N^{-1}\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + O_p(1/n_A) \qquad (10)
\end{aligned}
$$

From (8) and (10) we have:

$$
\begin{aligned}
\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})^2} = \frac{N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})(y_i - \bar{y})}{N^{-1}\sum_{i\in s_A} d_i^A(\hat{\mu}_i - \bar{\hat{\mu}})^2} \\
&= \frac{\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})(y_i - \bar{y}) + O_p\left(\frac{1}{\sqrt{n_A}}\right)}{\sum_{i\in s_A} d_i^A(\mu_i - \bar{\mu})^2 + O_p\left(\frac{1}{n_A}\right)} \\
&\to B^{MC} \quad \text{as } n_A \to \infty \qquad (11)
\end{aligned}
$$

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,

$$
\begin{aligned}
\hat{T}_y^{ECMC} &= N\hat{\bar{y}}^{ECMC} \\
&= N\left(N^{-1}\mathbf{d}^A\mathbf{y} + \left(N^{-1}\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} \mu(\mathbf{x}_i, \mathbf{B}) \right) B^{MC} + o_p\left(\frac{1}{\sqrt{n^*}}\right)\right) \\
&= \mathbf{d}^A\mathbf{y} + \left(\sum_{i\in s_B} d_i^B \mu(\mathbf{x}_i, \mathbf{B}) - \sum_{i\in s_A} \mu(\mathbf{x}_i, \mathbf{B}) \right) B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right) \qquad (12)
\end{aligned}
$$

□

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

Lemma 2.2. *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 same superpopulation model as in Lemma 1.*

Let $\mathbf{B}$ *be the finite-population quasilikelihood estimate of* $\boldsymbol{\beta}$, *such that* $\mathbf{B} \to \boldsymbol{\beta}$, *under conditions (1)-(6), the asymptotic ECLASSO calibration estimator of population total is:*

$$
\hat{T}_y^{ECLASSO} = \sum_{i\in s_A} d_i^A (y_i - \mu_i B^{MC}) + \sum_{i\in s_B} d_i^B \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right) \qquad (13)
$$

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

Proof. Under condition (6), the adaptive LASSO regression satisfies the oracle property through Theorems 1 and 4 in Zou (2006):

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

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 (1), it is asymptotically equivalent to estimated-control model-assisted calibration estimator of total. By Lemma 2.2:

$$\begin{aligned}\hat{T}_y^{ECLASSO} &= \hat{T}_y^{ECMC} \\ &= \sum_{i \in s_A} d_i^A \left(y_i - \mu(\mathbf{x}_i, \mathbf{B}) B^{MC}\right) + \sum_{i \in s_B} d_i^B \mu_i(\mathbf{x}_i, \mathbf{B}) B^{MC} + o_p\left(\frac{N}{\sqrt{n^*}}\right) \quad (14)\end{aligned}$$

□

2.2 Asymptotic expectation of $\hat{T}_y^{LASSO}$

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

Theorem 2.3. *$\hat{T}_y^{ECLASSO}$ 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 $\mathscr{B}$ 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 $\mathscr{B}$. 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, by Lemma 2,

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

$\square$

As long as LASSO regression parameters include the superpopulation parameters (i.e., the correct set of 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.

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

Theorem 2.4. *The asymptotic design variance of* $\hat{T}_y^{ECLASSO}$ *is given by*

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

$$\sum_{i \in s_B} \left(\frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \right)^2 (1 - \pi_i^B) + \sum_{i \in s_B} \sum_{j \neq i} \frac{\pi_{ij}^B - \pi_i^B \pi_j^B}{\pi_{ij}^B} \frac{\hat{\mu}_i \hat{B}^{MC}}{\pi_i^B} \frac{\hat{\mu}_j \hat{B}^{MC}}{\pi_j^B}$$

Proof. We derive the asymptotic linearized variance estimate by taking the variance of equation (13) directly:

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

We derive each component $V_{\mathscr{A}} E_{\mathscr{B}}$, $E_{\mathscr{A}} V_{\mathscr{B}}$ separately, assuming that $\mathscr{A}$ is single-stage probability-based sample:

$$\begin{aligned} V_{\mathscr{A}} E_{\mathscr{B}} &= V_{\mathscr{A}} \left(E_{\mathscr{B}} \left[\sum_{i \in s_A} d_i^A \left(y_i - \mu_i B^{MC} \right) + \sum_{i \in s_B} d_i^B \mu_i B^{MC} \right] \right) \\ &= V_{\mathscr{A}} \left(\sum_{i \in s_A} d_i^A \left(y_i - \mu_i B^{MC} \right) + \sum_{i=1}^{N} \mu_i B^{MC} \right) \end{aligned} \tag{15}$$

$$\begin{aligned} &= \sum_{i \in U} \left(\frac{y_i - \mu_i B^{MC}}{\pi_i^A} \right)^2 \pi_i^A (1 - \pi_i^A) + \\ &\quad \sum_{i \in U} \sum_{j \neq i} \left(\pi_{ij}^A - \pi_i^A \pi_j^A \right) \frac{(y_i - \mu_i B^{MC})}{\pi_i^A} \frac{(y_j - \mu_j B^{MC})}{\pi_j^A} \end{aligned} \tag{16}$$

We use sample estimates for population quantities in (16):

$$\widehat{V_{\mathscr{A}}E_{\mathscr{B}}} = \sum_{i\in s_A}\left(\frac{y_i-\hat{\mu}_i\hat{B}^{MC}}{\pi_i^A}\right)^2(1-\pi_i^A)+$$

$$\sum_{i\in s_A}\sum_{j\neq i}\frac{\pi_{ij}^A-\pi_i^A\pi_j^A}{\pi_{ij}^A}\frac{\left(y_i-\hat{\mu}_i\hat{B}^{MC}\right)}{\pi_i^A}\frac{\left(y_j-\hat{\mu}_j\hat{B}^{MC}\right)}{\pi_j^A} \tag{17}$$

Now for the second component, assuming that $\mathscr{B}$ is single-stage probability-based sample::

$$\begin{aligned}E_{\mathscr{A}}V_{\mathscr{B}} &= E_{\mathscr{A}}\left[V_{\mathscr{B}}\left(\sum_{i\in s_A}d_i^A\left(y_i-\mu_iB^{MC}\right)+\sum_{i\in s_B}d_i^B\mu_iB^{MC}\right)\right]\\ &= E_{\mathscr{A}}\left[V_{\mathscr{B}}\left(\sum_{i\in s_B}d_i^B\mu_iB^{MC}\right)\right]\end{aligned} \tag{18}$$

$$= \sum_{i\in U}\left(\frac{\mu_iB^{MC}}{\pi_i^B}\right)^2\pi_i^B(1-\pi_i^B)+$$

$$\sum_{i\in U}\sum_{j\neq i}\left(\pi_{ij}^B-\pi_i^B\pi_j^B\right)\frac{\mu_iB^{MC}}{\pi_i^B}\frac{\mu_jB^{MC}}{\pi_j^B} \tag{19}$$

We use sample estimates for population quantities in (19):

$$\widehat{E_{\mathscr{A}}V_{\mathscr{B}}} = \sum_{i\in s_B}\left(\frac{\hat{\mu}_i\hat{B}^{MC}}{\pi_i^B}\right)^2(1-\pi_i^B)+$$

$$\sum_{i\in s_B}\sum_{j\neq i}\frac{\pi_{ij}^B-\pi_i^B\pi_j^B}{\pi_{ij}^B}\frac{\hat{\mu}_i\hat{B}^{MC}}{\pi_i^B}\frac{\hat{\mu}_j\hat{B}^{MC}}{\pi_j^B} \tag{20}$$

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

$$\begin{aligned}v_{\mathscr{A}}(\hat{T}_y^{ECLASSO}) &\approx \widehat{V_{\mathscr{A}}E_{\mathscr{B}}}+\widehat{E_{\mathscr{A}}V_{\mathscr{B}}}\\ &= \sum_{i\in s_A}\left(\frac{y_i-\hat{\mu}_i\hat{B}^{MC}}{\pi_i^A}\right)^2(1-\pi_i^A)+\\ &\quad\sum_{i\in s_A}\sum_{j\neq i}\frac{\pi_{ij}^A-\pi_i^A\pi_j^A}{\pi_{ij}^A}\frac{(y_i-\hat{\mu}_i\hat{B}^{MC})}{\pi_i^A}\frac{(y_j-\hat{\mu}_j\hat{B}^{MC})}{\pi_j^A}+\\ &\quad\sum_{i\in s_B}\left(\frac{\hat{\mu}_i\hat{B}^{MC}}{\pi_i^B}\right)^2(1-\pi_i^B)+\\ &\quad\sum_{i\in s_B}\sum_{j\neq i}\frac{\pi_{ij}^B-\pi_i^B\pi_j^B}{\pi_{ij}^B}\frac{\hat{\mu}_i\hat{B}^{MC}}{\pi_i^B}\frac{\hat{\mu}_j\hat{B}^{MC}}{\pi_j^B}.\end{aligned} \tag{21}$$

□

Table 1: Governor election covariates and outcome variables, by sample type

			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%		
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%		
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%		

Table 2: Senate election covariates and outcome variables, by sample type

			Age						Gender		Race				Education			
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
GA	Analytical	3,020	12%	13%	23%	26%	23%	2%	49%	51%	74%	17%	2%	7%	8%	23%	33%	36%
	Benchmark	67	11%	13%	29%	22%	20%	6%	55%	45%	80%	11%	0%	9%	23%	37%	21%	19%
MN	Analytical	3,774	14%	14%	20%	28%	22%	2%	52%	48%	92%	2%	1%	5%	7%	27%	35%	31%
	Benchmark	57	6%	12%	18%	25%	21%	19%	50%	50%	83%	1%	4%	11%	34%	32%	18%	17%
TX	Analytical	3,273	14%	13%	21%	27%	22%	3%	49%	51%	74%	8%	11%	7%	8%	30%	33%	29%
	Benchmark	150	11%	15%	23%	17%	25%	9%	51%	49%	65%	16%	13%	7%	27%	38%	21%	14%
NC	Analytical	8,010	12%	12%	20%	26%	27%	4%	53%	47%	83%	9%	2%	6%	8%	25%	33%	34%
	Benchmark	90	6%	11%	19%	25%	29%	10%	49%	51%	74%	17%	5%	5%	27%	32%	21%	21%
VA	Analytical	6,875	15%	12%	19%	26%	25%	3%	49%	51%	82%	8%	2%	8%	8%	22%	30%	41%
	Benchmark	81	18%	12%	19%	18%	20%	13%	54%	46%	77%	16%	3%	3%	30%	26%	21%	24%
IL	Analytical	4,061	14%	13%	20%	26%	24%	3%	49%	51%	83%	7%	4%	6%	9%	26%	31%	34%
	Benchmark	78	3%	14%	21%	25%	19%	18%	53%	47%	86%	8%	4%	2%	27%	37%	21%	16%
MI	Analytical	7,820	15%	12%	19%	26%	25%	3%	51%	49%	87%	6%	1%	6%	8%	27%	31%	34%
	Benchmark	75	6%	16%	18%	21%	29%	11%	51%	49%	77%	21%	1%	1%	25%	44%	12%	19%
NJ	Analytical	1,120	18%	10%	18%	27%	24%	3%	45%	55%	78%	8%	5%	9%	13%	22%	34%	31%
	Benchmark	58	7%	12%	20%	35%	20%	6%	53%	47%	78%	11%	6%	5%	15%	31%	22%	33%

		Religion					Born again Evangelical Christian			Attend religion				Approve Obama			Party lean			House of Representative support			Outcome (whole state)	
state	sample	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	31%	8%	18%	9%	33%	28%	20%	51%	14%	25%	14%	47%	35%	54%	11%	32%	39%	29%	6%	54%	40%	41%	53%
	Benchmark	64%	12%	11%	7%	6%	51%	37%	11%	23%	34%	18%	25%	30%	68%	2%	42%	22%	37%	10%	60%	30%		
MN	Analytical	24%	18%	15%	12%	31%	15%	23%	62%	7%	25%	15%	54%	40%	45%	15%	38%	30%	32%	8%	44%	48%	49%	48%
	Benchmark	43%	28%	5%	19%	5%	18%	54%	28%	9%	37%	11%	43%	34%	59%	7%	26%	25%	49%	23%	37%	40%		
TX	Analytical	26%	17%	17%	9%	31%	25%	18%	57%	12%	26%	12%	51%	28%	58%	14%	26%	43%	31%	6%	60%	34%	51%	46%
	Benchmark	53%	13%	12%	18%	5%	47%	32%	21%	23%	25%	15%	38%	34%	59%	7%	42%	29%	29%	13%	47%	40%		
NC	Analytical	32%	9%	17%	11%	31%	27%	22%	51%	14%	25%	12%	48%	39%	54%	7%	34%	35%	31%	5%	51%	44%	55%	42%
	Benchmark	60%	11%	13%	12%	4%	50%	36%	13%	16%	28%	10%	46%	29%	64%	7%	39%	26%	35%	12%	49%	39%		
VA	Analytical	24%	11%	12%	11%	41%	16%	21%	64%	8%	21%	10%	61%	31%	43%	26%	33%	32%	35%	7%	48%	45%	48%	44%
	Benchmark	43%	12%	15%	25%	6%	32%	39%	29%	12%	29%	15%	43%	44%	51%	5%	33%	38%	29%	15%	40%	45%		
IL	Analytical	18%	25%	12%	12%	33%	13%	18%	70%	8%	23%	12%	57%	42%	52%	6%	36%	29%	34%	7%	46%	46%	49%	48%
	Benchmark	36%	21%	10%	27%	5%	36%	31%	33%	11%	19%	14%	57%	53%	46%	1%	24%	47%	29%	4%	44%	51%		
MI	Analytical	22%	20%	14%	13%	31%	17%	20%	64%	9%	24%	12%	54%	41%	54%	5%	33%	32%	36%	7%	47%	46%	51%	46%
	Benchmark	28%	25%	19%	25%	4%	32%	38%	31%	13%	24%	13%	49%	48%	44%	8%	19%	42%	39%	16%	30%	54%		
NJ	Analytical	11%	32%	9%	11%	38%	7%	13%	80%	7%	19%	11%	64%	39%	48%	13%	38%	26%	35%	7%	44%	49%	55%	41%
	Benchmark	13%	48%	13%	15%	11%	17%	56%	27%	7%	31%	22%	39%	47%	50%	4%	34%	32%	34%	23%	40%	38%		