Addressing Sample Selection Bias for Machine Learning Methods[*]

Dylan Brewer[†] and Alyssa Carlson[‡]

June 29, 2023

Abstract

We study approaches for adjusting machine learning methods when the training sample differs from the prediction sample on unobserved dimensions. The machine learning literature predominately assumes selection only on observed dimensions. Common approaches are to weight or include variables that influence selection as solutions to selection on observables. Simulation results show that selection on unobservables increases mean squared prediction error using popular machine-learning algorithms. Common machine learning practices such as weighting or including variables that influence selection into the training or prediction sample often worsens sample selection bias. We propose two control-function approaches that remove the effects of selection bias before training and find that they reduce mean-squared prediction error in simulations. We apply these approaches to predicting the vote share of the incumbent in gubernatorial elections using previously observed re-election bids. We find that ignoring selection on unobservables leads to substantially higher predicted vote shares for the incumbent than when the control function approach is used.

JEL classification: C13, C31, C55, D72

Keywords: sample selection, machine learning, control function, inverse probability weighting

[*] We thank Jeffrey Wooldridge, Ying Zhu, David Kaplan, Mark Watson, Christian Cox, Todd Elder, and seminar participants at Michigan State University and the Southern Economics Association meeting for comments and feedback. The authors have no declarations of interest. The supplemental appendix for this paper can be found at https://www.dylanbrewer.com/wp-content/uploads/2021/05/BrewerCarlson_SuppAppendix.pdf.

[†] School of Economics, Georgia Institute of Technology. Email: brewer@gatech.edu

[‡] Department of Economics, University of Missouri. Email: carlsonah@missouri.edu

1 Introduction

Machine learning (ML) has become increasingly popular in economics due in part to its impressive "off-the-shelf" prediction capabilities (see e.g., Varian 2014, Bajari et al. 2015, Athey 2017, Mullainathan & Spiess 2017, Athey 2018, for overviews of ML applications in economics). In the canonical supervised-learning case, a researcher trains a ML algorithm on a set of training data for which she observes both attributes and outcomes and uses the ML algorithm to predict outcomes for a set of prediction data for which she observes only the attributes. In economic applications, the data are often not randomly assigned to the training set and prediction set — instead, the researcher observes outcomes in the training set for a reason.

Consider the example of predicting election outcomes for incumbent candidates using past election data. Because past election data only include incumbents who chose to run again, the training data suffer from sample selection. The selection decision is likely based on the politician's private beliefs of the probability of re-election which will bias the distribution of observed outcomes. This is an example of selection on unobservables where the mechanism that determines selection is potentially correlated with the outcome conditional on observed attributes. The ML literature focuses on the more restrictive case of selection on observables in which the mechanism that determines selection is independent of the outcome conditional on observed attributes (Shimodaira 2000, Zadrozny et al. 2003, Zadrozny 2004, Sugiyama et al. 2007, 2008b). In many economic examples this setting is unrealistic, particularly when being in the observed set of training data is the outcome of an economic decision as in the incumbent election example.

In this paper, we review the leading ML approaches when there is selection and find that assuming selection on observables is the most common approach. We document that the literature justifies the selection on observables assumption by arguing that prediction is immune to problems of selection on unobservables, machine learning algorithms are complex enough to sort out selection effects automatically, and assuming selection on observables is better than doing nothing to address selection. We characterize these arguments as the *prediction-is-immune*, *smart-algorithm*, and *do-something* fallacies. In simulation, we refute these arguments, showing that prediction performs worse when there is selection, that ML algorithms cannot automatically address selection, and that addressing selection on observables when there is selection on unobservables can worsen prediction.

We then review two existing estimation strategies to sample selection. The first is the inverse probability weighting approach, popular in both the ML and econometrics literature as a method to addressing selection on observables. The second is the control function approach of Heckman (1979) as a method to address selection on unobservables. Although the control function approach has been well-studied for regression-based estimators, it has yet to be thoroughly considered in conjunction with ML algorithms. A contribution of this paper is to propose two Heckman-style control function approaches that can work with a variety of ML procedures. We then study whether the popular weighting approaches are sensitive to violations of selection on observables. In simulation, we find that the weighting approaches are quite sensitive to the violations of the selection on observables assumption. In contrast, we find that the control function approaches that address selection on unobservables generally improve the prediction performance.

Finally, we apply the different approaches to predict county-level vote shares for incumbent U.S. governors based on county-level economic conditions using ML. To study the incumbency advantage, we predict hypothetical election outcomes for incumbent governors who do not run again. Because incumbents decide whether to run again based on private information and preferences, the training set is likely to suffer from selection on unobservables. If incumbents decide to run again based on private information about the probability of winning, observed incumbent vote shares will overstate the expected vote share relative to incumbents who have not yet decided to run again. We find that ignoring selection, using a weighting approach to address selection on observables, and using a control function approach to address selection on unobservables result in substantially different predicted county vote shares. Addressing selection on unobservables reduces the proportion of predicted incumbent wins by as much as eight percentage points relative to ignoring selection and seven percentage points relative to the weighted approach, which is a large effect given the baseline predicted incumbent election rate is about 40%.

This paper adds to both the ML and econometrics literature on addressing sample selection. Many papers have proposed using weighting approaches to address sample selection in both the ML literature (Shimodaira 2000, Zadrozny et al. 2003, Zadrozny 2004, Huang et al. 2006, Sugiyama et al. 2007, 2008b) and the econometrics literature (Rosenbaum 1987, Robins et al. 1995, Wooldridge 2002, 2007, Hirano et al. 2003). It is well established that this approach requires the selection on observables assumption, however, it is often casually stated that the weighting approaches still perform well even if the selection on observables assumptions does not hold (e.g., Huang et al. 2006). This paper more thoroughly investigates this claim by analyzing the sensitivity of the weighting estimators to violations of the selection on observables assumption. We find that our work is complementary to Wooldridge (2007) who re-examines the underlying assumptions for weighting approaches for M-estimators to recover population objects of interest. He finds that using weights based only on observed attributes can result in poor estimation when the correct weights depend on unobserved information. However, he focuses on the case of missing covariates while we focus on the case of missing outcomes.

Although there is a long literature in econometrics on how to address selection on unobservables (Heckman 1979, Vella 1998, Das et al. 2003), there has only been a recent growth in interest in the ML literature.[1] For example in the the recommender systems literature, Steck (2010) Schnabel et al. (2016), Wang et al. (2019), and Zhang et al. (2020) propose weighting approaches, but they use known/observed weights that depend on unobserved information rather than using estimated weights that require the selection on observables assumption. In this paper, we assume that the probability of being in the observed set of training data is unknown and needs to be estimated. There has also been a recent surge in the deep generative models literature in using imputation to address missing data that may depend on unobserved attributes in the unsupervised learning setting (Ipsen et al. 2021, Ma & Zhang 2021, Gong et al. 2021, Ghalebikesabi et al. 2021). However, we are considering a supervised learning setting where the outcome is unobserved.

We are aware of only two papers that consider a similar control function approach to missing outcomes in the ML literature. Zadrozny & Elkan (2001) estimates a linear regression with a Heckman-style control function where the probability of selection is estimated using a variety of ML approaches. Zhu (2017) provides non-asymptotic properties of partial lin-

[1] There is also a recent and related strain of literature that addresses selection on unobservables with ML but in the context of treatment effect estimation (Belloni et al. 2017, Feng et al. 2021).

ear regressions with penalization, where addressing sample selection using a Heckman-style control function is the leading example. The novel contribution of our work is to propose using the a Heckman-style control function with different ML algorithms more generally and to evaluate two different implementation approaches.

Our paper also relates to a growing literature in econometrics on the sensitivity of ML methods to violations of underlying model assumptions. Angrist & Frandsen (2022) study the sensitivity of using ML for variable selection in a instrumental variables model where they find using ML methods like random forest for the first step can result in very poor treatment effects estimates in the final stage. In a working paper, Hünermund et al. (2021) study the effect of bad controls on the performance of double ML (Belloni et al. 2014). They find that double ML is particularly sensitive to bad controls and is more likely to select the bad controls when included in the set of potential control variables. Similarly, our paper also finds that the ML methods that depend on the selection on observables assumption are quite sensitive to the violations of that assumption.

The remainder of the paper is organized as follows. The next section describes the model setup and introduces the selection framework. The following section discusses three approaches to address selection in ML: ignoring selection, inverse-probability-weighting approaches, and a proposed Heckman control function approach from econometrics. The fifth section presents simulation results for the sensitivity of the weighting approaches to violation of the selection on observables assumption. In the simulations, we also show how the control function approaches can help improve prediction when there is selection on unobservables. The sixth section applies the considered methods to predicting incumbent gubernatorial elections. The final section concludes and discusses directions for further research.

2 Selection assumptions in ML

We begin with a general sample-selection framework for a continuous outcome of interest Y_i for observation i drawn from the population:

$$Y_i = f(X_i) + U_i, \tag{1}$$

where Y_i is only observed if $S_i = 1$ while the attributes X_i are always observed. We assume the unobserved components U_i are mean independent of the attributes X_i, such that $f(X_i) = E(Y_i|X_i)$ and is thus the best predictor in the mean-squared-error sense. The prediction goal is to train on the observed outcome, $\{Y_i : S_i = 1\}$ using attributes X_i and then apply the trained learner to predict the outcome for a separate prediction sample not subject to sample selection.[2]

When there is selection, the quality of prediction depends on two factors. First, prediction quality depends on identifying the conditional mean $f(X_i) = E(Y_i|X_i)$, the best mean squared predictor, when only the the mean conditional on selection $E(Y_i|X_i, S_i = 1)$ is recoverable from the data. These two conditional means can be represented in the following relationship

$$E(Y_i|X_i, S_i = 1) = f(X_i) + E(U_i|X_i, S_i = 1). \tag{2}$$

If (S_i, X_i, Y_i) are arbitrarily correlated with one another, i.e., there is selection on unobservables, then $E(U_i|X_i, S_i = 1) \neq 0$, and training on the selected sample will produce predictions that converge to $E(Y_i|X_i, S_i = 1)$ and not $E(Y_i|X_i)$, reducing prediction quality.

Second, prediction quality depends on the performance of the considered ML algorithm as a sample approximation to a population object of interest, such as $E(Y_i|X_i)$. A primary objective of this paper is to show that the ability of an algorithm to fit the selected training data cannot overcome the bias if $E(Y_i|X_i, S_i = 1) \neq E(Y_i|X_i)$. In other words, if $E(Y_i|X_i)$ is poorly identified from $E(Y_i|X_i, S_i = 1)$, then the quality of the ML algorithm cannot make up for it and prediction accuracy will suffer.

[2] An alternative goal is to predict the unobserved counterpart to the training set, conditioned on the information that $S_i = 0$, predicting a counterfactual outcome. For example, following the incumbent's decision to not run for re-election, we may be interested in predicting the vote share if the incumbent had made the opposite decision to run again. We provide the results of prediction on the unobserved sample in the online supplementary appendix.

2.1 The selection on observables assumption

One common approach to identifying the conditional mean $E(Y_i|X_i)$ when only $\{Y_i, X_i : S_i = 1\}$ is observed is to make the selection on observables assumption:

$$P(S_i = 1|X_i, Y_i) = P(S_i = 1|X_i), \tag{3}$$

which holds that the outcome is independent of selection conditional on the observed attributes so $E(Y_i|X_i, S_i = 1) = E(Y_i|X_i)$. In other words, the object we would like to base our predictions on but is generally not estimable, $E(Y_i|X_i)$, is assumed to be equivalent to the object that is estimable, $E(Y_i|X_i, S_i = 1)$.

In the selection on observables setting, importance weighting approaches (discussed in section 3.2) can improve prediction, but the assumption of selection on observables is fundamental. In the literature, we identify three main arguments made in order to ignore selection on unobservables and to proceed with estimation as if selection was on observables or as if there was no selection. We refer to these arguments as the *prediction-is-immune fallacy*, the *smart-algorithm fallacy*, and the *do-something fallacy*.

The misconception that the approaches of econometrics and ML are at odds leads to the *prediction-is-immune fallacy*. While the econometrics field tends to focus on consistent estimation and inference of parameters or attributes of a conditional distribution, the ML field tends to focus on developing algorithms proven to have strong predictive performance (Kleinberg et al. 2015, Athey 2018, Mullainathan & Spiess 2017).[3] In one of the most highly-cited treatments of selection in ML, Zadrozny (2004) claims,[4] "In econometrics, the usual assumption is [selection on unobservables] because the goal is to estimate the parameters of a model... In classifier learning, this is not a concern, because we are mostly interested in the predictive performance of the model and not in making conclusions about the underlying mechanisms that generate the data." However, a better understanding of the data generating process is important for both parameter estimation and prediction. As discussed earlier, quality prediction is a combination of identifying a population object, such as a con-

[3] Alternatively, Farrell et al. (2021) incorporates ML into econometrics models with goals beyond prediction.

[4] Zadrozny (2004) has over 937 citations on Google Scholar as of March 6, 2023.

ditional mean, and accurate sample approximation to the population object, which is often better achieved by ML algorithms. Without the underlying assumptions that justify identification, one may be accurately approximating the wrong population object. In simulation, we show that without the selection on observables assumption, predictive accuracy worsens, demonstrating that prediction is not immune to endogenous selection.

The increase in complexity of ML algorithms has led to a shift of modelling responsibility from the researcher to the algorithm and a false sense of security that the complexity of the algorithm can compensate for a lack of understanding of the specific problems that plague the data set or economic question, which leads to the *smart-algorithm fallacy*. When addressing sample selection, understanding the role of the observed attributes is fundamental to accurate prediction. Zadrozny (2004) states "In order to make the condition $P(S_i = 1|X_i, Y_i) = P(S_i = 1|X_i)$ true in practice, the input to the classifier X_i has to include all the variables that affect the sample selection." This would suggest to the reader that including any observable attribute that affects the sample selection process would be beneficial. However, this is not true and can cause more harm than good.

For example, suppose we have access to an additional attribute Z_i, an instrument, that is only informative to the selection process such that it impacts selection but does not impact the outcome, $E(Y_i|X_i, Z_i) = E(Y_i|X_i)$. Note that inclusion of Z_i does not make the selection on observables assumption hold, i.e. $P(S_i = 1|X_i, Z_i, Y_i) \neq P(S_i = 1|X_i, Z_i)$, when selection on unobservables is true. Thus, contradictory to the quote above, the inclusion of attributes that affect the sample selection process only will not lead to the selection on observables assumption holding. In fact, we find in simulation that including the instrument, Z_i, when selection on unobservables is true can worsen prediction. This reiterates the findings in Wooldridge (2007), Bhattacharya & Vogt (2007), and Wooldridge (2016) that weighting and matching methods can be inconsistent if instruments are misused in estimation. Angrist & Frandsen (2022) similarly argues against the *smart-algorithm* fallacy in the instrumental variables setting. They find that when it is left to the ML procedure to decide which variables to include as IV controls, the ML procedure creates artificial exclusion restrictions and spurious results.

Finally, the temptation to do something rather than nothing when the treatment may be worse than the cure leads to the *do-something fallacy.* For example, Huang et al. (2006)[5] proposes a weighting approach which they acknowledge requires a selection on observables assumption that is unlikely to hold, but they justify their assumption stating "we will see experimentally that even in situations where our key assumption is not valid, our method can nonetheless perform well." In an experimental simulation, Huang et al. (2006) showed in a single setting that weighting performs better than ignoring the sample selection when selection on unobservables is true. However, this is not a general result.

[Figure 1 about here]

We replicated the simulated experiment in Huang et al. (2006) using the same UCI breast cancer data considered in that paper. Huang et al. evaluates the mean squared prediction error on the prediction sample for an unweighted estimator, OLS, and two weighted estimators, WLS - Logit and WLS - KMM, where the former uses logit to estimate the probability of being selected and the later uses their proposed kernel means matching approach to estimate weights.[6] The validity of both weighting approaches depend on the selection on observables assumption. The left plot in Figure 1 reports the results for the same data generating process considered in Huang et al.[7] Because the weighted approaches appeared to still have smaller MSPE compared to the unweighted approach, they state: "remarkably, despite our assumption regarding the difference between the training and test distributions being violated, our method still improves the test performance." But this result was very specific to the chosen parameters. For example, a slight change to the probabilities of selection parameters produces the data generating process on the right plot.[8] Now the weighted approaches do not improve relative to the unweighted approach and on average perform slightly worse. This reiterates a point made by Wooldridge (2007) that weighting approaches perform poorly if

[5] Huang et al. (2006) has 1849 citations on Google Scholar as of March 6, 2023.

[6] The KMM approach was implemented in Matlab using the code provided by http://www.gatsby.ucl.ac.uk/∼gretton/covariateShiftFiles/covariateShiftSoftware.html, accessed March 2023.

[7] The data was split into a training and test set with 0.5 probability. Within the training set, each observation is selected according to $P(S_i = 1|Y_i = 1) = 0.1$ and $P(S_i = 1|Y_i = 0) = 0.9$.

[8] Each observation is selected according to $P(S_i = 1|Y_i = 1) = 0.9$ and $P(S_i = 1|Y_i = 0) = 0.45$.

the true weights depend on unobserved quantities. Therefore, choosing a weighting approach when the underlying assumptions are not true comes with a cost to prediction quality.

Zadrozny (2004) also makes similar claims: “Even if [selection on observables] is not true in practice (either because we do not have access to all the variables that control the selection or because it truly depends directly on y), assuming [section on observables] is more realistic than the usual [missing completely at random].” To the contrary, we demonstrate in simulation that addressing selection on observables with a weighting approach can worsen prediction when there is selection on unobservables, refuting the *do-something fallacy*.

3 Strategies to address selection

This section reviews three estimation strategies to address selection in ML. The first strategy ignores sample selection altogether. The second strategy assumes selection on observables is true and uses an importance weighting approach. We propose a final strategy which allows for selection on unobservables and adopts the control function approach from Heckman (1979) to be incorporated in ML algorithms.

3.1 Ignoring sample selection

An ML algorithm that ignores sample selection minimizes the following empirical loss function

$$\min_{\theta} \sum_{i=1}^{n} S_i L\left(\hat{f}(X_i,\theta), Y_i, \alpha\right), \tag{4}$$

where $L(\cdot,\cdot,\alpha)$ is the loss function that can depend on nuisance parameters α, $\hat{f}(X_i,\theta)$ represents a model for the conditional mean indexed by unknown parameters θ which can be either finite or infinite dimensional, and S_i emphasizes that the loss function is only computed for $S_i = 1$. For example, LASSO would use a linear model, $\hat{f}(X_i,\theta) = X_i\theta$, and a quadratic loss function with L_2-norm regularization, $L(X_i\theta, Y_i, \alpha) = (Y_i - X_i\theta)^2 + \alpha||\theta||_2$.

This approach of ignoring sample selection will be valid if the following two conditions hold. First, the selection on observables assumption, equation (3), must hold. Second, the modeling space of $\hat{f}(X_i,\theta)$ must well approximate the true conditional mean,

$\hat{f}(X_i, \theta^*) = f(X_i)$ for some θ^* in the parameter space. If the second condition does not hold, then the solution to equation (4) does not converge to the true parameters θ^*.[9] To circumvent this issue, an importance probability weighting approach by weighting loss functions proportional to the inverse of the probability of being in the selected sample will recover the best approximation when the modeling space does not contain the true conditional mean.

3.2 Importance probability weighting

Zadrozny (2004) showed that even when the modeling space of $\hat{f}(X_i, \theta)$ does not well approximate the true conditional mean, the solution to the weighted loss function

$$\min_{\theta} \sum_{i=1}^{n} \frac{P(S_i = 1)}{P(S_i = 1|X_i)} S_i L\left(\hat{f}(X_i, \theta), Y_i, \alpha\right), \tag{5}$$

will converge to the optimal parameter.[10] Importance weighting approaches have been adapted into many algorithms in the ML literature to address selection on observables with a major focus on how to estimate the weights. Shimodaira (2000) proposes kernel density estimation, Zadrozny (2004) and Bickel et al. (2007) focus on classification approaches to estimate probability of selection, Huang et al. (2006) uses kernel means matching, and Sugiyama et al. (2007) minimizes the Kullback-Leibler divergence between the test distribution and the weighted training distribution. However, all of these weighting approaches rely on the selection on observables assumption (equation (3)) holding.

[9] Fan et al. (2005) explains that when there is selection on observables, identification of the population objects of interest depends on whether or not the true conditional distribution lies within the model space. Wooldridge (2007) similarly shows that under selection on observables, the population minimization problem for misspecified M-estimators does not identify components of the conditional distribution $D(Y_i|X_i)$.

[10] The optimal parameter under misspecification solves $\theta = \arg\min_\theta E\left(L\left(\hat{f}(X_i, \theta), Y_i\right)\right)$.

3.3 Control function

In the econometrics literature, Heckman (1979) proposes a control function approach to address selection on unobservables in a linear regression framework. For example, let

$$S_i = 1\{g(X_i, Z_i, \delta) + V_i > 0\}, \tag{6}$$

where $V_i \perp (X_i, Z_i)$ and Z_i are the instruments, attributes that contribute to the sample selection rule but are irrelevant in the outcome equation.[11] The control function approach assumes $E(U_i|X_i, Z_i, S_i = 1) = E(U_i|g(X_i, Z_i, \delta))$, so the conditional mean of the unobserved error conditional on selection can be controlled for through the model for the sample selection process. For example, if (U, V) was bivariate normal and independent of the attributes, Heckman (1979) shows

$$E(Y_i|X_i, Z_i, S_i = 1) = f(X_i) + \rho\lambda(g(X_i, Z_i, \delta)), \tag{7}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ denotes the inverse Mills ratio. He proposes a two-step estimator that first estimates the parameters δ from the model for the sample selection process, $g(X_i, Z_i, \delta)$, and second includes $\hat{\lambda}_i = \lambda(g(X_i, Z_i, \hat{\delta}))$ as an additional covariate in a linear regression to remove the bias induced by selection on unobservables.

Ahn & Powell (1993), Das et al. (2003), and Newey (2009) (among others) extend his approach to a distribution-free setting, where they model the conditional mean as

$$E(Y_i|X_i, Z_i, S_i = 1) = f(X_i) + m(g(X_i, Z_i, \delta)), \tag{8}$$

where $m(\cdot)$ is an unknown function that can be approximated. Let $m^k(\cdot) = (m_{1k}(\cdot), ..., m_{kk}(\cdot))$ be a vector of functions increasing in complexity that approximate $m(\cdot)$, such as polynomials

[11] Throughout this paper, we assume that the instrument satisfies the exclusion restriction. Without the exclusion restriction, the two-step Heckman estimator has been shown to perform poorly in terms of parameter estimates (Wolfolds & Siegel 2019). Alternative identification strategies such as Chamberlain (1986), Lewbel (2007), or d'Haultfoeuille & Maurel (2013) could be used instead. We investigate the sensitivity of the proposed approaches to violations of the exclusion restriction in the supplemental appendix section B.

or splines.[12]

For the control function approach, the strength of an instrument, Z_i, is crucial to recover $f(X_i)$. For example, if the instrument does not strongly predict selection, i.e., $g(X_i, Z, \delta) = g(X_i, Z', \delta)$ for $Z \neq Z'$ in the support, then $f(X_i)$ in the second step is not identified (Leung & Yu 1996, Puhani 2000). However, the strength of an instrument can be tested in a first stage.

How to implement a control function for a more general ML algorithm is an open question. We propose two different implementation approaches inspired by two equivalent control function approaches in the linear least squares setting. In both approaches, the parameters δ from the model for the sample selection process, $g(X_i, Z_i, \delta)$ are estimated in a first stage.

The first approach includes the control function as an additional attribute in the following minimization

$$\min_{\gamma} \sum_{i=1}^{n} S_i L\left(\hat{h}([X_i, \hat{m}_i], \gamma), Y_i, \alpha\right), \tag{9}$$

where $\hat{m}_i = m^k(g(X_i, Z_i, \hat{\delta}))$. In this approach, $\hat{h}([X_i, \hat{m}_i], \gamma)$ models $E(Y_i|X_i, Z_i, S_i = 1)$ in equation (8) as a function of both X_i and $\hat{m}_i$. An estimate of $f(X_i)$ is recovered from $\hat{h}(X_i, 0, \hat{\gamma})$, since $E(Y_i|X_i, Z_i, S_i = 1) = f(X_i)$ when $m(g(X_i, Z_i, \delta)) = 0$. If normality is assumed, as in Heckman (1979), then $\hat{m}_i$ can be replaced by $\hat{\lambda}_i$. We refer to this approach as the Variable Addition Control Function (CF-VA).

The second approach partials out the control function from the attributes and the outcome. Let $\tilde{Y}_i = Y_i - \hat{m}_i(\hat{M}'\hat{M})^{-1}\hat{M}'Y$ and $\tilde{X}_i = X_i - \hat{m}_i(\hat{M}'\hat{M})^{-1}\hat{M}'X$ where $\hat{M}$, Y, and X are the vertical stacking over i of $\hat{m}_i$, Y_i, and X_i respectively. Then the partialled out attributes and outcome are plugged into following minimization

$$\min_{\theta} \sum_{i=1}^{n} S_i L(\hat{f}(\tilde{X}_i, \theta), \tilde{Y}_i, \alpha). \tag{10}$$

We refer to this approach as the Partialled Out Control Function (CF-PO). Similar to the CF-VA approach, if normality is assumed, then $\hat{m}_i$ in the partialling out step can be replaced by $\hat{\lambda}_i$.

[12] Alternatively, Newey (2009) also suggests using series of the inverse Mills function or the normal CDF evaluated at $g(X_i, Z_i, \delta)$ as approximations of $m(g(X_i, Z_i, \delta))$.

When estimating a linear regression ($f(X_i, \theta) = X_i\theta$) using least squares ($L(X_i\theta, Y_i, \alpha) = (Y_i - X_i\theta)^2)$), these two approaches result in identical estimates. However, for more general models and loss functions, these two approaches need not produce the same result. For example, if the loss function includes regularization, like in LASSO, then the CF-VA approach may not select the control function attributes or may shrink their contribution to the prediction and as a result, the bias induced by selection on observables may not be fully removed.[13] On the other hand, the CF-PO approach is best suited for linear models ($f(X_i, \theta) = X_i\theta$) where partialling out fully removes the bias induced by selection on unobservables.

We are aware of only two papers that consider these approaches for ML. Zadrozny & Elkan (2001) use the CF-VA approach with decision trees in the first step and OLS in the second step while Zhu (2017) advocates the CF-PO approach with LASSO. The next section investigates the performance of these two CF approaches across ML algorithms with a Monte Carlo simulation.

4 Simulation

In the simulation, we study the performance of three popular supervised learners for continuous outcomes when there is selection on unobservables: LASSO, random forest, and neural nets.[14] We show for these learners that selection on unobservables increases mean-squared

[13] A variation on the CF-VA approach would be to "force" the learner to include the CF by not penalizing its effect. This variation would differ from both the CF-VA and CF-PO approaches. We are grateful to a referee for pointing this out. However, we find that this variation is not straightforward to implement with out of the box ML commands in all statistical software. In Stata, this can be implemented using the `lasso` command with the `(alwaysvars)` option. When tested on our application data, we find that this approach is nearly identical to the CF-VA approach, but we suspect this result is specific to the application data context. For example, if the CF was unlikely to be selected, it may make a larger difference. We leave it to future research to more systematically investigate the performance of this third approach.

[14] We surveyed articles published in the top 5 Economics Journals plus highly rated Econometrics field journals (Journal of Econometrics, Journal of Applied Econometrics, Journal of Business Economics and Statistics, and Review of Economics and Statistics) published between 2017 and 2022 that included "Machine Learn" in the title or abstract. Of the 80 articles recorded, we found that LASSO (21.6%), Random Forest (16.2%), and Neural Nets (10.8%) were the most popular ML algorithms. For a detailed discussion of these and other learners refer to Friedman et al. (2009).

prediction error when ignoring selection (refuting the idea that *prediction-is-immune*), and that adding instruments for selection as additional covariates can worsen prediction (refuting the idea that a *smart algorithm* can overcome selection). Next, we show that assuming selection is on observables and using a weighting approach can also worsen prediction relative to simply ignoring selection (refuting the idea that the researcher should *do something* to address selection even if the assumptions required are untrue). Finally, we apply the CF-VA and CF-PO approaches and evaluate their relative performances, finding that CF-PO performs best in the considered simulation.

We simulate selection on observables and unobservables within the framework described by equations (1) and (6). We generate 100 attributes, X_i, in the outcome equation drawn from a standardized multivariate normal distribution with $Cov(X_{ki}, X_{ji}) = 0.5^{|k-j|}$. The instrument that determines selection, Z_i, is generated from

$$Z_i = \sum_{j=1}^{100} 0.05 X_{ji} + e_{zi} \tag{11}$$

where e_{zi} is drawn from a $N(0, 0.75)$ resulting in an overall variance of approximately 1.5.

We consider three DGP designs. The outcome in the first design is generated from a linear model,

$$f(X_i) = \sum_{j=1}^{100} X_{ji}\theta_j \tag{12}$$

where $\theta_j = 0.4/j^2$ so there is a squared decay in the importance of the covariates. The sample selection is generated from a linear index model,

$$g(X_i, Z_i, \delta) = 1.25 + \sum_{j=1}^{100} X_{ji}\delta_j + \delta_z Z_i \tag{13}$$

where $\delta_j = 0.1/(10.5 - j)^2)$ and $\delta_z = 1$ resulting in a variance of $g(X_i, Z_i, \delta)$ equal to approximately 2.3.[15] We draw unobserved errors from a bivariate standard normal with

[15] The first DGP is similar to the designs considered in Belloni et al. (2014) and Bia et al. (2020). A quadratic decay in the parameter strength means there is not exact sparsity, but the coefficients decay quickly enough to produce an approximate sparse representation.

covariances, ρ, that vary from 0 to 0.9. This selection specification results in approximately 85% of the observations being selected into the training sample.

The second DGP uses the same sample selection equation as DGP 1 but introduces a more complex relationship between Y_i and X_i:

$$f(X_i) = -0.25 + 1.25 * \sin(\pi/4 + 0.75\pi \sum_{j=1}^{100} X_{ji}\theta_j) \tag{14}$$

where π is the mathematical constant equal approximately to 3.14. We again draw the unobserved errors from a bivariate standard normal distribution with covariances varying from 0 to 0.9.

The third DGP investigates the consequence of using a weak instrument. In this DGP, the outcome and sample selection are generated the same as DGP 1 but $\delta_z = 0.001$ and has a first stage F-statistic of approximately 0.15.

Each simulation draws 25,000 observations where on average 20,000 observations are selected into the training sample and 5,000 observations have an unobserved outcome. We train learners on the sample-selected training data ($\{(Y_i, W_i) : S_i = 1\}$) and predict outcomes on separately drawn data of sample size 5,000 that is not subject to sample selection.

For each learner, we use off-the-shelf versions from the Statistics and Machine Learning Toolbox for MATLAB R2022a.[16] We believe that most applied researchers utilize the off-the-shelf versions of the ML algorithms so our analysis is most useful in that context. For all the considered ML algorithms, we select nuisance parameters using 5-fold cross validation.[17] We

[16] We obtain LASSO estimates using the `fitrlinear` function with the least squares learner and the LASSO regularization. We obtain random forest estimates using the `fitrensemble` function with the bagging method, with 100 bins for the continuous predictors, and 300 learning cycles. MATLAB recommends binning continuous predictors to save on computational time. We also chose a relatively high number of learning cycles as increasing the number of learning cycles does not not result in over-fitting (Friedman et al. 2009, pg. 596). The trees have a minimum leaf size of 1 and the maximum number of splits is the number of observations minus 1. The neural net estimates are obtained using the `fitrnet` function with two fully connected hidden layers, each of size 50. (Friedman et al. 2009, pg. 400) recommends to specify relatively large layers and then allow for regularization.

[17] Nuisance parameters are the regularization parameter λ for LASSO, number of predictor variables to sample for each cycle of the random forest, and the regularization parameter λ for neural nets.

then train each learner with each of the following strategies to addressing sample selection:

1. Ignore Sample Selection (1): Only attributes X_i are included in the unweighted loss function in equation (4).
2. Ignore Sample Selection (2): The attributes X_i, instruments Z_i, Z_i^2, and interactions between X_i and Z_i are included in the unweighted loss function in equation (4).
3. Weighted: Both the attributes X_i and the instruments Z_i are used to estimate weights, then only X_i is included in the loss function in equation (5).
4. CF-VA: Both the attributes X_i and the instruments Z_i are used to estimate the probability of selection, then the control function is included as an additional attribute with X_i in the loss function in equation (9).
5. CF-PO: Both the attributes X_i and the instruments Z_i are used to estimate the probability of selection, then the control function is partialled out of the attributes X_i and outcome Y_i, which are used in the loss function in equation (10).

For both the weighted and CF approaches, the probability of selection, $P(S_i = 1|X_i, Z_i)$, needs to be estimated in a first stage. We estimate the probability of selection using an L1-penalized Probit,[18] where the regularization parameter is determined through 5-fold cross-validation.

4.1 Simulation results

[Figure 2 about here]

[18] Both the weighted and CF approaches can be sensitive to poor estimation of the probability of selection. Sugiyama et al. (2007) and Sugiyama et al. (2008b) show the sensitivity of the weighting approach to different weight estimation approaches while there is mixed results for the CF estimators. Arabmazar & Schmidt (1982) and Goldberger (1983) show the sensitivity of the CF estimators to incorrect distributional assumptions while Schaffner (2002) and Van der Klaauw & Koning (2003) find the CF estimators to be quite robust to distributional misspecification if the sample selection mechanism is modelled flexibly (e.g., including quadratic terms). To isolate the sensitivity to the selection on observables assumption, we have a correctly specified estimator for the probability of selection for the weights and use the correctly specified inverse Mills ratio for the control function. In the supplemental appendix section B, we investigate the performance of the CF approaches under distributional misspecification.

Figure 2 displays the mean squared prediction errors for the ignoring sample section and weighting approaches. Panel (a) displays MSPE for the first DGP, panel (b) for the second DGP, and panel (c) for the third DGP. We compare MSPE to an unbiased baseline equal to the median MSPE of the Ignore SS(1) prediction when $\rho = 0$. All three approaches rely on the selection on observables assumption which only holds when $\rho = 0$.

We identify three main takeaways. First, the presence of selection on unobservables degrades prediction quality, contradicting the *prediction-is-immune fallacy*. As the level of selection on unobservables increases (i.e., for greater values of ρ), both ignoring sample selection and weighting for selection on observables perform worse relative to the selection on observables benchmark.

Second, we find that including the instrument Z_i as an attribute in the outcome equation (Ignore SS(2)) does not help and in some cases is a detriment in terms of prediction quality (compared to Ignore SS(1)). This refutes the *smart-algorithm fallacy* because ML algorithms can actually misappropriate the effects of the instrument, Z_i, as a strong predictor of the outcome while in reality it is only relevant in the selection equation. This appears to have a particularly strong impact for neural nets, where even when selection on observables holds, including the instrument as an attribute has detrimental effects on prediction.

Third, contrary to the *do-something fallacy*, weighting does not necessarily produce better MSPE compared to ignoring sample selection when there is selection on unobservables. This is contrary to the statements in Huang et al. (2006) which suggests weighting approaches may still offer prediction improvements under selection on unobservables. This result is consistent with Wooldridge (2007) who showed that weighting estimators will be inconsistent when the estimated probability weights depend on the wrong predictors of selection (in this case only X_i and Z_i but not Y_i).

[Figure 3 about here]

Figure 3 reports the MSPE results for out of sample prediction performance of the two CF approaches as well as the Ignore SS (1) as a comparison. Panel (a) displays MSPE for the first DGP, panel (b) for the second DGP, and panel (c) for the third DGP. The results suggest two main takeaways. First, the performance of the CF approaches relative to ignoring sample selection varies with the level of selection on unobservables. For low correlation, the CF approaches perform similarly to ignoring sample selection, but as the correlation increases, both CF approaches perform better than ignoring sample selection in all of the DGPs and learners considered. In particular, the CF-PO approach tends to be fairly flat across all levels of correlation. This implies that the CF-PO approach is appropriately capturing the effect of selection on unobservables as correlation increases.

Second, in all considered settings, the CF-PO approach performs similarly across all levels of correlation with the smallest MSPE, while in some settings, the CF-VA approach offers little improvement over Ignore SS (1). Specifically, in DGP3 where the instrument is weak, the CF-VA approach does not provide much improvement over Ignore SS (1) while the CF-PO approach still manages to remove the bias due to selection on unobservables. This is likely because with a weak instrument, the control function is highly multicollinear with the included attributes which means the algorithm shrinks or prunes the CF's corrective effect in the CF-VA approach. Consequently, the CF-PO approach generally performs just as well and in some instances much better than CF-VA approach in terms of MSPE.

Additional simulations are included in the appendix that further investigate the performance of the CF methods when the instrument is invalid and errors are not normal. We find that the control function approach performs surprisingly well given the considered types of misspecifications and simulation context.

5 Incumbent election application

[Figure 4 about here]

We apply the approaches to correcting for sample selection to predicting the hypothetical gubernatorial election outcomes for incumbent governors who do not run again. Modeling election outcomes is a common goal for pure forecasting purposes (see e.g., Kennedy et al. 2017) as well as estimation of causal parameters (see e.g., Stegmaier et al. 2017). For instance, a practitioner may be interested in hypothetical election outcomes to aid an incumbent's decision to enter the race, or a practitioner may be interested in the hypothetical counterfactual election outcome if an incumbent had run for re-election rather than retiring. We consider the latter, seeking to understand the role of selection in determining incumbency advantage, using variation in term limits as the instrument for selection.[19] Scholars argue that name recognition, ability to raise funding, and ability to discourage entry of quality challengers provides an "incumbency advantage" in elections (Levitt & Wolfram 1997, Ashworth & Bueno de Mesquita 2008, Gowrisankaran et al. 2008, Hall & Snyder Jr 2015, Lopes da Fonseca 2017). Figure 4 plots the distribution of the incumbent's vote share and major-party non-incumbent candidate shares for U.S. gubernatorial elections from 1972-2010 (CQ Press 1967-2019), clearly showing the relative success of incumbents.

In response to the apparent incumbency advantage, term limits are often seen as necessary to increase the competitiveness of elections. However, the success of incumbents in observed prior elections is a misleading estimate of the expected success of an incumbent who has not yet made the decision to run again. Prior elections where the incumbent is present are fundamentally different from open-seat elections due to the incumbent's endogenous choice to run again for election. Thus, there is a selection process so that every incumbent election available in the training data is conditional upon the incumbent's willing participation.

We are interested in training a learner to predict the log odds of the vote share $Y_{i,j}$ of the incumbent i in county j as a function of the incumbent's performance in the previous term $X_{i,j}$.[20] The training data of prior elections featuring the incumbent are only observed when the incumbent chooses to run again. Thus, we model the incumbent's participation and subsequent vote share to understand how selection into the training set influences prediction. To simplify notation, we suppress the county subscript j.

Participation in an election is costly. When considering a potential repeat run for governor, the incumbent i compares the expected payoff of the campaign with the value of his or her outside option ε_i, which we assume is $\perp (X_i, Z_i)$, where Z_i is a binary variable that is equal to one for a governor who was subject to a term limit, which is an instrument for selection. The expected payoff is the incumbent's payoff from winning π_w or losing π_l weighted by the incumbent's belief about the likelihood of winning $B(X_i, U_i)$. Without loss of generality, we normalize the payoff of losing to zero, $\pi_l = 0$. The incumbent's belief about

[19] Besley & Case (1995), Escaleras & Calcagno (2009), and Alt et al. (2011) use variation in term limits to study gubernatorial quality, expenditures, and shirking.

[20] We model the log odds of the vote share because the off-the-shelf ML algorithms we consider are designed for continuous dependent variables rather than a limited dependent variable like vote share.

the likelihood of winning is based on observed public information about the incumbent's past performance X_i and unobserved (potentially private) information U_i.

We can write the incumbent's participation decision as a binary variable S_i that is equal to one if the expected payoff net of the barrier to entry exceeds the value of the outside option, $S_i = 1\{B(X_i, U_i)\pi_w - \delta_Z Z_i - \varepsilon_i > 0\}$. We model the candidate's beliefs as linear in parameters with a separable error term so that $B(X_i, U_i)\pi_w = \delta_X X_i + \pi_w U_i$ where the parameter β encompasses both the payoff of winning and the marginal effect of past performance on belief of winning. Thus, we model the incumbent's participation decision in the first stage as $S_i = 1\{\beta X_i - \delta_Z Z_i + V_i > 0\}$, where we assume $V_i = \pi_w U_i - \varepsilon_i$ is normally distributed.

For a potential election, the log odds of the vote share of the incumbent is a latent variable $Y_i = f(X_i) + U_i$ that is a function of the incumbent's past performance X_i and other unobserved information U_i. Y_i is only observed when the incumbent chooses to run. Thus, the training data Y_i are determined by a process that matches the canonical sample selection framework introduced in equations 1-6. The unobserved error U_i is present in both the model for the outcome of interest and the model for the participation (and selection) decision, $E(Y_i|X_i, S_i = 1) \neq E(Y_i|X_i)$, a clear case of selection on unobservables. The only training examples we observe are for past cases where the incumbent decided the expected payoff exceeded the costs—potentially due to unobserved private information U_i that affects the vote share of the incumbent.

5.1 Data

We study US state gubernatorial elections from 1972-2010. Our data on county election returns comes from the Voting and Elections Collection maintained by the CQ Press (1967-2019). We include all elections for which county returns are available, resulting in a sample of 502 state-level races (county returns are not available for Alaska, and Virginia does not allow incumbents to run again so these states are not included). The election returns data includes the presence of the incumbent and tallies of the number of votes for candidates by county. Figure 4 plots the distribution of the incumbent's vote share and major-party non-incumbent candidate shares, clearly showing the relative success of gubernatorial incumbents.

A broad range of county-level economic conditions serve as attributes that may predict the incumbent's county vote share. Economic conditions come from the Economic Profile by County series compiled by the US BEA (1969-2017).[21] We include election-year and two-year lags of county-level population, employment, and total and per-capita income. Employment is divided into wage and salary, farm proprietors, and non-farm proprietors. Income is divided into total and per-capita earnings, wage and salary income, dividends and interest income, farm proprietors' income, non-farm proprietors' income, employer retirement contributions, employer pension contributions, employer social security contributions, unemployment insurance, and welfare. We include state and presidential-election-cycle indicators.

The term limit variable serves as the excluded variable Z_i that affects whether the incumbent runs again, but does not affect the incumbent's vote share. Term limits and succession data come from the Klarner Governor's Dataset 2013, the Rutgers Center on the American Governor list of gubernatorial elections (McDowell 1948-2013), the Council of State Government's *Book of the States* information on historical state term-limits (The Council of State Governments 1960-2019), Ballotpedia (1969-2019), and the National Governors Association (1969-2019). The right panel of figure 4 shows the number of incumbent and open races per election cycle as well as the reason the incumbent did not run. In the final sample, the incumbent is present in 56 percent of races. Of open races, 54 percent are due to term limits and the remaining 46 percent are due to the incumbent's choice. In two cases of our sample, an existing term limit is eliminated or changed to allow the sitting governor to run again—thus, we classify a governor as term limited based on the presence of a binding term limit when the term begins.[22] In a sense, we define the term limit assignment as an "intent to treat," allowing us to avoid any manipulation of the term limit correlated with unobserved or private information correlated with the vote share.

The unit of observation is a county-level election. The final sample includes 30,940 county-level observations from 512 elections after generating lagged variables. Appendix table A1 displays summary statistics for open and incumbent races. The sample includes 217

[21] We chain all dollar amounts to 2000 using the Consumer Price Index (US BLS 1950-2020) and standardize the attributes so that each is in units of the z-score.

[22] An amendment to the 1976 Georgia State Constitution and a 1980 ballot initiative in South Carolina allowed the sitting term-limited governors to run again.

open races (representing 12,895 county-race observations) and 285 incumbent races (representing 17,095 county-race observations). We train our algorithms using the incumbent races and use information from the open and incumbent races to generate the control function. The feature space includes 239 potential attributes. Most earnings and employment statistics are balanced across the groups, but open races are associated with greater unemployment payments and entry of third-party challengers.

5.2 Results

[Figure 5 about here.]

In the first stage, we use an L1-penalized Probit to generate the weights for the weighting approach and the control function for CF-PO. The excluded instrument Z_i, whether the incumbent is term-limited, is highly relevant for prediction. When we increase the penalization term (λ) to increase the number of parameters excluded from the Probit model, the term limit variable is the last to drop out. We estimate and report the post-selection coefficients and standard errors in table 2 in the supplemental appendix. The LASSO-selected coefficients have the expected signs—term limits, presence of a third party challenger, and greater unemployment insurance payments per capita are correlated with the incumbent not seeking re-election.

To understand how the different approaches affect prediction, we plot the predicted vote shares when ignoring selection versus the predicted vote shares using the weighted and CF-PO approach. We display the results in terms of vote share rather than log odds for ease of interpretation.[23] The predictions using the weighting approach are plotted against the predictions ignoring sample selection in the top row of figure 5. If the weighting approach predicts a higher vote share, the points will lie above the $y = x$ line, and vice-versa. From the figures, it is clear that weighting changes the prediction patterns, with a slight favor towards *increasing* the predicted vote share relative to ignoring selection. This runs counter to our hypothesis that selection would cause prediction to be biased towards overly favorable vote shares if better-performing candidates choose to run for re-election more often. The bottom

[23] We re-transform the predictions using the non-parametric smearing estimator (Duan 1983).

row of figure 5 plots the predicted vote share when correcting for selection on unobservables using the CF-PO approach versus the predicted vote share when ignoring selection. These figures show that when correcting for selection on unobservables using the CF-PO approach, the predicted vote shares are *lower* relative to ignoring selection. This finding confirms the hypothesis that selection would cause prediction to be biased toward high vote shares.

We present the mean-squared prediction errors estimated from the training sample in table 3 in the supplementary appendix. Both weighting and using the CF-PO methods had ambiguous effects on the mean-squared prediction errors within the training sample. Weighting and CF-PO worsened the cross-validation fit for the LASSO algorithm but improved the fit for the random forest and neural net algorithms. We caution that it is difficult to interpret these changes to the MSPE in the training data to understand which algorithm will perform better in the prediction data because both weighting and CF-PO may worsen the fit within the training sample while improving the fit out of sample.

Finally, we use the predictions to forecast whether the incumbent would have received more than 50 percent of the vote at the state level in each election in which the incumbent did not run again. To do so, we used the predicted incumbent vote share for each county and aggregated to the state level by assuming an average turnout for each county based on historic turnout rates. Figure 6 plots the results by algorithm and by approach to selection. Ignoring selection or merely addressing selection on observables would lead us to predict the incumbent receiving the majority of votes more often using each algorithm. The neural net is particularly strongly affected by the approach to selection bias whereas random forest is somewhat less affected—a result consistent with the findings of the simulation exercises. These differences are practically meaningful—addressing selection on unobservables in the neural net case reduces the proportion of predicted incumbent wins by seven percentage points relative to ignoring selection and nine percentage points relative to the weighted approach, which is large considering a predicted incumbent win rate of about 32% when ignoring selection.

[Figure 6 about here.]

6 Conclusion

Contrary to popular belief, we show that for ML algorithms, selection on unobservables adversely affects prediction and cannot be solved by weighting schemes designed for selection on observables. More optimistically, we show that partialling out a control function term before training the ML algorithm can improve predictive performance, particularly at medium and high levels of selection. Merely including the control function as an additional covariate is less effective due to variable selection and shrinkage in the ML algorithms. In the context of our simulations, we show that it is important for the researcher to a priori determine the roles of the attributes as either variables that affect the outcome and selection mechanism or variables that only affect the selection mechanism. We show that treating both selection and outcome variables the same will result in poor prediction quality.

Our empirical application demonstrates that accounting for selection on unobservables can substantially change prediction results. When training ML learners to predict the vote share of the incumbent, addressing selection on unobservables reduces the proportion of predicted incumbent wins by as much as seven percentage points relative to ignoring selection and nine percentage points relative to the weighted approach, which is large compared to an average predicted incumbent win rate of 32% when ignoring selection. These findings are consistent with incumbents choosing to run again when the probability of success is greater.

This paper provides a preliminary exploration into selection on unobservables with ML procedures, but there are still many questions left unanswered for further research. For example, although we provide intuition as well as simulation to examine prediction quality, we do not derive proper inference for these approaches. Moreover, we only investigated the performance of the three most commonly used ML algorithms in economics research. Additional investigation on the performance of other ML algorithms could be warranted. Finally, we considered a regression setting where the outcome is continuous. Since many data applications have binary or discrete outcomes, determining how control function approaches could be implemented with classification ML algorithms is an important next step.

References

Ahn, H. & Powell, J. L. (1993), ‘Semiparametric estimation of censored selection models with a nonparametric selection mechanism’, *Journal of Econometrics* **58**(1-2), 3–29.

Alt, J., Bueno de Mesquita, E. & Rose, S. (2011), ‘Disentangling accountability and competence in elections: Evidence from U.S. term limits’, *Journal of Politics* **73**(1), 171–186.

Angrist, J. D. & Frandsen, B. (2022), ‘Machine labor’, *Journal of Labor Economics* **40**(S1), S97–S140.

Arabmazar, A. & Schmidt, P. (1982), ‘An investigation of the robustness of the tobit estimator to non-normality’, *Econometrica: Journal of the Econometric Society* pp. 1055–1063.

Ashworth, S. & Bueno de Mesquita, E. (2008), ‘Electoral selection, strategic challenger entry, and the incumbency advantage’, *Journal of Politics* **70**(4), 1006–1025.

Athey, S. (2017), ‘Beyond prediction: Using big data for policy problems’, *Science* **355**(6324), 483–485.

Athey, S. (2018), The impact of machine learning on economics, *in* ‘The Economics of Artificial Intelligence: An Agenda’, University of Chicago Press, pp. 507–547.

Bajari, P., Nekipelov, D., Ryan, S. P. & Yang, M. (2015), ‘Machine learning methods for demand estimation’, *American Economic Review* **105**(5), 481–85.

Ballotpedia (1969-2019), ‘Encyclopedia of american politics’. https://ballotpedia.org/Ballotpedia:About, accessed 06-2020.

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017), ‘Program evaluation and causal inference with high-dimensional data’, *Econometrica* **85**(1), 233–298.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), ‘Inference on treatment effects after selection among high-dimensional controls’, *The Review of Economic Studies* **81**(2), 608–650.

Besley, T. & Case, A. (1995), 'Does electoral accountability affect economic policy choices? evidence from gubernatorial term limits', *Quarterly Journal of Economics* **110**(3), 769–798.

Bhattacharya, J. & Vogt, W. B. (2007), Do Instrumental Variables Belong in Propensity Scores?, NBER Technical Working Papers 0343, National Bureau of Economic Research, Inc. https://www.nber.org/papers/t0343.

Bia, M., Huber, M. & Lafférs, L. (2020), Double machine learning for sample selection models, Papers 2012.00745v5, arXiv.org. https://arxiv.org/abs/2012.00745.

Bickel, S., Brückner, M. & Scheffer, T. (2007), Discriminative learning for differing training and test distributions, *in* 'Proceedings of the 24th international conference on Machine learning', pp. 81–88.

Chamberlain, G. (1986), 'Asymptotic efficiency in semi-parametric models with censoring', *Journal of Econometrics* **32**(2), 189–218.

CQ Press (1967-2019), 'Voting and elections collection'. Governor election returns, county detail by year, http://library.cqpress.com/elections/download-data.php, accessed 05-2020.

Das, M., Newey, W. K. & Vella, F. (2003), 'Nonparametric estimation of sample selection models', *The Review of Economic Studies* **70**(1), 33–58.

Duan, N. (1983), 'Smearing estimate: A nonparametric retransformation method', *Journal of the American Statistical Association* **78**(383), 605–610.

d'Haultfoeuille, X. & Maurel, A. (2013), 'Another look at the identification at infinity of sample selection models', *Econometric Theory* **29**(1), 213–224.

Escaleras, M. & Calcagno, P. (2009), 'Does the gubernatorial term limit type affect state government expenditures?', *Public Finance Review* **37**(5), 572.

Fan, W., Davidson, I., Zadrozny, B. & Yu, P. S. (2005), An improved categorization of classifier's sensitivity on sample selection bias, *in* 'Fifth IEEE International Conference on Data Mining', IEEE.

Farrell, M. H., Liang, T. & Misra, S. (2021), Deep learning for individual heterogeneity: An automatic inference framework, Papers 2010.14694v2, arXiv.org. https://arxiv.org/abs/2010.14694v2.

Feng, Y. et al. (2021), Causal inference in possibly nonlinear factor models, Papers 2008.13651v3, arXiv.org. https://arxiv.org/abs/2008.13651v3.

Friedman, J., Hastie, T. & Tibshirani, R. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer: New York, NY.

Ghalebikesabi, S., Cornish, R., Holmes, C. & Kelly, L. (2021), Deep generative missingness pattern-set mixture models, *in* 'The 24th International Conference on Artificial Intelligence and Statistics', pp. 3727–3735.

Goldberger, A. S. (1983), Abnormal selection bias, *in* 'Studies in econometrics, time series, and multivariate statistics', Elsevier, pp. 67–84.

Gong, Y., Hajimirsadeghi, H., He, J., Durand, T. & Mori, G. (2021), Variational selective autoencoder: Learning from partially-observed heterogeneous data, *in* 'The 24th International Conference on Artificial Intelligence and Statistics', pp. 2377–2385.

Gowrisankaran, G., Mitchell, M. F. & Moro, A. (2008), 'Electoral design and voter welfare from the us senate: Evidence from a dynamic selection model', *Review of Economic Dynamics* **11**, 1–17.

Hall, A. & Snyder Jr, J. (2015), 'How much of the incumbency advantage is due to scare-off?', *Political Science Research and Methods* **3**(3), 493–514.

Heckman, J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–161.

Hirano, K., Imbens, G. W. & Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**(4), 1161–1189.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. & Smola, A. (2006), Correcting sample selection bias by unlabeled data, *in* 'Advances in Neural Information Processing Systems 19'.

Hünermund, P., Louw, B. & Caspi, I. (2021), Double machine learning and bad controls–a cautionary tale, Papers 2108.11294, arXiv.org. https://arxiv.org/abs/2108.11294.

Ipsen, N. B., Mattei, P.-A. & Frellsen, J. (2021), not-miwae: Deep generative modelling with missing not at random data, *in* 'The 2021 International Conference on Learning Representations'.

Kennedy, R., Wojcik, S. & Lazer, D. (2017), 'Improving election prediction internationally', *Science* **355**(6324), 515–520.

Klarner, C. (2013), 'Governors dataset'. https://doi.org/10.7910/DVN/PQ0Y1N, accessed 05-2020.

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review Papers & Proceedings* **105**(5), 491–95.

Leung, S. F. & Yu, S. (1996), 'On the choice between sample selection and two-part models', *Journal of Econometrics* **72**(1-2), 197–229.

Levitt, S. & Wolfram, C. (1997), 'Decomposing the sources of incumbency advantage in the U.S. house', *Legislative Studies Quarterly* **22**(1), 45–60.

Lewbel, A. (2007), 'Endogenous selection or treatment model estimation', *Journal of Econometrics* **141**(2), 777–806.

Lopes da Fonseca, M. (2017), 'Identifying the source of incumbency advantage through a constitutional reform', *American Journal of Political Science* **61**(3), 657–670.

Ma, C. & Zhang, C. (2021), Identifiable generative models for missing not at random data imputation, *in* 'Advances in Neural Information Processing Systems 34', pp. 27645–27658.

McDowell, C. (1948-2013), 'Gubernatorial elections'. http://governors.rutgers.edu/testing/wp-content/uploads/2014/09/Incumb_Chart_Word_2013.pdf, accessed 06-2020.

Mullainathan, S. & Spiess, J. (2017), 'Machine learning: An applied econometric approach', *Journal of Economic Perspectives* **31**(2), 87–106.

National Governors Association (1969-2019), ‘Former governors’. https://www.nga.org/former-governors/, accessed 06-2020.

Newey, W. K. (2009), ‘Two-step series estimation of sample selection models’, *The Econometrics Journal* **12**(S1), S217–S229.

Puhani, P. (2000), ‘The heckman correction for sample selection and its critique’, *Journal of Economic Surveys* **14**(1), 53–68.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995), ‘Analysis of semiparametric regression models for repeated outcomes in the presence of missing data’, *Journal of the American Statistical Association* **90**(429), 106–121.

Rosenbaum, P. R. (1987), ‘Model-based direct adjustment’, *Journal of the American Statistical Association* **82**(398), 387–394.

Schaffner, J. A. (2002), ‘Heteroskedastic sample selection and developing-country wage equations’, *Review of Economics and Statistics* **84**(2), 269–280.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N. & Joachims, T. (2016), Recommendations as treatments: Debiasing learning and evaluation, *in* ‘The 33rd International Conference on Machine Learning’, pp. 1670–1679.

Shimodaira, H. (2000), ‘Improving predictive inference under covariate shift by weighting the log-likelihood function’, *Journal of Statistical Planning and Inference* **90**(2), 227–244.

Steck, H. (2010), Training and testing of recommender systems on data missing not at random, *in* ‘The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, pp. 713–722.

Stegmaier, M., Lewis-Beck, M. S. & Park, B. (2017), The VP-Function: A Review, *in* K. Arzheimer, J. Evans & M. S. Lewis-Beck, eds, ‘The SAGE Handbook of Electoral B Behaviour’, Vol. 2, SAGE Publications Ltd, 55 City Road, London, chapter 25, pp. 584–605.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V. & Kawanabe, M. (2007), Direct importance estimation with model selection and its application to covariate shift adaptation, *in* ‘Advances in Neural Information Processing Systems 20’.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. & Kawanabe, M. (2008b), ‘Direct importance estimation for covariate shift adaptation’, *Annals of the Institute of Statistical Mathematics* **60**(4), 699–746.

The Council of State Governments (1960-2019), ‘The book of the states’. http://knowledgecenter.csg.org/kc/category/content-type/content-type/book-states, accessed 06-2020.

US BEA (1969-2017), ‘Economic profile by county (CAINC30)’. Regional Economic Accounts, https://apps.bea.gov/regional/downloadzip.cfm, accessed 03-2019.

US BLS (1950-2020), ‘CPI for all urban consumers (CPI-U)’. https://data.bls.gov/timeseries/CUUR0000SA0, accessed 07-2020.

Van der Klaauw, B. & Koning, R. H. (2003), ‘Testing the normality assumption in the sample selection model with an application to travel demand’, *Journal of Business & Economic Statistics* **21**(1), 31–42.

Varian, H. R. (2014), ‘Big data: New tricks for econometrics’, *Journal of Economic Perspectives* **28**(2), 3–28.

Vella, F. (1998), ‘Estimating models with sample selection bias: a survey’, *Journal of Human Resources* **33**(1), 127–169.

Wang, X., Zhang, R., Sun, Y. & Qi, J. (2019), Doubly robust joint learning for recommendation on data missing not at random, *in* ‘The 36th International Conference on Machine Learning’, pp. 6638–6647.

Wolfolds, S. E. & Siegel, J. (2019), ‘Misaccounting for endogeneity: The peril of relying on the heckman two-step method without a valid instrument’, *Strategic Management Journal* **40**(3), 432–462.

Wooldridge, J. M. (2002), 'Inverse probability weighted m-estimators for sample selection, attrition, and stratification', *Portuguese Economic Journal* **1**(2), 117–139.

Wooldridge, J. M. (2007), 'Inverse probability weighted estimation for general missing data problems', *Journal of Econometrics* **141**(2), 1281–1301.

Wooldridge, J. M. (2016), 'Should instrumental variables be used as matching variables?', *Research in Economics* **70**(2), 232–237.

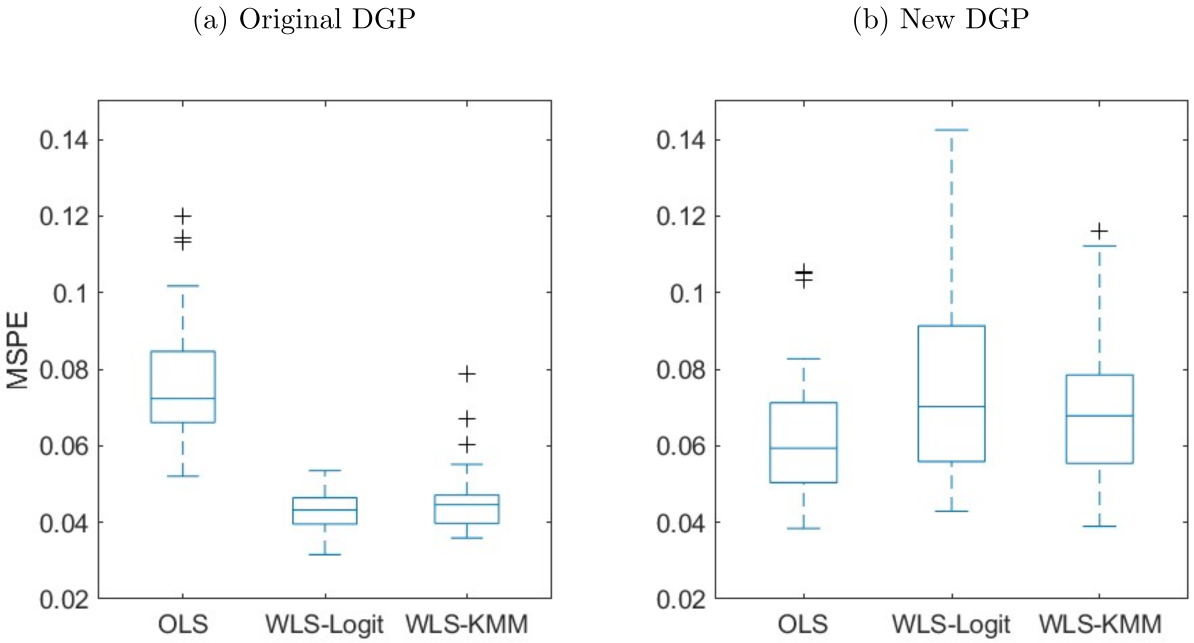
Zadrozny, B. (2004), Learning and evaluating classifiers under sample selection bias, *in* 'The 21st International Conference on Machine Learning'.

Zadrozny, B. & Elkan, C. (2001), Learning and making decisions when costs and probabilities are both unknown, *in* 'The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 204–213.

Zadrozny, B., Langford, J. & Abe, N. (2003), Cost-sensitive learning by cost-proportionate example weighting, *in* 'Third IEEE International Conference on Data Mining', pp. 435–442.

Zhang, W., Bao, W., Liu, X.-Y., Yang, K., Lin, Q., Wen, H. & Ramezani, R. (2020), Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning, *in* 'The Web Conference 2020', pp. 2775–2781.

Zhu, Y. (2017), 'Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients', *The Annals of Statistics* **45**(5), 2274–2298.

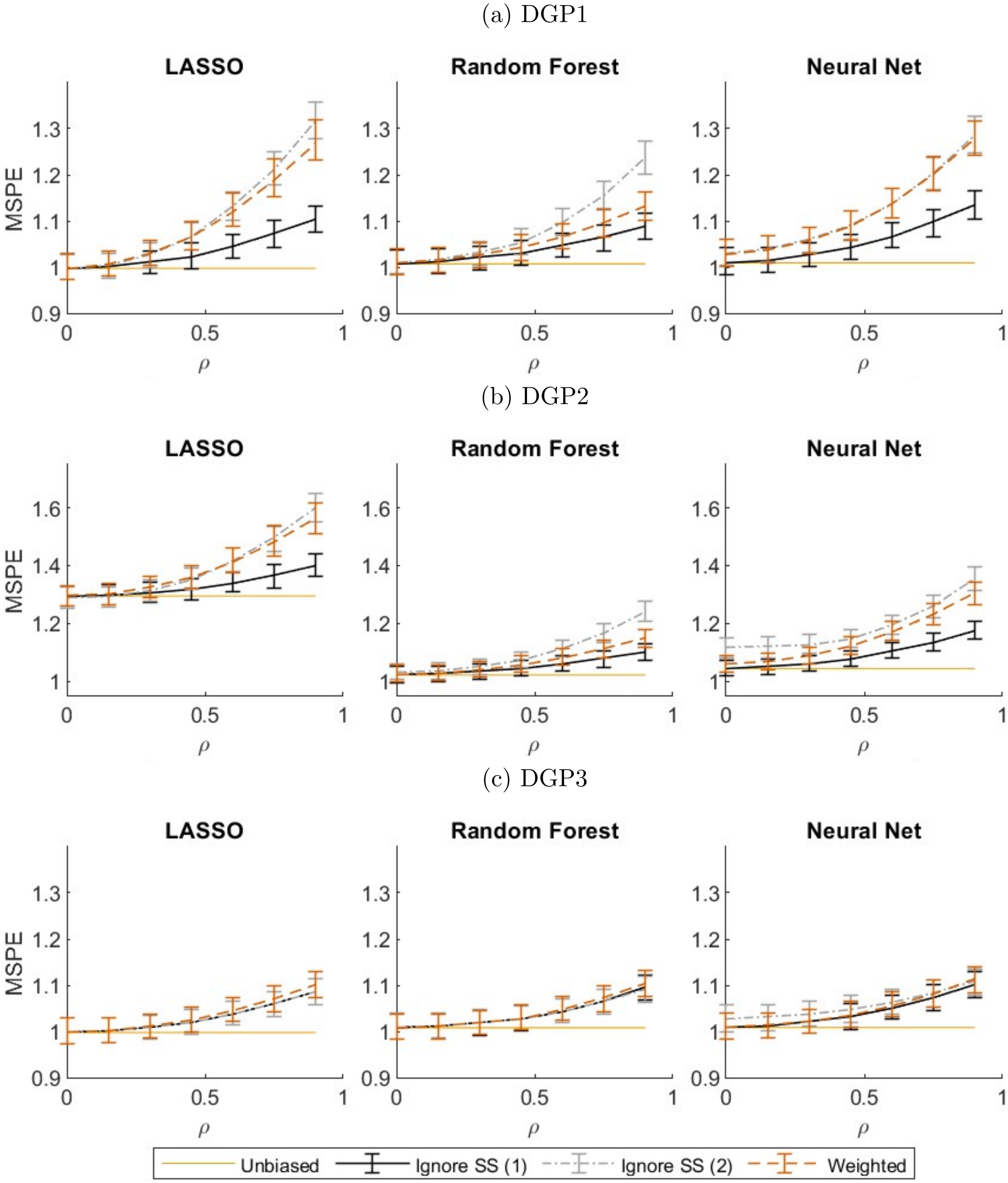


(a) Original DGP

(b) New DGP

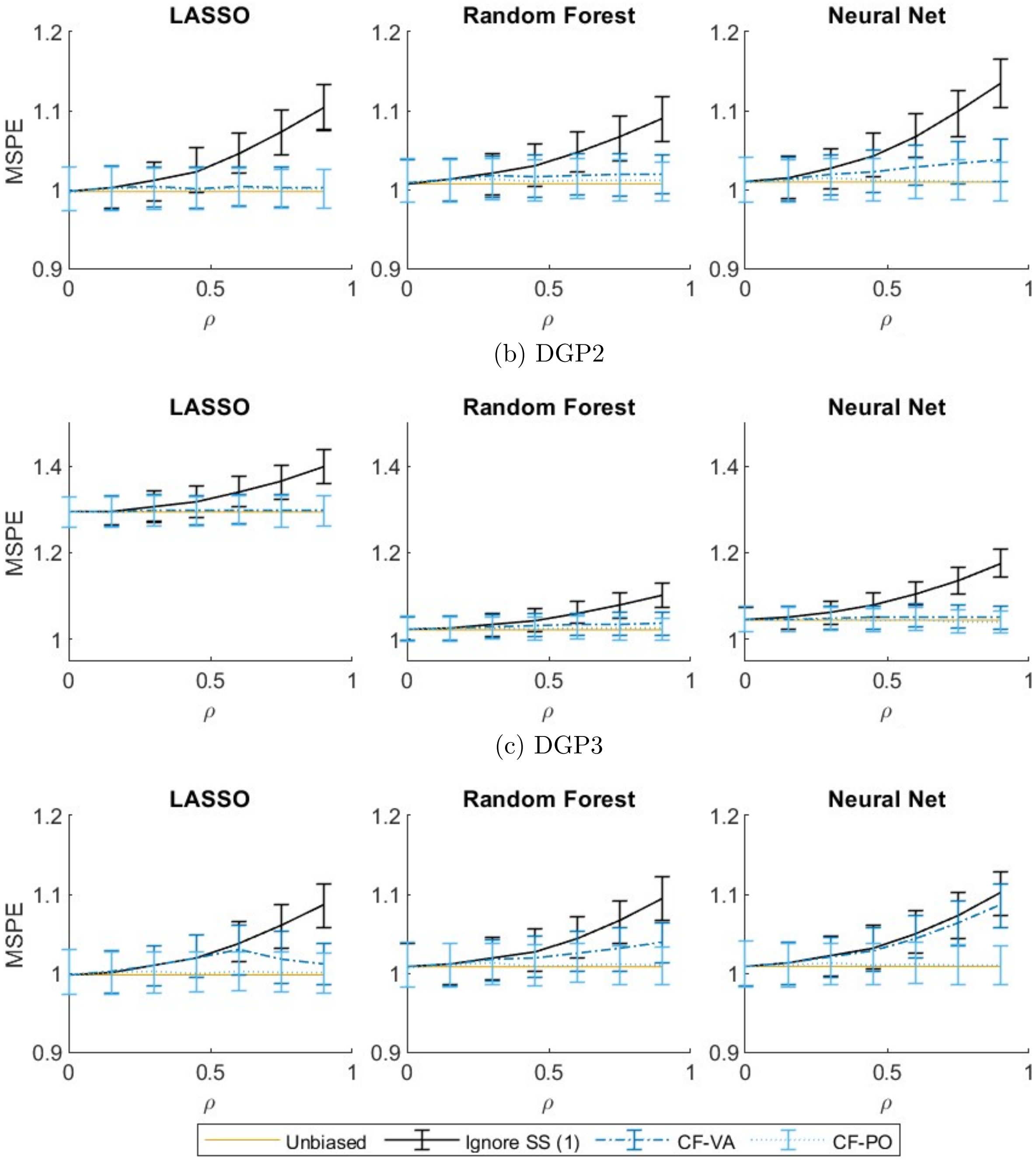
Boxplot of mean squared prediction error for 30 simulated draws of the breast cancer data from UCI following (a) the Huang et al. (2006) data generating process and (b) an alternative data generating process.

Figure 1: Replication of simulation from Huang et al. (2006).



Compares MSPE for weighted and un-weighted approaches to sample selection with error bars corresponding to 10^{th} and 90^{th} percentiles across simulations for (a) DGP1 , (b) DGP2, and (c) DGP3.

Figure 2: Weighted and un-weighted approaches to sample selection.



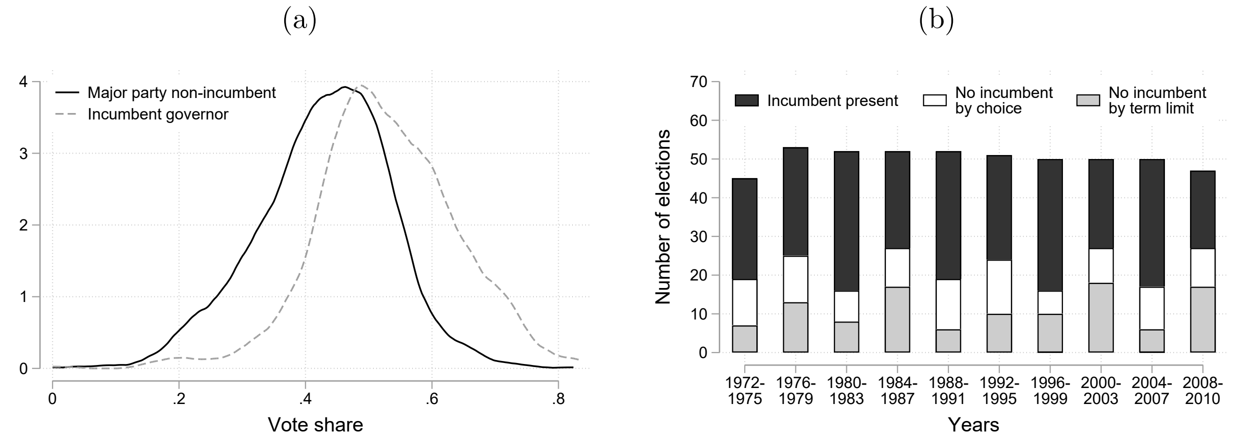
(a) DGP1

(b) DGP2

(c) DGP3

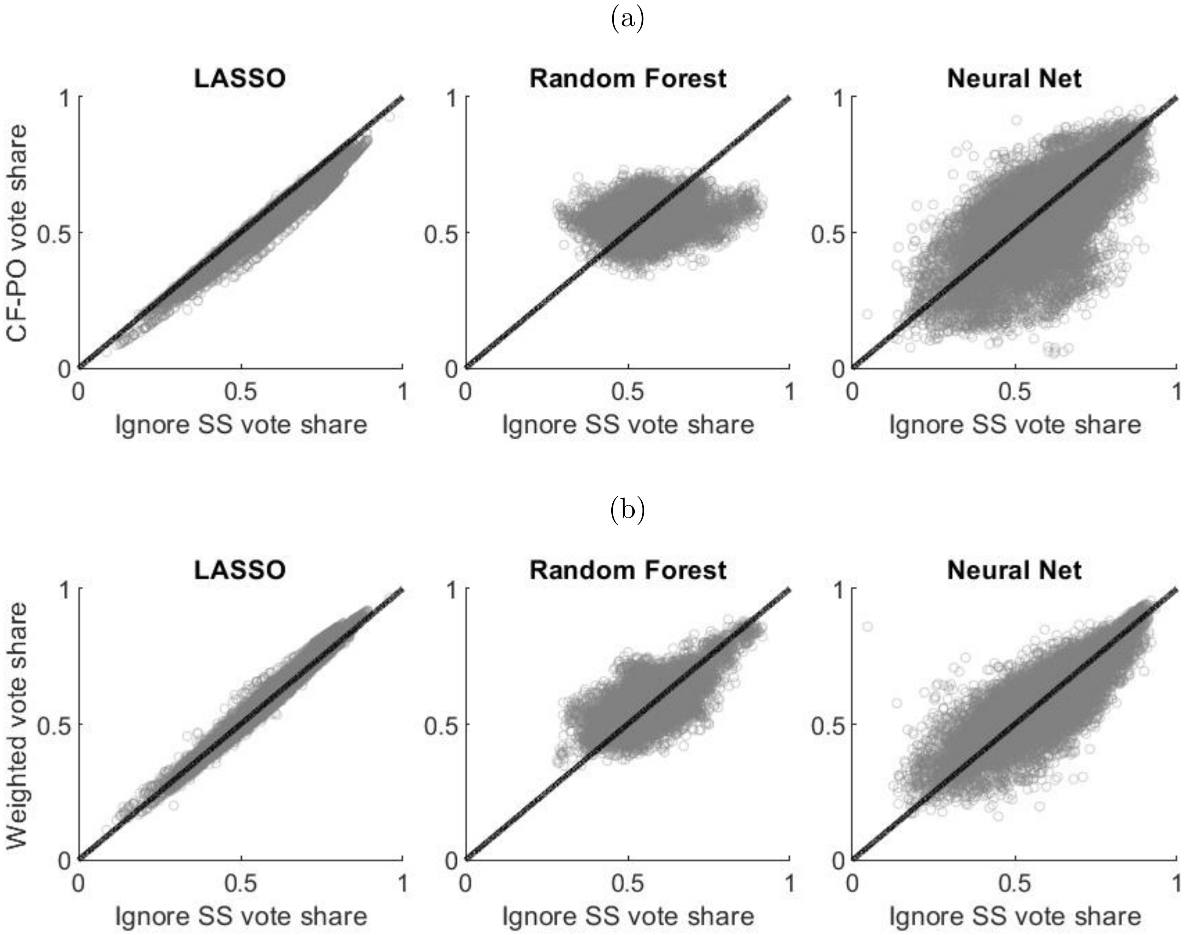
Compares MSPE for Heckman CF approaches to sample selection with error bars corresponding to 10^{th} and 90^{th} percentiles across simulations for (a) DGP1 , (b) DGP2, and (c) DGP3.

Figure 3: CF approaches to sample selection.



(a): Kernel density plots of the distribution of the incumbent's vote share and major-party non-incumbent challengers, 1972-2010. (b): Types of election by presidential cycle.

Figure 4: Gubernatorial elections.

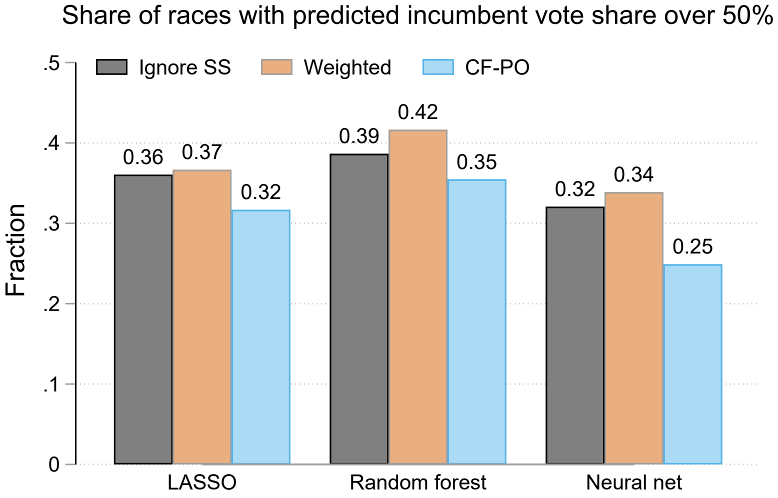


(a)

(b)

Plots predicted vote shares for each out-of-sample county race ignoring selection versus using the (a) weighting method to address selection on observables or (b) using the CF-PO method to address selection on unobservables.

Figure 5: Predicted vote share by method.



The fraction of races the algorithms forecast the incumbent receiving greater than 50% of the vote based on ignoring selection, using weighting to address selection on observables, and using the CF-PO approach to address selection on unobservables.

Figure 6: Share of incumbent winners by method.