Rehabilitating the Regression: Honest and Valid Causal Inference through Machine Learning*

Marc Ratkovic†

July 11, 2019

Abstract

The linear regression suffers from several well-known flaws. First, specification choices by the researcher can affect inference. Second, the regression is primarily a correlative tool and generally does not estimate an average causal effect. We introduce a method that overcomes both shortcomings. First, the method combines a machine learning approach to control for background covariates with a regression to estimate the coefficient on the treatment variable of theoretical interest. Second, the method models the treatment variable as well as the outcome, allowing for the recovery of a causal effect. The method applies regardless of whether the treatment variable is a binary, continuous, or count variable. We prove that the method's estimate is consistent for the average causal effect and that its standard errors are asymptotically valid and semiparametrically efficient. A simulation study and application to real-world datasets illustrates the method's utility.

Key Words: average treatment effect, causal inference, partially linear regression, machine learning

*I would like to thank Soichiro Yamauchi and Max Gopelrud for work on developing the software, John Londregan, Scott de Marchi, Brandon Stewart, Kevin Munger, Curtis Signorino, and Christopher Lucas for helpful comments; Camille DeJarnett for excellent research assistance; and Stefan Wager for guidance in implementing his software. Presented at the Midwest Political Science Association Annual Meeting, April 7, 2018 and at the Duke University Methods Seminar. Not for citation or distribution without permission from the author.

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://scholar.princeton.edu/ratkovic

1 Introduction

The linear regression is a central part of the quantitative researcher's toolkit, used to estimate effect sizes, adjust for background covariates, and conduct inference. We introduce a method that improves on two of its most pronounced shortcomings. First, inference with a regression can be *model-dependent*, meaning our inference may be sensitive to the inclusion or exclusion of control variables (Lenz and Sahn, 2017; Samii, 2016; Schrodt, 2014; King and Zeng, 2006; Achen, 2005; Berk, 2004; King, Keohane and Verba, 1994; Leamer, 1983). In response, scholars have advocated for machine learning and nonparametric methods as means to avoid linearity and additivity assumptions (Montgomery and Olivella, 2018; Hill and Jones., 2014; Hainmueller and Hazlett, 2013; Beck, King and Zeng, 2000; Grimmer, Messing and Westwood, 2017; Mohanty and Shaffer, 2018). These methods, though, are tuned for prediction and their naive use will lead to invalid inference (Chernozhukov et al., 2018).

A recent body of work has combined machine learning with the familiar regression, using machine learning to control for background variables but estimating a regression coefficient on the key variable of interest (Chernozhukov et al., 2018; Belloni, Chernozhukov and Hansen, 2014*b*,*a*). These methods do allow for valid *statistical* inference, but fall prey to the second, and deeper, critique: the regression is fundamentally a correlative measure and does not estimate an average *causal* effect (Aronow and Samii, 2016; Angrist and Pischke, 2009; Berk, 2004). A causal effect is identified by random fluctuations in the treatment variable, and therefore causal estimation requires modeling not just the mean of the treatment variable but also its variance. Failing to model the variance of the treatment (for example, Chernozhukov

et al., 2018; Fong, Hazlett and Imai, 2018; Belloni, Chernozhukov and Hansen, 2014*b*) will produce estimates that are not consistent for the average causal effect of the treatment on the outcome. Taken together, these two critiques carry deep implications for how our field accumulates knowledge, suggesting that the linear regression is not a reliable guide to how we assess our theories and test our hypotheses (Samii, 2016). While we can use machine learning methods to reduce model dependence, we cannot expect this approach to return credible causal estimates.

We utilize five key principles from the methodological literature to overcome these critiques. This first, termed *Neyman Orthogonalization* by Chernozhukov et al. (2018), requires removing the impact of the control variables prior to estimating the effect of interest. Before even considering the variable of primary interest, we need a carefully developed model of how the control variables affect both the outcome and the variable of primary interest. This requires more than simply including a battery of controls on the righthand-side, but instead constructing a flexible, yet reliable, model of how the control variables drive both the treatment and the outcome. The second principle, termed *honesty* by Wager and Athey (2017), requires that we not use the same set of data to both learn about the control variables *and* to conduct inference.[1] Honesty is most easily achieved by a *split-sample* approach (e.g. van der Vaart, 1998, ch. 25), where the data is split into equally sized subsets, with one subset used to model the covariates and the other to conduct inference. The split-sample approach creates a wall between the two processes, controlling for variables and estimating an effect, that prevents overfitting or mistakes on the former from polluting the latter. A split-sample

[1] A separate, but closely related, notion of honest inference was introduced by Li (1989); we discuss both below.

approach may inflate the estimate's variance, since the effects are estimated only off part of the data. We restore efficiency to our estimates through *repeated cross-fitting* (Chernozhukov et al., 2018), where we rotate whether each split of the data is for modeling controls or the causal effect. By this means, all the data is used to estimate the treatment effect at some point. We then repeat this procedure of splitting the data and cross-fitting, averaging estimates over the multiple splits; we illustrate below how doing so recovers efficient estimates. The fourth principle comes from *causal inference*, that our parameter of interest should be defined as an average of observation-level causal effects. The final principle is our primary theoretical result, that combining these four generates a *semiparametrically efficient* estimate of the average causal effect.

We utilize these principles in order to offer an improvement to the standard regression model. Our approach does not rely on model specifications by the researcher, returns an estimate of the average causal effect of a treatment variable on the outcome, and allows for valid inference on this estimated effect. To reduce model specification error, we implement a machine learning method to adjust for omitted variable and confounding bias attributable to the background covariates. A regression is then used for estimating the effect of the treatment on the outcome. The method returns a coefficient and standard error on the treatment variable; this standard error can be used to construct p-values and confidence intervals in the normal way. This coefficient is consistent for the average causal effect of the treatment on the outcome, regardless of whether the treatment variable is binary, continuous, or count data. We prove that the estimator is a *semiparametrically efficient* (SPE) estimate for the true causal effect, meaning that it is consistent, asymptotically normal, and has a variance no larger than the best possible parametric model.

2 Estimation Principles

We estimate the average causal effect of a treatment variable of primary theoretical interest on an outcome, after controlling for a set of background covariates. In constructing such an estimator, we rely on five principles. The first, termed *Neyman Orthogonalization* by Chernozhukov et al. (2018),[2] guides how we handle control variables. The concern is that a function of the covariates may bias our desired inference; this function is termed a *nuisance function* since it is not of direct interest but it must be modeled well in order to estimate the parameter of interest. In our setting, Neyman Orthogonalization requires generating a valid, reliable model for how the covariates impact the treatment variable and the outcome. By removing the impact of the control variables prior to estimation, we are able to conduct inference that, at least asymptotically, is independent of these covariates. As we illustrate below, this allows us to recover an estimate of the average causal effect that is asymptotically indistinguishable from an estimate were the true nuisance function known.[3]

Our second estimation principle requires conducting *honest inference*. Honest inference, in the sense of Wager and Athey (2017), requires not using the same data to both learn how the covariates impact the treatment and outcome as well as to estimate the average causal effect. We provide a mathematical justification below in Appendix B, but the heuristic argument carries the same insight. Were we to use the same data to use a machine learning

[2] After Neyman (1979), the method involves solving an efficient score condition, which is any score condition minus its expectation given covariates (see, e.g. Tsiatis, 2007; van der Vaart, 1998). The method of partialling out (Robinson, 1988) requires working with the outcome and treatment both minus their expectation given covariates. The two processes lead to estimates that are first-order equivalent (e.g. the $DML-1$ and $DML-2$ estimators of Chernozhukov et al., 2018). In our implemented model the two estimates are numerically identical, so we use the terms interchangeably.

[3] This is what is meant by an "adaptive" estimator in the sense of Bickel (1982), building on Stein (1956).

method to learn how the covariates impact the treatment and outcome *and* use this model to adjust for the covariates *and* to conduct inference, we are prone to over-fitting. A similar impulse underlies hypothesis-testing in general: using the same data to produce a hypothesis and to test it is simply data-dredging. In order to maintain honesty, we generate three separate splits of the data: one to model the level of variance in the treatment assignment, one to learn how the covariates impact the treatment and outcome, and a third to estimate and conduct inference on the average causal effect. By creating a wall between these three processes, we can guard against overfitting in any one polluting our causal inference.

Our third estimation principle is using *repeated cross-fitting* to restore efficiency (see, e.g., Theorem 4.1 of Chernozhukov et al., 2018). Sample-splitting eliminates important sources of bias in estimation, but the estimated effect will have a larger variance than a full-sample analysis since the effect is estimated off only a subset of the data. When cross-fitting, the roles of the splits are switched, generating multiple unbiased estimates that are then averaged. Repeated cross-fitting involves multiple iterations of cross-fitting, and doing so returns a series of unbiased estimates that are also then be averaged (see Chernozhukov et al., 2018, Sec. 3.4).[4] Repeated cross-fitting helps return an efficient estimate, since all of the data is used at some point to estimate the causal effect, while maintaining honesty.

Our fourth principle is that the estimand should correspond with an *average causal effect* (see, e.g. Imbens and Rubin, 2015). Intuitively, the causal effect for a single observation is the expected impact of random fluctuations in the treatment on the outcome, for that observation. An average causal effect is the effect of these individual causal effects averaged

[4]Wager and Athey (2017); Athey, Tibshirani and Wager (2019) follow a similar procedure, fitting each tree in their forest to a bootstrapped sample, then estimating effects on those "out-of-bag" observations not selected in the bootstrap, and this process is repeated over multiple iterations in creating a forest.

over the sample. Doing so requires both the magnitude of the random fluctuations in the treatment variable as well as the impact of these fluctuations on the outcome. The standard regression model treats the data, including the treatment variable as, is fixed, and will therefore not generally recover an average causal effect (Aronow and Samii, 2016; Hirano and Imbens, 2005). This insight has been lacking in much of the recent literature using machine learning methods to estimate average effects (Chernozhukov et al., 2018; Belloni, Chernozhukov and Hansen, 2014*b*; Fong, Hazlett and Imai, 2018, e.g.), though we address it here.

The final estimation principle follows from our primary analytic result, that an estimator combining the four previous principles is *semiparametrically efficient* (SPE) for the average causal effect. We prove this in two steps. First, we characterize an estimate of the average causal effect that depends on knowing the true nuisance functions, i.e. the true treatment assignment mechanism and how the covariates impact the outcome. This estimator is *infeasible*, since it requires knowing the true nuisance functions. Second, we show that these principles can be combined to generate a *feasible* estimator that is asymptotically indistinguishable from the infeasible estimate, which is what is meant by SPE. SPE establishes that, at least asymptotically, our estimator will be as efficient as a model with the controls properly specified, though we provide evidence from simulated and actual data that there is little to no efficiency loss, and possible efficiency gains, from using this method in settings where least squares regression is also valid.

A well-constructed regression model can improve the validity of our inference. By well-constructed, we mean one that learns the control specification from the data, estimates the average causal effect of the treatment on the outcome, and allows for valid inference using

confidence intervals and p-values. The need for such a tool is clear: the regression serves as the primary method for inference in our field. To gauge the prevalence of the regression model, we conducted a review of all articles published in the *American Political Science Review*, *American Journal of Political Science*, and *World Politics* from 2014-2018. We find that 78% of published articles utilize quantitative data, with the AJPS and World Politics at about 86% and the APSR at 66%. Of the quantitative studies, 40% use a least squares regression to isolate the effect of the variables of theoretical interests, and over 99% of these presents a p-value or confidence interval as a central piece of testing the underlying theory.

A substantively important proportion of the field's theory-testing depends on getting p-values and confidence intervals correct. Yet, for a variety of reasons, the standard regression model fails to do so. We review two of these reasons next, and move on to our proposed alternative.

3 Overview and Existing Work

We consider the setting where the researcher has theoretical expectations about the relationship between some variable of primary interest, a treatment variable denoted T_i, and an observed outcome, which we denote Y_i, for observations indexed by $i \in \{1, 2, 3, \ldots, n\}$. Our method applies to general treatment regimes, so the treatment variable T_i may be continuous, binary, or count data; we are not restricted to considering a binary treatment. We also assume the researcher has observed a set of k background covariates for each individual, denoted by the vector X_i. These covariates are assumed causally prior to the treatment and outcome.

We presume the researcher is interested in conducting inference on the average causal

effect of the treatment on the outcome, which we denote as θ, after adjusting for the background covariates.[5] We focus on constructing a valid confidence interval around our estimate, $\widehat{\theta}$, while using a machine learning method to control for the covariates. To do so, we generate an estimate that is consistent and asymptotically normal (CAN), such that

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, \sigma_\theta^2). \tag{1}$$

With a CAN estimate, we can construct valid confidence intervals[6] and z-statistics[7] in the standard way.

Far and away, the most common means of estimating θ is the linear regression. The regression model takes the form

$$Y_i = \mu + \theta T_i + X_i^\top \gamma + \epsilon_i;\ \mathbb{E}(\epsilon_i | T_i, X_i) = 0 \tag{2}$$

Under standard assumptions (See, e.g. Wooldridge, 2013), the least squares estimate is consistent and asymptotically normal (CAN). Under an the assumption that the errors are of equal variance, the least squares estimate is efficient among unbiased estimators (Wooldridge, 2013).

The standard regression model, despite its ubiquity, is known to be flawed. First, inference may depend on the particular model specification and how control variables are entered

[5] For a formal characterization of the average causal effect θ in a general treatment regime, see A.1.

[6] The $100 \times (1-\alpha)\%$ confidence interval $[\widehat{\theta} - C_{1-\alpha/2}\widehat{\sigma}_\theta^2/\sqrt{n}, \widehat{\theta} - C_{1-\alpha/2}\widehat{\sigma}_\theta^2/\sqrt{n}]$ will fail to cover the true value no more than $100 \times \alpha\%$ of the time, asymptotically. Taking $\alpha = .05$ gives the standard critical value $C_{0.975} = 1.96$.

[7] Under the null hypothesis of $\mathcal{H}_0 : \theta = \theta_0$, we can construct the z-statistic $\widehat{z} = \sqrt{n}|\widehat{\theta} - \theta_o|/\widehat{\sigma}_\theta$ and estimate a p-value as $2\Phi(-|\widehat{z}|)$. Taking the standard cutoff of $\widehat{z} = 1.96$ gives a p-value of 0.05.

into the model (Leamer, 1983, e.g.). Second, even if properly specified, the simple regression model does not return an estimate of the average causal effect (Aronow and Samii, 2016; Samii, 2016). We move onto these concerns next.

3.1 Risk 1: Reducing Model Dependence through the Partially Linear Model

First, inference in the regression model is *model-dependent*, in that inference on θ depends on the control set specification (King and Zeng, 2006). If there are omitted nonlinearities or interaction terms, our estimate of θ can be biased.

We can reduce concerns over model-dependence by relaxing how the control variates enter into the model, using the partially linear model (PLM, see, e.g. Härdle et al., 2012):

$$Y_i = \mu_Y + T_i\theta + f(X_i) + \epsilon_i;\ \mathbb{E}(\epsilon_i|X_i, T_i) = 0; \tag{3}$$

$$T_i = \mu_T + g(X_i) + v_i;\ \mathbb{E}(v_i|X_i) = 0. \tag{4}$$

Here f, g are nonparametric functions of the data, and therefore do not make linearity or additivity assumptions required by the standard linear model. Because the PLM contains both nonparametric components ($\{f, g\}$) and a parametric component (θ), it is an example of a *semiparametric* model. The PLM addresses two possible threats to inference: f adjusts for omitted variable bias, where a predictor is not modeled properly, and g adjusts for confounding bias, which is induced by a misspecification of a term that impacts the treatment and the outcome. The functions f, g may in-truth be linear in the covariates, but under the PLM, they need not be.

A central goal is deriving an estimate $\widehat{\theta}$ that is *semiparametrically efficient* (SPE). An estimator is efficient if it has the smallest achievable variance.[8] A semiparametrically efficient estimator is one that not only allows for inference on θ but has a variance no larger than that achieved by the best possible parametric model that perfectly controls for the nuisance functions.[9]

A SPE estimator must satisfy two conditions. First, it must be CAN, so we can use a central limit theorem to conduct inference. Second, $\widehat{\theta}$ must be estimated efficiently, were the nuisance functions $\{f, g\}$ that might bias our estimates known. In the PLM, this means that the estimate of θ should collapse onto regressing $Y_i - f(X_i)$ on $T_i - g(X_i)$, if f, g were both known without error.[10] Subtracting the effect of f, g from the outcome and treatment prior to estimating θ breaks any unwanted correlation between the covariates and our estimate of θ, which provides the motivation for Neyman orthogonalization. Neyman orthogonalization, combined with a split-sample approach, allows the researcher to generate SPE estimates.

Recent Work Using the PLM Recent years have seen a flurry of interest in the PLM. Belloni, Chernozhukov and Hansen (2014*b*) assume that the researcher has a large set of covariates, maybe hundreds or thousands, but that there is a subset that can model f and another that can model g. A machine learning method method is used to select these variables, and they are included as controls in the regression. In a second, and related work, Chernozhukov et al. (2018) introduce a "Double Machine Learning" approach. The method

[8]For example, the standard result from the Gauss-Markov Theorem that the least squares estimate is minimum variance among unbiased estimators is an efficiency result (Wooldridge, 2013, Sec. 3.5)

[9]By "best possible parametric model," we mean the least-favorable parametric submodel. For a more complete overview van der Vaart (1998, ch. 25).

[10]Formally we are describing conditions 25.55-25.56 in van der Vaart (1998) in the context of the PLM, using the approximately least favorable formulation of Sec 25.11, since we use moment conditions and not a full probability model. SPE follows under an assumption of equivariant errors; for SPE under heteroskedasticity in the outcome model, see Chernozhukov et al. (Section 2.2.4 2018).

uses any machine learning method[11] in a split-sample approach, as described above. Once an estimate of θ is recovered, the role of the estimation and auxiliary samples are split, and the two subsequent estimates are averaged, an example of cross-fitting. Repeated cross-fitting involves averaging cross-fit estimates over several different splits of the data.

A separate body of work uses random forests to estimate a treatment effect in the PLM and, in a follow up work, extend the method to instrumental variable, mediation, and quantile regression (Wager and Athey, 2017; Athey, Tibshirani and Wager, 2019). The methods work by using a tree to identify locally homogenous subsets of the data (Athey and Imbens, 2016), eliminating the effect of confounding variables. Rather than use a single tree, the authors utilize a forest (see, e.g. Hill and Jones., 2014; Montgomery and Olivella, 2018).

An additional approach put forth by Hainmueller and Hazlett (2013) uses a machine learning method to fit a single, complex, smooth curve to the outcome as a function of the treatment, and the average causal effect is estimated as the average partial derivative of this function with respect to the treatment variable. The authors do not utilize either a split-sample or Neyman orthogonalization, and thereby do not generate valid confidence intervals or p-values.[12] In a related work, Fong, Hazlett and Imai (2018) construct a set of weights that eliminate the effect of the function g in the partially linear model, under the assumption of normal, homoskedastic errors in the treatment. A bootstrap is used for uncertainty.

Like these works, our focus will be on recovering a CAN estimate of θ in a semiparametric model, where we allow the control variables to enter the model without linearity or additivity

[11] Specifically, any method that achieves a $n^{-1/4}$ uniform convergence rate on the estimation errors on f, g, which we explain more below. This condition is crude, but sufficient; see Chernozhukov et al. (2018) for details.

[12] Notably, Mohanty and Shaffer (2018) acknowledge the miscalibration of the method's confidence intervals in Appendix C.2.

assumptions. Of these works, all save the causal forest approach adjust for the mean, but fail to adjust for the variance, of the treatment variable. We next move onto the second critique, highlighting the importance of modeling the variance of the treatment.

3.2 Risk 2: Regression Estimates are not Causal

While promising, work using the PLM fails to address an important critique from the theory of causal inference. In a recent, prominent work, Aronow and Samii (2016) ask a simple question: do regression coefficients recover an estimate of an average causal effect? They show that, in general, the answer is no. The gap between the regression estimate and the causal effect is driven by not properly modeling the randomness in the treatment variable (see, e.g. Angrist and Pischke, 2009; Hirano and Imbens, 2005). Intuitively, the causal effect for a single observation is the expected impact of random fluctuations in the treatment on the outcome, for that observation. The larger the variance in these fluctuations, the more informative the observation is for the regression estimate. If heteroskedasticity in the treatment assignment covaries with the treatment effect across observations, the regression effect and causal effect will diverge.

Formally, denote θ_i as the average causal effect of the treatment on the outcome for observation i.[13] Our interest is in recovering the average effect $\theta = \mathbb{E}(\theta_i)$. We denote the systematic fluctuation of θ_i around θ as $\tau(X_i) = \mathbb{E}(\theta_i - \theta|X_i)$, θ^{LS} the parameter estimated by least squares, and $\sigma_T^2(X_i) = \text{Var}(T_i|X_i)$. The central result in Aronow and Samii (2016)

[13]We provide a formal description of the estimands in Appendix A.1.

is that the least squares regression estimate is a variance weighted average treatment effect,

$$\theta^{LS} = \frac{\mathbb{E}\left\{\sigma_T^2(X_i)\theta_i\right\}}{\mathbb{E}\left\{\sigma_T^2(X_i)\right\}}. \tag{5}$$

The gap between the two can be characterized as

$$\theta^{LS} - \theta = \frac{\mathbb{E}\left\{\sigma_T^2(X_i)\tau(X_i)\right\}}{\mathbb{E}\left\{\sigma_T^2(X_i)\right\}}, \tag{6}$$

which we will refer to as *treatment variance bias*. Inspection reveals that either one of two conditions are sufficient to guarantee that the treatment variance bias is zero, leaving the regression coefficient consistent for the true average causal effect. The first occurs when there is no treatment effect heterogeneity ($\tau(X_i) = 0$ for all observations), the second when there is no treatment assignment heterogeneity ($\sigma_T^2(X_i)$ is constant across observations). As observational studies rarely justify either assumption (see Samii, 2016, for a more complete discussion), we are left with a gap between the causal parameter θ and the parameter estimated by the regression, θ^{LS}.

4 A Partially Linear Model for Causal Effect Estimation

We introduce a model that extends the PLM model, but properly adjusts for variance in the treatment, which we call the Partial Linear Causal Effect (PLCE) model:

$$Y_i = \mu_Y + \theta T_i + \tau(X_i)T_i + f(X_i) + \epsilon_i \tag{7}$$

$$T_i = \mu_T + g_1(X_i) + g_2(X_i)\widetilde{v}_i + v_i \tag{8}$$

Our primary aim is in estimating θ, but doing so requires adjusting for several sets of nuisance functions. The first arises from misspecifying predictors (f), an example of omitted variable bias. The second bias arises through misspecifying the assignment mechanism, which can lead to confounding bias (g_1). The third form of bias, treatment variance bias, comes from not adjusting for the interaction between the treatment effect heterogeneity (τ) and heteroskedasticity in treatment assignment, (g_2).

We make the following assumptions on the PLCE:

Assumption 1 (PLCE Assumptions)

1. *Random Sample:* $\{Y_i, T_i, X_i^\top\}_{i=1}^n$ *is a random sample.*

2. *Conditional independence:* $\mathbb{E}(\epsilon_i|T_i, X_i) = \mathbb{E}(v_i|X_i) = \mathbb{E}(\widetilde{v}_i|X_i) = 0$

3. *Treatment heterogeneity:* $\mathbb{E}(v_i^2|X_i) = \sigma_v^2$, $\mathbb{E}(\widetilde{v}_i^2|X_i) = \sigma_{\widetilde{v}}^2$, $\mathrm{Cov}(v_i, \widetilde{v}_i|X_i) = 0$

4. *Positivity:* $\sigma_v^2 > C > 0$ *for some constant* C

5. *Bounded functions:* $\mathbb{E}(Y_i^4), \mathbb{E}(T_i^4) < C_1 < \infty$ *for some constant* C_1

The assumptions are either analogous to the classic least squares assumptions or are dictated by the nature of our causal estimand (see, e.g., Assumption PLM of Cattaneo, Jansson and Newey, 2018, for related assumptions). The first two assumptions are the same as the least squares assumptions, though we require conditional independence assumptions on multiple error terms since we consider the treatment variable as random. The third assumption simply splits the composite error into a component that is a function of the covariates and one that does not. The fourth requires random variance in the treatment for each observation, ensuring a counterfactual exists.[14] The fifth, while technical, relaxes the least squares assumption that the population model is linear in the covariates to simply that the population model not vary too wildly to preclude inference.

4.1 Overview of Strategy

Our strategy is to estimate a function that will take the observed covariates in X_i and observed treatment T_i and generate a new set of covariates that will adjust for each of the three forms of bias encountered when conducting causal estimation. We will denote the true function as U and our estimate of this function as $\widehat{U}$. The function will require estimating the treatment errors in order to adjust for the treatment variance. We will denote the composite error $\widetilde{u}_i = g(X_i)\widetilde{v}_i + v_i$, and its estimate $\widehat{u}_i$. Given these two estimates, we are going to generate a new set of covariates $\widehat{U}_{i,\widehat{u}_i}$ such that entering them into the model

$$Y_i = \mu_Y + \theta T_i + \widehat{U}_{i,\widehat{u}_i}^{\top}\gamma + e_i \tag{9}$$

[14] See Aronow and Samii (2016) for a discussion of and diagnostics for this assumption.

will allow for efficient estimation of the causal effect θ. We turn next to the formal framework, key assumptions, and our implementation.

4.2 The Infeasible Estimator

First, we characterize an infeasible estimator, which characterizes a control vector that would account for all the biases in the PLCE model. This vector, which we denote $U_{i,\widetilde{u}_i}$ is a function, denoted U, of the covariates (X_i), treatment (T_i), and treatment error term $(\widetilde{u}_i)$,

$$U_{i,\widetilde{u}_i} = U(X_i, T_i, \widetilde{u}_i) \tag{10}$$

where the function U maps the value data and treatment error term to the control vector. This estimate is infeasible, since we know neither the function U, its true value evaluated at the data $U_{i,\widetilde{u}_i}$, nor the composite treatment error $\widetilde{u}_i$.

If we had these values, though, we could estimate the model

$$Y_i = \mu_Y + \theta T_i + U_{i,\widetilde{u}_i}^{\top}\gamma + e_i \tag{11}$$

and return a CAN and efficient estimate for the average causal effect, θ. We denote this estimate as $\widehat{\theta}_{U,\widetilde{u}}$, and by construction

$$\sqrt{n}(\widehat{\theta}_{U,\widetilde{u}} - \theta) \rightsquigarrow \mathcal{N}(0, \sigma_u^2) \tag{12}$$

We estimate θ after partialing out the covariates (see, e.g. Robinson, 1988; Cattaneo, Jansson and Newey, 2018). Denote $\widehat{T}_{i,U_i,\widetilde{u}_i}$ and $\widehat{Y}_{i,U_i,\widetilde{u}_i}$ as our estimate of T_i and Y_i given

the true $U_{i,\widetilde{u}_i}$. We partial out the effect of the $U_{i,\widetilde{u}_i}$ by subtracting the estimated values off the observed values, considering $Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i}$ and $T_i - \widehat{T}_{i,U_i,\widetilde{u}_i}$. Our infeasible esitmate is then constructed by regressing $Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i}$ on $T_i - \widehat{T}_{i,U_i,\widetilde{u}_i}$,

$$\widehat{\theta}_{U,\widetilde{u}} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})(T_i - \widehat{T}_{i,U_i,\widetilde{u}_i})}{\sum_{i=1}^{n}(T_i - \widehat{T}_{i,U_i,\widetilde{u}_i})^2} \tag{13}$$

4.3 The Feasible Estimator

We never know U_i or $\widetilde{u}_i$, so we can never calculate $\widehat{\theta}_{U,\widetilde{u}}$. We can, though, derive a feasible estimator constructed from our estimates of the treatment error $\widehat{u}_i$, control function $\widehat{U}$, and control vector $\widehat{U}_{i,\widehat{u}_i}$. We construct an estimator, $\widehat{\theta}_{\widehat{U},\widehat{u}}$, with the important property that it is asymptotically indistinguishable from the infeasible estimate $\widehat{\theta}_{U,\widetilde{u}}$.

By asymptotically indistinguishable, we mean that the difference between the feasible estimator and infeasible estimator is $\sqrt{n}$-negligible,[15]

$$\sqrt{n}\left|\widehat{\theta}_{\widehat{U},\widehat{u}} - \widehat{\theta}_{U,\widetilde{u}}\right| \xrightarrow{\text{u}} 0. \tag{14}$$

Establishing that the feasible estimator is asymptotically indistinguishable from the optimal, infeasible estimator is at the core of constructing a SPE estimate.

4.3.1 Sample Splitting

Generating this feasible estimator requires a mixture of Neyman orthogonalization and split-sample strategies. Formally, we denote the sample $\mathcal{S}$ and split it at random into three

[15] We use $a \xrightarrow{\text{u}} b$ to denote uniform convergence of random variable a to scalar b; we provide a brief overview of this mode of convergence in Appendix A.2

approximately equal sized subsamples $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$ of size n_0, n_1, n_2. We use the first, $\mathcal{S}_0$, to model the treatment error term; the second, $\mathcal{S}_1$, to the effect of the covariates on the outcome and treatment, using the models from $\mathcal{S}_0$ and $\mathcal{S}_1$; and the third $\mathcal{S}_2$, to estimate the average causal effect, given the two previous models.

4.3.2 Modeling the Treatment Error

We use data from $\mathcal{S}_0$ to estimate the function $\widehat{u}$, our estimate of the error term in the treatment. The function takes four arguments, the treatment and covariates for evaluating the error, the sample on which the function is evaluated, and the sample on which the function is estimated, denoted

$$\widehat{u}(T_i, X_i, \mathcal{S}_a, \mathcal{S}_0) = T_i - \widehat{\mathbb{E}}^{\mathcal{S}_0}(T_i|X_i). \tag{15}$$

This function, then, is the predicted error term using a model fit to the data in $\mathcal{S}_0$ and evaluated on data in $\mathcal{S}_a$, where at different points we may wish to evaluate on $\mathcal{S}_1$ or $\mathcal{S}_2$; the important point is that it is estimated on a separate split from which it is evaluated.

4.3.3 Modeling the Effects of the Covariates

The true control function U is unknown and must be estimated. We evaluate this function on $\mathcal{S}_2$ in order to estimate θ, but we estimate the effect of the covariates from split $\mathcal{S}_1$ and the treatment error function from $\mathcal{S}_0$. This gives an estimate

$$\widehat{U}_{i,\widehat{u}_i} = \widehat{U}(X_i, \widehat{u}(T_i, X_i, \mathcal{S}_2, \mathcal{S}_0), \mathcal{S}_2, \mathcal{S}_1). \tag{16}$$

We next provide conditions on the estimation of the outcome and treatment such that we

can use $\widehat{U}_{i,\widehat{u}_i}$ as a control set in our regression. We denote the fitted values from estimating Y_i and T_i in split $\mathcal{S}_2$ given $\widehat{U}_{i,\widehat{u}_i}$ as $\widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i}$ and $\widehat{T}_{i,\widehat{U}_i,\widehat{u}_i}$. Establishing that our estimator is SPE requires the following assumption beyond those in Assumption 1:[16]

Assumption 2 (Estimation Error) *The conditional means of the outcome and the treatment* $\mathbb{E}(Y_i|U_i,\widetilde{u}_i), \mathbb{E}(T_i|U_i,\widetilde{u}_i)$ *are well-approximated such that the estimates converge uniformly to the true values at rate* $n^{-1/4}$.

This assumption requires that $\widetilde{u}$, Y_i, and T_i are sufficiently well-approximated that we can estimate their conditional means with error going to zero at rate $n^{-1/4}$ or faster, uniformly. The reduction from $n^{-1/4}$ is important; absent the split-sample approach, we would require an $n^{-1/2}$ uniform convergence, which is only obtainable by parametric methods, not a flexible machine learning method. As we discuss in Appendix A.2, many machine learning methods can indeed achieve the $n^{-1/4}$ rate, allowing us to use the machine learning method to "control" for the different biases. The technical details are in Appendix B, but intuitively, by estimating $\widetilde{u}$ off one split, U off a second, and θ off a third, any overfitting attributable to one split will not have an impact on another.

We can now construct our feasible estimator,

$$\widehat{\theta}_{\widehat{U},\widehat{u}} = \frac{\sum_{i\in\mathcal{S}_2}(Y_i - \widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(T_i - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})}{\sum_{i\in\mathcal{S}_2}(T_i - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})^2} \tag{17}$$

Our next proposition guarantees that, under the previous assumptions, the feasible estimator is SPE.

[16] See Appendix A.2 for a discussion of uniform convergence. We give the mathematical version of this assumption in Appendix B.

Proposition 1 (Split-Sample Semiparametric Estimation)

Under Assumptions 1 and 2, the estimate $\widehat{\theta}_{\widehat{U},\widehat{u}}$ *is a semiparametrically efficient estimate of the causal effect* θ*.*

Proof: *See Appendix B.*

We have just established the theoretical properties such that we can, in fact, use a machine learning method to control for our covariates while still returning an honest confidence interval on the average causal effect.

5 Our Implementation

Proposition 1 establishes that any estimator that approximates the true confounding functions sufficiently well, by satisfying Assumption 2, estimated following the proposed split-sample strategy will recover a SPE estimate of the causal effect. As discussed by Chernozhukov et al. (2018), any number of machine learning methods can achieve this rate. We implement a method that satisfies these properties, allowing the researcher to estimate the average causal effect and recover valid confidence intervals and p-values.

5.1 Assumptions for the Implemented Method

We next give our estimation procedure. Starting with our estimates of $\widetilde{u}_i$, we first use split $\mathcal{S}_0$ to estimate the model

$$T_i = B_{i;g'}^{\top} c_{g'} + \epsilon_i. \tag{18}$$

In this setting, $B_{i,g'}$ is a vector consisting of three different sets of variables, each a function of the covariate vector X_i. The first are the original covariates themselves, included as linear terms.[17] The second consist of smooth functions of each individual covariate, such that along with each term, we are also fitting a smoothing spline in each covariate (e.g. Beck and Jackman, 1998). Third, we include all two-way interactions among the smooth components.[18] The goal is to account for all linear and nonlinear main effects and interactions. We give a complete discussion of how the bases are constructed and the parameters $c_{g'}$ estimated in Appendix D.

From this, we turn to split $\mathcal{S}_1$, where we can then generate

$$\widehat{u}_i = T_i - B_{i;g'}^{\top}\widehat{c}_{g'} \tag{19}$$

using the coefficents $\widehat{c}_g$ estimated from sample $\mathcal{S}_0$. We then construct the new set of bases in $\mathcal{S}_1$ that will adjust for the heteroskedasticity in the treatment by combining the original set $B_{i;g'}$ with their interactions with the estimated error $\widehat{u}_i$ evaluated on $\mathcal{S}_1$,

$$B_{i;g} = [B_{i;g'}^{\top} : \widehat{u}_i B_{i;g'}^{\top}]^{\top} \tag{20}$$

Interacting the estimated error $\widehat{u}_i$ with the covariates will adjust for heteroskedasticity in the treatment, while the original set of bases will adjust for confounding bias.

[17] We automatically include linear covariates. Any quadratic term, discontinuity, or nonlinear term suggested by theory should be included. Though the this term would be learned by the method asymptotically, a decent specification of controls will always improve finite-sample performance.

[18] The model with two-way, nonparametric interactions is a "tensor-product smoothing spline;" see James et al. (2013) for an accessible introduction.

Staying in $\mathcal{S}_1$, we then model the outcome and the treatment as

$$Y_i = \mu_Y + B_{i,f}^\top c_f + \epsilon_i \tag{21}$$

$$T_i = \mu_T + B_{i,g}^\top c_g + \widetilde{u}_i \tag{22}$$

and estimate $\widehat{c}_f, \widehat{c}_g$.

In generating the estimated control set $\widehat{U}_{i,\widehat{u}_i}$, the primary issue is that our vectors $B_{i,f}$ and $B_{i,g}$ may contain hundreds or thousands of variables. We transform this set into a smaller set by combining a bootstrap strategy, to estimate which elements of the vectors have some explanatory power, with a principal components approach to reduce the number of covariates to a manageable number. We provide the technical details in Appendix C, but summarize the key points here. First, using data from split $\mathcal{S}_1$, we generate a bootstrap of the variance of $\widehat{c}_f$ and $\widehat{c}_g$. We then take the principal components of these variance matrices to characterize which elements of $B_{i,f}$ and $B_{i,g}$ are useful in modeling the outcome and the treatment. Intuitively, if an element of the basis vector helps model the outcome or treatment, we would expect its point estimate to vary over the bootstrap samples. If it does not, then it is not relevant for generating our control set. In order to summarize the structure of this bootstrap variance matrix, we take its principal components. Specifically, we take these principal components estimated off the split $\mathcal{S}_1$ and multiply them by the covariates from $B_{i,f}$ and $B_{i,g}$ from split $\mathcal{S}_2$. Combining the two sets gives our estimate of $\widehat{U}_{i,\widehat{u}_i}$.

In order to prove that the implemented method gives a SPE estimate, we rely on one additional assumption:

Assumption 3 (Bounded, Continuous, and Finite-Dimensional Control Set) *The functions*

$$U(X_i, \widetilde{u}_i, \mathcal{S}_2)$$

$$u(T_i, X_i, \mathcal{S}_1, \mathcal{S}_0)$$

are bounded and continuous in the covariates. The vector $U_{i,\widetilde{u}_i} = U(X_i, \widetilde{u}_i, \mathcal{S}_2)$ *is finite dimensional.*

These assumptions are in addition to Assumptions 1-2, and help us establish a feasible means of estimation. Boundedness simply requires that the control functions not go off to infinity over the range of the data. The finite dimensional assumption will help ensure that we can enter these covariates as controls into a regression and recover a valid standard error on the effect estimate. The finite dimensional assumption may seem restrictive, but it is a genuine relaxation of the standard "sparsity assumption" encountered in this literature (Belloni, Chernozhukov and Hansen, 2014*b*,*a*), where it is assumed that the nuisance functions are each a sum of some sparse subset of the covariates. We allow the estimates to not be sparse, but instead to be some flexible weighted average of all the covariates. The assumption enforces finite-dimensional *variance* on the nuisance functions.[19] The primary advantage of this assumption is that it will allow us to enter $\widehat{U}_{i,\widehat{u}_i}$ into a regression of the outcome on the treatment and recover a valid standard error from running a regression.

We can then give our final proposition.

Proposition 2 (Estimation and Inference on the Average Causal Effect) *Under*

[19]This is a "sufficient dimension reduction" assumption, as used in inverse regression (e.g. Taddy, 2013).

Assumptions 1-3, using data from split $\mathcal{S}_2$ but $\widehat{U}_{i,\widehat{u}_i}$ constructed from splits $\mathcal{S}_1, \mathcal{S}_0$, the regression estimate of θ from the model

$$Y_i = \mu_Y + \theta T_i + \widehat{U}_{i,\widehat{u}_i}^{\top}\gamma + e_i \tag{23}$$

will be semiparametrically efficient for the average causal effect of the treatment on the outcome. A robust standard error from this regression will lead to honest confidence intervals and inference.

5.2 Repeated Cross-Fitting

Our theory is derived for cases where we split the data once, in order to recover a SPE estimate of the causal effect in split $\mathcal{S}_2$. Since the size of each split goes to infinity as the sample size grows, the estimator is asymptotically efficient.

In practice, we confront a finite dataset, not an asymptote. In order to restore efficiency, we cross-fit–rotating the roles of the $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$ for a given split of the data, resplit the data, and repeat (see Chernozhukov et al., 2018, Sec. 3.4). Formally, denote the separate splits of the full data as $s \in \{1, 2, \ldots, S\}$. Each split denotes a separate splitting of the data into $\mathcal{S}_0^{(s)}, \mathcal{S}_1^{(s)}, \mathcal{S}_2^{(s)}$, as given above. From this, we then generate an estimate $\widehat{\theta}^{(s)}$ and then $\widehat{\sigma}^{2,(s)}_{\widehat{\theta}}$ is the robust standard error on the coefficient[20] (e.g. Long and Ervin, 2000). The estimates $\widehat{\theta}^{(s)}, \widehat{\sigma}^{2,(s)}$ are calculated after cross-fitting, rotating the roles of $\mathcal{S}_0^{(s)}, \mathcal{S}_1^{(s)}, \mathcal{S}_2^{(s)}$. The estimate of $\widehat{\theta}$ is simply an average over multiple splits of the data; the estimate of the variance comes

[20] We implement the "HC0" version at default but the software allows the researcher to specify others

from the law of total variance

$$\widehat{\theta} = \frac{1}{S}\sum_{i=1}^{S} \widehat{\theta}^{(s)} \tag{24}$$

$$\widehat{\sigma}^2 = \frac{1}{S}\sum_{i=1}^{S} \widehat{\sigma}^{2,(s)}_{\widehat{\theta}} + \frac{1}{S}\left\{\sum_{i=1}^{S} \frac{(\widehat{\theta}^{(s)} - \widehat{\theta})^2}{S-1}\right\}. \tag{25}$$

Though repeated cross-fitting reuses the data, each single estimate it generates for the average causal effect and its variance is honest. Averaging over these reduce the variance attributable to conditioning on a single split. We provide evidence below that repeated cross-fitting results in standard errors comparable to least squares regression when the model is in-truth linear.

5.3 Performance Expectations and Simulation Evidence

Our central goal is valid causal inference in the setting where the background variables may enter the model in nonlinear, interactive means. Along with the proposed method, we have discussed four different machine learning methods that purport to offer causal estimates in a general regression setting: Kernel Regularized Least Squares ((KRLS) Mohanty and Shaffer, 2018; Hainmueller and Hazlett, 2013), CBPS for continuous treatments ((CBPS) Fong, Hazlett and Imai, 2018), Double Machine Learning (DML) of Chernozhukov et al. (2018), and the generalized random forest (GRF) of Athey, Tibshirani and Wager (2019).

Each offers different expectations in terms of both producing unbiased causal estimates and allowing for valid causal inference. Each of the methods are explicitly designed to learn and adapt to complex models in the outcome or treatment. Three of the additional methods are still valid in the presence of nonlinearities, but fail to adjust for treatment variance bias:

KRLS, CBPS, and DML. These methods should only be used for causal effect estimation if the researcher knows treatment variance bias is not present.

The final method, GRF, uses a collection of trees to identify similar observations. To the extent that the trees select on variables that drive the treatment heteroskedasticity, we expect GRF to return unbiased estimates and valid inference. Yet, when constructing the trees, it selects on variables that only influence the mean of the treatment. GRF has a "blind spot," missing variables that drive treatment variance bias but do not influence the mean of the treatment variable, suggesting that it will not return valid standard errors in the presence of this form of treatment heteroskedasticity. We verify these theoretical expectations in a simulation study; see Appendix E for a complete description of the settings and results.

6 Extensions

The proposed method, at heart, involves generating a set of covariates that flexibly incorporate control variables and adjust for confounding bias attributable to nonrandom treatment assignment. As such, we inherit much of the flexibility, intuitions, but also shortcomings of the familiar regression model. We discuss several below.

6.1 Incorporating Multiple Variables of Theoretical Interest

The method extends naturally to testing multiple treatment variables. Assuming instead that the treatment is a vector

$$T_i = [T_{i1}, T_{i2}, \ldots, T_{iJ}]^\top$$

The method would require estimating a separate control set for each, say

$$U_{i,\widehat{U}_i,\widehat{u}_i} = [U^\top_{i,\widehat{U}_{i1},\widehat{u}_{i1}}, U^\top_{i,\widehat{U}_{i2},\widehat{u}_{i2}}, \ldots, U^\top_{i,\widehat{U}_{iJ},\widehat{u}_{iJ}}]^\top \tag{26}$$

and then entering this control set along with the treatment variables. An important assumption in this case is that, after conditioning on covariates, the treatment variables have no independent causal relationship on each other; relaxing this assumption would require more careful modeling of the cross-treatment interactive effect, which can quickly grow challenging (e.g. Imai and Yamamoto., 2013).

6.2 Random Effects

An advantage of situating the method with a regression framework, as opposed to using a random forest or other machine learning method, is the ease with which we can incorporate random effects. We illustrate with a single random effect. Denote $j \in \{1, 2, \ldots J\}$ some set of known groupings or clustering in the data. Following Gelman and Hill (2007), denote $j[i]$ as the cluster associated with observation i. Then, for random effects denoted a, b in the two models, we can estimate

$$Y_i = \mu_Y + \theta T_i + \tau(X_i)T_i + f(X_i) + a_{j[i]} + \epsilon_i \tag{27}$$

$$T_i = \mu_T + g_1(X_i) + g_2(X_i)\widetilde{u}_i + b_{j[i]} + u_i. \tag{28}$$

Additional random effects can be added. The clustering requires care in how we split the samples, ensuring that each split is populated by observations from each cluster.

6.3 Problems Inherited from the Linear Regression that We Do Not Address

As the method is an adaptation of the standard regression, it inherits both strengths and weaknesses. We discuss several of the weaknesses here that are not solved through integrating machine learning with the regression model. First, omitted variable bias can still show when an important variable is not included in the model. The method can adjust for biases attributable to unspecified nonlinearities and interactions, but no statistical method can compensate for the omission of a theoretically relevant variable.

A second form of bias shows when the causal structure is not properly specified. For example, the method does not adjust for simultaneity bias, where the treatment and outcome co-occur. Additional issues, like post-treatment bias and amplification bias are not corrected for (Acharya, Blackwell and Sen, 2016; Middleton et al., 2016).

We next move on to empirical applications.

7 Empirical Applications

The proposed method allows for several advances over both the standard regression and cutting-edge machine learning methods. We illustrate using two datasets. In the first, we address concerns that the split-sample approach embedded in repeated cross-fittng may inflate the variance or our estimated effects. We consider a survey experiment, where we know least squares provides an unbiased estimate of the average treatment effect, and show that the proposed method returns practially identical point estimates and standard errors. In the second application, we reanalyze data from a study where a continuous treatment variable

	PLCE	Least Squares without Covariates	Least Squares with Covariates
Hawks	12.18	11.98	11.97
s.e.	(3.79)	(3.80)	(3.80)
Doves	34.37	35.43	35.19
s.e.	(3.10)	(3.12)	(2.85)

Table 1: **PLCE versus Least Squares Estimates:** Estimate effect

was artificially dichotomized in order to estimate a causal effect. The proposed method handles continuous treatment variables, allowing us to maintain the treatment variable on its original scale.

7.1 Maintaining Efficiency

The split-sample and repeated cross-fitting approach raises concerns over the efficiency of the PLCE estimates. We illustrate here using data from a survey experiment, where we know least squares provides is unbiased and minimum variance among linear unbiased estimates.

Mattes and Weeks (2019) conduct a survey experiment in the United States, asking respondents about a hypothetical foreign affairs crisis involving China and military presence in the Arctic. Varied is whether the hypothetical President is a hawk or dove, whether the policy is conciliatory or maintains status quo military levels, the party of the President, and whether the policy is effective in reducing Chinese military presence in the Arctic. The outcome is whether the respondent disapproves of the President's behavior; controls consist of measures of the respondent's hawkishness, views on internationalism, trust in other nations, previous vote, age, gender, education, party ID, ideology, interest in news, and importance of religion in their life.

We focus on how the estimated causal effect of conciliation varies between hawks and

doves, as reported in Table 2 of the original work. Results appear in Table 1. For the two estimated effects, we find that PLCE returns results quite similar to least squares in an experimental setting. Importantly, the standard errors are comparable between the two methods, suggesting that the proposed method does not result in a loss of efficiency relative to standard methods.

7.2 Estimating a Causal Effect in the Presence of a Continuous Treatment

We next reanalyze data from a recent study that estimated the causal effect of racial threat on voter turnout (Enos, 2015). The author operationalizes racial threat by distance to a public housing project, a continuous measure, and racial threat by changes in voting behavior. The demolishment of a subset of the projects in the early 2000s in Chicago provides a natural experiment used for identifying the causal effect. The author implements a difference-in-difference (DiD) analysis (e.g Angrist and Pischke, 2009). Unfortunately, the treatment variable of interest, distance, is continuous, while the standard DiD approach requires a binary treatment. The author artificially dichotomizes the continuous treatment variable, considering all observations closer than some threshold distance to the projects as exposed to racial threat and observations further away as not.

The threshold is not actually known, or even estimable, given the data. There is no reason to suspect that racial threat only extends, say, 0.3 kilometers, and drops off precipitously after.[21] The proposed method allows us to estimate the average causal effect of distance on the outcome. We conduct four separate analyses. For the first, we estimate the causal effect

[21] We emphasize that this problem was not an issue with the original author's design or insights, but instead of the methodological standing of our day.

of distance on change in turnout for white residents within one kilometer of a demolished housing project. The treatment variable is Euclidean distance to the housing project, and the control variables consist of turnout in the previous two elections (1998, 1996), age, squared age, gender, median income for the Census block, value of dwelling place, and whether the deed for the residence is in the name of the voter.[22] We next generate three matched samples for further analysis, matching on all covariates.[23] The first are black voters within one kilometer of a demolished housing project. As argued in the original piece (p. 11), this group will not face racial threat, and so it provides a measure of the secular trend in turnout absent racial threat. The next two samples consist of white and black voters, but both further than one kilometer from any housing project, either demolished or not. The latter two groups serve as placebo groups, since they are sufficiently far from a demolished project that any threat should be muted.

Figure 1 presents the estimated causal effect of distance from a demolished public housing unit, in kilometers, on turnout for different subsets of the data. In the top two row, we consider residents within 1 km of a demolished unit; in the bottom two rows, we consider residents greater than 1 km from any public housing unit, demolished or not.

We estimate that living adjacent to a public housing unit, rather than 1 km away, causes a decrease in turnout of about 13 percentage points, an effect in line with the results from the original analysis (see Figure 1 there). The remaining results suggest that race may be a factor. The estimated effect 0.018(0.025) for blacks near the projects is both substantively

[22] See the supplemental materials of Enos (2015) for more details.

[23] We estimate the treatment level, distance, as a function of all covariates for white residents within one kilometer of a demolished project. We then use this model to predict the treatment level, using black residents within one kilometer and then white and black residents greater than one kilometer away. Nearest neighbor matching is implemented to construct the three additional datasets (see, e.g. Imai and Van Dyk, 2004).

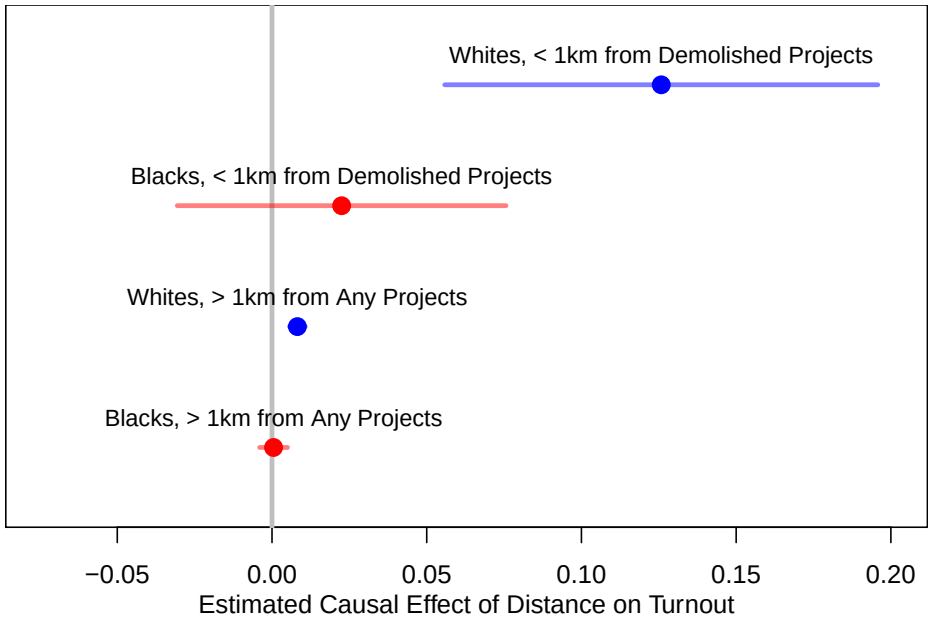

Figure 1: **Causal Effect Estimate.** Estimated causal effect of distance from a demolished public housing unit, in kilometers, on turnout for different subsets of the data. The effect on whites near the project is substantively and statistically significant. For blacks and whites further away, the effects are substantively or statistically insignifcant.

and statistically insignificant, suggesting a discernible effect for nearby whites but not nearby blacks. The bottom two lines consider distal blacks and whites. In both cases, we estimate no substantively important effects of distance on turnout, as the effect for whites is statistically significant, though miniscule.

8 Conclusion

The naive use of the regression, particularly on observational data, has been under attack from several different angles, and for quite a while (e.g. Leamer, 1983). We both suspect, and in some cases have proven, that our standard regression-based approach is not the *exact* right way to conduct inference, but the familiar tools of the regression coefficient, standard error, and p-value return such sensible results that we move ahead regardless.

Our goal has been to use recent advances in political methodology, econometrics, and

statistics to correct several of these issues. Importantly, we are not trying to replace the regression with a fancy machine-learning method, but to keep the parts we like (coefficients, standard errors, p-values) while using a machine learning method to improve the parts we do not like (control specifications for the treatment and outcome). The method learns how the covariates impact the treatment variable and outcome, before attempting inference on the treatment variable. Central to the approach is a split-sample approach, creating a wall between modeling the covariates and modeling the treatment, using different subsets of the data for modeling randomness in the treatment assignment, how the covariates impact the treatment and outcome, and conducting inference, respectively. Doing so generates *honest* confidence intervals, in that causal inference and modeling are not done using the same data.

While we have used statistical theory to resolve several issues, there are some we simply cannot address. Missing a key variable, of course, admits no statistical solution. As well, issues of reverse causality, post-treatment bias, augmentation bias, and other biases induced by misspecifying the underlying causal structure are not resolved, though we do plan to extend this method to instrumental variable, mediation, and other settings.

The hope is that the regression can return as a respectable means for causal inference in observational data. Field, laboratory, and natural experiments are an indispensable tool, offering studies with high internal validity. We argue here that a thoughtful regression analysis, constructed to control for a wide variety of interactions and nonlinearities as well as return a causal estimate, can be used to help augment and reenforce results from these strategies.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3).

Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22(4):327–339.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Aronow, Peter and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1):250–267.

Athey, Susan and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences of the United States of America* 113(27):7353–7360.

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* . Forthcoming.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):21–35.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond linearity by default: Generalized additive models." *American Journal of Political Science* pp. 596–627.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014*a*. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28(2):29–50.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014*b*. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81(2):608–650.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Sage.

Bickel, P. J. 1982. "On Adaptive Estimation." *Annals of Statisitcs* 10(3):647–671.

Cattaneo, Matias D., Michael Jansson and Whitney K. Newey. 2018. "Alternative Asymptotics and the Partially Linear Model with Many Regressors." *Econometric Theory* 34(2):277–301.

Chernozhukov, Victor, Denis Chetverikov, Esther Demirer, Mertand Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* .

Enos, Ryan D. 2015. "What the Demolition of Public Housing Teaches Usabout the Impact of Racial Threaton Political Behavior." *American Journal of Political Scince* 60(1):12.

Fong, Christian, Chad Hazlett and Kosuke Imai. 2018. "Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements." *The Annals of Applied Statistics* 12(1):156–177.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambrdige University Press.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods." *Political Analysis* 25(4):1–22.

Hainmueller, Jens and Chad Hazlett. 2013. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Härdle, Wolfgang Karl, Marlene Müller, Stefan Sperlich and Axel Werwatz. 2012. *Nonparametric and semiparametric models*. Springer Science & Business Media.

Hill, Daniel and Zachary Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):661–687.

Hirano, Keisuke and Guido Imbens. 2005. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley and Sons, Ltd, Chichester, UK chapter The Propensity Score with Continuous Treatments.

Imai, Kosuke and David A Van Dyk. 2004. "Causal inference with general treatment regimes: Generalizing the propensity score." *Journal of the American Statistical Association* 99(467):854–866.

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21(2):141–172.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biometical sciences.* Cambridge University Press.

James, G., D. Witten, T. Hastie and R. Tibshirani. 2013. *An Introduction to Statistical Learning.* New York: Springer.

King, Gary and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.

King, Gary, Robert Keohane and Sidney Verba. 1994. *Designing Social Inquiry.* Princeton:Princeton University Press.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73(1):31–42.

Lenz, Gabriel and Alexander Sahn. 2017. "Achieving Statistical Significance withCovariates and without Transparency." Working Paper.

Li, Ker-Chau. 1989. "Honest Confidence Regions for Nonparametric Regression." *The Annals of Statistics* 17(3):1001–1008.

Long, L.S. and L.H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *The American Statistician* 54:217–224.

Mattes, Michaela and Jessica L. P. Weeks. 2019. "Hawks, Doves, and Peace: An Experimental Approach." *American Journal of Political Science* 63(1):53–66.

Middleton, Joel A., Marc A. Scott, Ronli Diakow and Jennifer L. Hill. 2016. "Bias Amplification and Bias Unmasking." *Political Analysi* 24(3):307–323.

Mohanty, Pete and Robert Shaffer. 2018. "bigKRLS: Optimized Kernel Regularized Least Squares." *Political Analysis* Forthcoming. R package version 2.0.4.
URL: *https://CRAN.R-project.org/package=bigKRLS*

Montgomery, Jacob M. and Santiago Olivella. 2018. "Tree-based models for political science data." *American Journal of Political Science* .

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Newey, Whitney K and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics* 4:2111–2245.

Neyman, Jerzy. 1979. "C(α) Tests and Their Use,." *Sankhya: The Indian Journal of Statistics* 41:1–21.

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 1(25):1–40.

Robinson, Peter. 1988. "Root-N Consistent Semiparametric Regression." *Econometrica* 56(4):931–954.

Samii, Cyrus. 2016. "Causal Empricism in Quantitative Research." *Journal of Politics* 78(3):941–955.

Schrodt, Philip A. 2014. "Seven deadly sins of contemporary quantitative political analysis." *Journal of Peace Research,* 51(2):287–300.

Stein, Charles. 1956. Efficient Nonparametric Testing and Estimation. In *Berkeley Sympo-*

sium on Mathematical Statistics and Probability: Volume 1: Contributions to the Theory of Statistics, ed. Jerzy Neyman.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108(503):755–770.

Tsiatis, Anastasios. 2007. *Semiparametric Theory and Missing Data*. Springer Series in Statistics Springer Science & Business Media.

van der Vaart, A. W. and J. H. van Zanten. 2008. Reproducing kernel Hilbert spaces ofGaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor ofJayanta K. Ghosh*. Vol. 3 of *IMS Collections* pp. 200–222.

van der Vaart, Aad. 1998. *Asymptotic Statistics*. Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics* Cambridge University Press.

Wager, Stefan and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* .

Wahba, Grace. 1990. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Wooldridge, Jeffrey M. 2002. *Economic Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach*. 6 ed. Cincinnati, OH: South-Western College Publishing.

Word Count Abstract:142; Body: 10100

A Background and Preliminaries

A.1 The Average Causal Effect

To formalize our notion of causal effect, we utilize the "potential outcomes" notation popular in our field (See Imbens and Rubin, 2015, for an introduction and overview). We assume each observation has a deterministic potential outcome function which maps a treatment level t to outcome $Y_i(t)$. We characterize the average treatment effect for observation i as

$$\theta_i = \frac{\text{Cov}_{\mathcal{T}}(Y_i(t), t)}{\text{Var}_{\mathcal{T}}(t)} \tag{29}$$

$$= \frac{\text{Cov}_{\mathcal{T}}(Y_i(t), t)}{\sigma_T^2(X_i)} \tag{30}$$

which is the best linear approximation to the potential outcome function in terms of the treatment, where the $\mathcal{T}$ denotes that the operations are over the treatment. The average effect for the sample is then $\theta_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i$ and the population level effect is

$$\theta = \lim_{n\to\infty}\theta_n \tag{31}$$

$$= \mathbb{E}\left(\frac{\text{Cov}_{\mathcal{T}}(Y_i(t), t)}{\sigma_T^2(X_i)}\right) \tag{32}$$

where the outer expectation is over the sample. Note that the positivity assumption $\sigma_T^2(X_i) > 0$ guarantees θ_i, and hence θ, is a sensible quantity of interest.

For comparison, the regression estimate is

$$\theta^{LS} = \frac{\mathbb{E}(\text{Cov}(Y_i, T_i|X_i))}{\mathbb{E}(\text{Var}(T_i|X_i)}) \tag{33}$$

$$= \frac{\mathbb{E}(\text{Cov}(Y_i, T_i|X_i))}{\mathbb{E}(\sigma_T^2(X_i))} \tag{34}$$

Under an ignorability assumption on the treatment assignment that $Y_i(t) \perp\!\!\!\perp t|X_i$, the previous display equals

$$\theta^{LS} = \frac{\mathbb{E}(\text{Cov}_{\mathcal{T}}(Y_i(t), t))}{\mathbb{E}(\sigma_T^2(X_i))} \tag{35}$$

and direct substitution gives

$$\theta^{LS} = \frac{\mathbb{E}(\sigma_T^2(X_i)\theta_i)}{\mathbb{E}(\sigma_T^2(X_i))}, \tag{36}$$

showing that the least squares estimate is a variance weighted average treatment effect.

A.2 Convergence

We deal here with inference in a model where the variable of primary interest enters as through a normal linear regression but the controls are adjust for using a machine learning method. The regression component can be addressed using the standard tools found in introductory undergraduate textbooks (e.g. Wooldridge, 2013). The theory for the machine learning method requires a more advanced theory; we suggest Newey and McFadden (1994); Wooldridge (2002, ch. 12) for a concise description or Chernozhukov et al. (2018) for a self-contained overview aimed at the PLM.

We say that an estimator $\widehat{\theta}$ achieves the parametric rate of convergence if

$$\lim_{n\to\infty} \sqrt{n}(\widehat{\theta} - \nu) \rightsquigarrow \mathcal{N}(0, \sigma_\theta^2). \tag{37}$$

for some constant ν and variance $\sigma_\theta^2 < \infty$. A parametric model is consistent and asympotically normal (CAN) if it achieves the parametric rate and $\nu = \theta$, the target parameter of interest. We say an estimator achieves a nonparametric rate if it converges as n^{-a} for $a < 1/2$.

If a parameter of interest, θ, is estimated in conjunction with a nonparametric component, the slower convergence rate of the nonparametric component will pollute our parametric estimate, causing the parametric term to diverge:

$$\lim_{n\to\infty} \sqrt{n}|\widehat{\theta} - \theta| \xrightarrow{\text{p}} \infty \tag{38}$$

Breaking the pathways by which the nonparametric component can lead to poor performance of the parametric component is at the heart of semiparametric estimation.

Most of our theory using the standard linear regression requires that estimates are *consistent*, meaning they converge in probability. In a parametric model, if every parameter but θ is estimated consistently, we can recover a CAN estimate of θ using standard least squares or maximum likelihood methods. To handle the nonparametric, machine learning component, we need a stronger form of convergence. Assume the conditional mean of an

outcome $\mathbb{E}(Y_i|X_i) = m(X_i)$ and an estimate $\widehat{m}(X_i)$. The absolute bias of the estimator is

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{m}(X_i) - m(X_i)\right| \tag{39}$$

We say $\widehat{m}$ converges in probability if the absolute bias tends to zero, as the sample size grows. This suffices to characterize convergence in the linear regression, because the model is sufficiently simple that the average bias characterizes the distance between $\widehat{m}$ and m. The machine learning estimate, though, can be quite complex (e.g. Beck, King and Zeng, 2000). The only way to guarantee that convergence in the estimated machine learning component allows for inference on θ is to require *uniform convergence in probability*. Uniform convergence requires characterizing the maximal absolute bias, with least upper bound

$$\sup_{\widehat{m}\in\mathcal{M}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{m}(X_i) - m(X_i)\right| = |\widehat{m} - m_n| \tag{40}$$

where the supremum is taken over a compact set $\mathcal{M}_n$, which is the smallest set that contains any $\widehat{m}$ we might observe with probability tending to one. It is sometimes useful to subscript m to change in sample size, but we generally omit this subscript for the sake of reducing notation.

We also say, for some exponent a, that $\widehat{m}$ converges to m uniformly at rate n^{-a} if

$$\lim_{n\to\infty} n^a|\widehat{m} - m| \xrightarrow{\text{u}} 0 \tag{41}$$

We say an estimator is asymptotically $\sqrt{n}$-negligible if it converges to zero uniformly at rate $n^{-1/2}$.

An estimate $\widehat{\theta}$ allows for *honest inference* in a semiparametric model if we can recover a CAN estimate of θ in the presence of a nonparametric estimate $\widehat{m}$ that converges uniformly at some rate achievable by the estimator (Li, 1989). An estimator is not honest if it is valid only when the nonparametric component converges at a parametric rate. This definition differs slightly from Wager and Athey (2017) who define an honest estimator as one where the nonparametric control functions are estimated on one split of the data and inference conducted on the other half. The split-sample is a central tool in achieving honest confidence intervals (e.g. van der Vaart, 1998), so the Wager-Athey characterizes the *means* while the (Li, 1989) is the *end* goal. The two are clearly tightly connected.

B Proof of SPE

We start with the formal statement of Assumption 2

ASSUMPTION 4 *The treatment and outcome are well-approximated such that the estimates converge as*

$$\lim_{n\to\infty} n^{1/4}|\widehat{Y}_{\widehat{U},\widetilde{u}} - \widehat{Y}_{U,\widetilde{u}}| \xrightarrow{\text{u}} 0 \tag{42}$$

$$\lim_{n\to\infty} n^{1/4}|\widehat{Y}_{\widehat{U},\widehat{u}} - \widehat{Y}_{\widehat{U},\widetilde{u}}| \xrightarrow{\text{u}} 0 \tag{43}$$

$$\lim_{n\to\infty} n^{1/4}|\widehat{T}_{\widehat{U},\widetilde{u}} - \widehat{T}_{U,\widetilde{u}}| \xrightarrow{\text{u}} 0 \tag{44}$$

$$\lim_{n\to\infty} n^{1/4}|\widehat{T}_{\widehat{U},\widehat{u}} - \widehat{T}_{\widehat{U},\widetilde{u}}| \xrightarrow{\text{u}} 0. \tag{45}$$

We are interested in analyzing the empirical process,

$$\sqrt{n_2}(\widehat{\theta}_{\widehat{U},\widehat{u}} - \theta) \tag{46}$$

taken on split $\mathcal{S}_2$. We proceed by decomposing

$$\sqrt{n_2}(\widehat{\theta}_{\widehat{U},\widehat{u}} - \theta) = \sqrt{n_2}(\widehat{\theta}_{U,\widetilde{u}} - \theta) + \sqrt{n_2}(\widehat{\theta}_{\widehat{U},\widehat{u}} - \widehat{\theta}_{U,\widetilde{u}}). \tag{47}$$

The first term on the righthand side in the previous display is the infeasible estimator, were U and $\widetilde{u}$ known. The strategy involves showing that the sample splitting and assumption on the error rates sends the second term to zero uniformly. Establishing the second term goes to zero uniformly requires expanding the difference between the feasible and infeasible estimator, to which we now turn. We start with $\sqrt{n_2}\widehat{\theta}_{\widehat{U},\widehat{u}}$

$$\sqrt{n_2}\widehat{\theta}_{\widehat{U},\widehat{u}} = \sqrt{n_2}\frac{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i - \widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(T_i - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})}{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(T_i - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})^2} \tag{48}$$

Starting with the denominator,

$$\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(T_i - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})^2 \xrightarrow{\text{u}} \mathbb{E}(u_i^2) \tag{49}$$

by convergence in $\widehat{U}$ and $\widetilde{u}$, linearity in $\widetilde{u}$, and the continuous mapping theorem, which is the same limit as the denominator of $\widehat{\theta}_{U,\widetilde{u}}$. We will combine this component with the numerator via Slutsky's theorem later.

Moving on to the numerator, we expand it as

$$\sqrt{n_2}\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i-\widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(T_i-\widehat{T}_{i,\widehat{U}_i,\widehat{u}_i}) \tag{50}$$

$$\begin{aligned}=&\sqrt{n_2}\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i-\widehat{Y}_{i,U_i,\widetilde{u}_i}+\widehat{Y}_{i,U_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i}+\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})\times\\&(T_i-\widehat{T}_{i,U_i,\widetilde{u}_i}+\widehat{T}_{i,U_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}+\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})\end{aligned} \tag{51}$$

which splits into nine terms

$$\begin{aligned}=\sqrt{n_2}\Bigg\{&\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i-\widehat{Y}_{i,U_i,\widetilde{u}_i})(T_i-\widehat{T}_{i,U_i,\widetilde{u}_i})}_{=(a)}+\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i-\widehat{Y}_{i,U_i,\widetilde{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})}_{=(b)}+\\&\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i-\widehat{Y}_{i,U_i,\widetilde{u}_i})(\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})}_{=(c)}\end{aligned} \tag{52}$$

$$\begin{aligned}+&\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,U_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i})(T_i-\widehat{T}_{i,U_i,\widetilde{u}_i})}_{=(d)}+\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,U_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})}_{=(e)}+\\&\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,U_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i})(\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})}_{=(f)}\end{aligned} \tag{53}$$

$$\begin{aligned}+&\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(T_i-\widehat{T}_{i,U_i,\widetilde{u}_i})}_{=(g)}+\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})}_{=(h)}+\\&\underbrace{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(\widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}-\widehat{T}_{i,\widehat{U}_i,\widehat{u}_i})}_{=(i)}\Bigg\}\end{aligned} \tag{54}$$

Term (a) is exactly the numerator in $\widehat{\theta}_{U,\widetilde{u}}$, so we next show that the remaining terms go to zero uniformly. These will do so for two reasons: either by the split-sample approach or

the assumption on the $n^{-1/4}$ rate on the convergence.

To illustrate the logic of sample splitting, we give the sources of variance for each of the six terms and the split from which its variance depends below:

Term	Source fo Variance
$Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i},\ T_i - \widehat{T}_{i,U_i,\widetilde{u}_i}$	$\mathcal{S}_2$
$\widehat{Y}_{i,U_i,\widetilde{u}_i} - \widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i},\ \widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}$	$\mathcal{S}_1$
$\widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i} - \widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i},\ \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i}$	$\mathcal{S}_0$

The first cross term is attributable to variance in $\mathcal{S}_2$, which is what we want: this is the term that will drive our estimation and inference. For the remaining terms, the cross-product terms containing variance generated by different splits are zero, since the two forms of variance come from different data and are conditionally independent. This includes terms (b), (c), (d), (f), (g), (h) above. We illustrate with term (b):

$$\lim_{n\to\infty} \sqrt{n_2}\left\{\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})\right\} \tag{55}$$

$$= \lim_{n\to\infty} \sqrt{n_2}\mathbb{E}\left\{(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})\right\} \tag{56}$$

$$= \lim_{n\to\infty} \sqrt{n_2}\mathbb{E}\left\{\mathbb{E}\left((Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})|i\in\mathcal{S}_1\right)\right\} \tag{57}$$

$$= \lim_{n\to\infty} \sqrt{n_2}\mathbb{E}\left\{(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})\mathbb{E}\left((\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})|i\in\mathcal{S}_1\right)\right\} \tag{58}$$

$$= \lim_{n\to\infty} \sqrt{n_2}\mathbb{E}\left\{(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})\mathbb{E}\left((\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i})\right)\right\} \tag{59}$$

$$= \lim_{n\to\infty} \sqrt{n_2}\mathbb{E}(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})\mathbb{E}(\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}) \tag{60}$$

$$= \lim_{n\to\infty} \sqrt{n_2}\times 0\times 0 \tag{61}$$

$$\xrightarrow{\text{u}} 0 \tag{62}$$

where the lines follow from the law of large numbers, law of iterated expectations, the fact that any variance in $Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i}$ is due to $\mathcal{S}_2$, the random splitting of the sample into the three splits, taking the constant expectation outside of the outer expectation, uniform convergence, and then multiplying by zero.

The remaining cross-terms where both terms contain variance from the same sample are terms (e) and (i), and these go away by assumption. We illustrate with (e):

$$\lim_{n\to\infty} \sqrt{n_2}\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(\widehat{Y}_{i,U_i,\widetilde{u}_i} - \widehat{Y}_{i,\widehat{U}_i,\widetilde{u}_i})(\widehat{T}_{i,U_i,\widetilde{u}_i} - \widehat{T}_{i,\widehat{U}_i,\widetilde{u}_i}) \tag{63}$$

$$\leq \lim_{n\to\infty} \sqrt{n}\left|(\widehat{Y}_{U,\widetilde{u}} - \widehat{Y}_{\widehat{U},\widetilde{u}})(\widehat{T}_{U,\widetilde{u}} - \widehat{T}_{\widehat{U},\widetilde{u}})\right| \tag{64}$$

$$\leq \lim_{n\to\infty} \sqrt{n}\left|\widehat{Y}_{U,\widetilde{u}} - \widehat{Y}_{\widehat{U},\widetilde{u}}\right| \times \left|\widehat{T}_{U,\widetilde{u}} - \widehat{T}_{\widehat{U},\widetilde{u}}\right| \tag{65}$$

$$= \lim_{n\to\infty} \left(n^{1/4}\left|\widehat{Y}_{U,\widetilde{u}} - \widehat{Y}_{\widehat{U},\widetilde{u}}\right|\right) \times \left(n^{1/4}\left|\widehat{T}_{U,\widetilde{u}} - \widehat{T}_{\widehat{U},\widetilde{u}}\right|\right) \tag{66}$$

$$\overset{\text{u}}{\to} 0 \tag{67}$$

where the lines follow because $n_2 \approx n/3$ and the definition of uniform convergence, the Cauchy-Schwarz inequality, distributing $\sqrt{n}$, and then Assumption 2.

With this argument, we have completed our argument that terms $(b) - (i)$ are uniformly $\sqrt{n}$-negligible. Since (a) is the only term that does not disappear, it follows that

$$\lim_{n\to\infty} \sqrt{n_2}\left|\frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i - \widehat{Y}_{i,\widehat{U}_i,\widehat{u}_i})(T_i - \widehat{T}_{i,\widehat{U}_i,\widehat{u}_i}) - \frac{1}{n_2}\sum_{i\in\mathcal{S}_2}(Y_i - \widehat{Y}_{i,U_i,\widetilde{u}_i})(T_i - \widehat{T}_{i,U_i,\widetilde{u}_i})\right| \overset{\text{u}}{\to} 0. \tag{68}$$

Combined with Slutsky's theorem for the denominator, it follows that

$$\lim_{n\to 0} \sqrt{n}|\widehat{\theta}_{\widehat{U},\widehat{u}} - \widehat{\theta}_{U,\widetilde{u}}| \xrightarrow{\text{u}} 0 \tag{69}$$

meaning any difference between our feasible estimator and the infeasible SPE estimator is uniformly $\sqrt{n}$-negligible. This implies our estimator is SPE.

C Derivations for Our Implemented Method

We are now ready to construct the control set U_i and its feasible counterpart $\widehat{U}_i$. In order to do so, we first describe the nonparametric function space within which we estimate.

C.1 The Structure of the Nonparametric Function Space

In order to develop a feasible estimation strategy, we need to add some structure to the nonparametric nuisance functions. We assume the conditional means of the outcome and treatment live in a Reproducing Kernel Hilbert Space (RKHS, see Murphy, 2012; Hainmueller and Hazlett, 2013; Wahba, 1990; van der Vaart and van Zanten, 2008, for a refresher on RKHS theory and Gaussian processes).

We use two key results from RKHS theory. Each involves the "reproducing kernel," (RK) a function that measures the similarity between two points. The key insight is that the basis representation may involve thousands or more possible bases, but the function can be equivalently expressed in terms of this $n \times n$ RK matrix.[24] Modeling in terms of the RK instead of the basis representation is known as the "kernel trick," a well-established tool in

[24]The RK is better thought of as a generalized covariance matrix rather than a kernel, as in kernel density estimation.

machine learning (e.g. Murphy, 2012; Hainmueller and Hazlett, 2013). Our chief innovation here is using a basis representation to estimate the reproducing kernel, a process known as "deconvolution." Our assumption that $U_{i,\widetilde{u}_i}$ is finite dimensional is equivalent to assuming the RK is finite dimensional, or in our setting, the covariance of the basis-representation coefficient fitted values is finite dimensional. By estimating and including principal components of the RK as covariates, we are able to adjust for a broad class of functions in our regression. The split sample, as described above, will guard against overfitting, ensuring an honest confidence interval.

C.1.1 Formalities

An RKHS is a space of functions denote $\mathcal{H}_m$ with generic element m and inner product between between elements m, m' as $\langle m, m' \rangle_{\mathcal{H}_m}$. Functions in an RKHS are bounded and continuous and have finite norm, $0 < \langle m, m \rangle_{\mathcal{H}_m} < C_m < \infty$, for all $m \in \mathcal{H}_m$, and $\langle m, m \rangle_{\mathcal{H}_m} = 0$ for all m not in $\mathcal{H}_m$.

Every RKHS has a function called the "reproducing kernel" (RK) a symmetric function that takes as its arguments two covariates from space $\mathcal{X}$ and returns a real number, $K : \mathcal{X} \times \mathcal{X} \mapsto \Re$.[25].

We utilize two properties of the reproducing kernel. First, the Riesz Representation Theorem guarantees that any function $m \in \mathcal{H}_m$ can be expressed in two equivalent ways.

$$m(x) = \sum_{k=1}^{\infty} B_{m,k}(x) c_k = B_{i,m}^{\top} c_m \tag{70}$$

$$= \sum_{i=1}^{n} K(X_i, x) \alpha_k = K_{i,m}^{\top} \alpha \tag{71}$$

[25]For a gentle introduction to the intuition between an RKHS, see Hainmueller and Hazlett (2013)

By this means, we can see that adjusting for the RK will adjust for any function with that basis representation.

Second, the RK has the property that it "reproduces" the inner product for any function in the space,

$$\langle m, m' \rangle_{\mathcal{H}_m} = \int_{x \in \mathcal{X}} \int_{x' \in \mathcal{X}} m(x) m'(x') K(x, x') dF_x dF_{x'} \tag{72}$$

which, combined with the law of large numbers, will give us the estimate

$$\widehat{\langle m, m' \rangle}_{\mathcal{H}_m} = \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} m(X_i) m'(X_{i'}) K(X_i, X_{i'}) \tag{73}$$

Third, we connect these two using the celebrated Representation Theorem of Craven and Wahba (see Wahba, 1990). Considering the regression problem where $m(x) = \mathbb{E}(Z_i | X_i = x)$, the best estimate of $m(x)$ (in a mean square sense) can be found through solving

$$\widehat{c}_m = \underset{\widetilde{c}_m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Z_i - B_{i,m}^{\top} \widetilde{c}_m)^2 + \frac{1}{n} \lambda_m \widetilde{c}_m^{\top} B_m^{\top} K_m B_m \widetilde{c}_m \tag{74}$$

for some $\lambda_m > 0$, where the first term is the residual sum of squares and the second is the estimate of the norm of the functional. Through this representation, we are able to derive an estimate of the principal components of the $n \times n$ RK evaluated on the data. We then use these principal components to generate a set of controls that will adjust for a broad range of functions.

C.2 Formal Statement of Results

First, we derive closed form representation of the span of the RK for each nuisance space, which will give us the true control set U_f and U_g. We use $\text{sp}(A)$ to denote the span of matrix A, and we say $A \overset{\text{sp}}{\asymp} A'$ if matrices A and A' have the same span. All proofs are in the following section.

Lemma 1 (Connecting the Kernel, Error Variance, and Data) *For $j \in \{f, g\}$ and $\epsilon_j \in \{\epsilon, u\}$, define*

$$M_{\widehat{U},\widehat{u},j} = sp\left\{\frac{1}{n}\left(\text{Var}(\widehat{c}_j)\right)^{-} B_{j,\widehat{u}}^{\top}\text{Var}(\widehat{\epsilon}_j)B_{j,\widehat{u}}\right\} \tag{75}$$

$$M_{U,\widetilde{u},j} = \lim_{n\to\infty} sp\left\{M_{\widehat{U},\widehat{u},j}\right\} \tag{76}$$

Under Assumptions 1 - 3, we can express $U_{j,\widetilde{u}} \overset{\text{sp}}{\asymp} B_{j,\widetilde{u}}M_{U,\widetilde{u},j}$.

These estimates are infeasible, since we observe neither variance in $M_{\widehat{U},\widehat{u},j}$. We next derive a feasible estimate. Doing so requires sample-splitting subset $\mathcal{S}_1$ into $\mathcal{S}_{1a}$ and $\mathcal{S}_{1b}$, creating a crucial conditional independence between the coefficient estimates and estimated residuals. We use cross-fitting to recover an M matrix. We denote as $\text{Var}^{*,C}(A)$ the bootstrap estimated variance of vector A on split C. Note that $\widehat{u}$ is estimated on split $\mathcal{S}_0$ but evaluated on $\mathcal{S}_1$.

Lemma 2 (Feasible Estimation of the Kernel) *For $j \in \{f, g\}$ and $\widehat{\epsilon}_j \in \{\widehat{\epsilon}, \widehat{u}\}$,*

$$M^{*}_{\widehat{U},\widehat{u},j} = sp\left\{\frac{1}{n_1}\left(\text{Var}^{*,\mathcal{S}_{1a}}(\widehat{c}_j)\right)^{-} B_{j,\widehat{u}}^{\top}\text{Var}^{*,\mathcal{S}_{1b}}(\widehat{\epsilon}_j)B_{j,\widehat{u}}\right\} \tag{77}$$

Denote a matrix $\widehat{U}_{j,\widehat{u}}$ *such that* $\widehat{U}_{j,\widehat{u}} \overset{\text{sp}}{\asymp} B_{j,\widehat{u}} M^*_{\widehat{U}_{j,\widehat{u}}}$, *and* $\Pi_{\widehat{U}_{j,\widehat{u}}}$ *and* $\Pi_{U_{j,\tilde{u}}}$ *the least-squares projection matrices of an arbitrary unit vector of length* n_1 *onto the matrices in the subscript.*

Under Assumptions 1 - 3 and Lemma 1,

$$\lim_{n\to\infty} n^{1/4} \left| \Pi_{\widehat{U}_{j,\widehat{u}}} - \Pi_{U_{j,\tilde{u}}} \right| \overset{\text{u}}{\to} 0 \tag{78}$$

in operator norm.

Our final proposition is given in text as Proposition 2. It follows directly from combining Lemma 2, which guarantees that we get $n^{-1/4}$ rates on estimation error, and Proposition 1.

The formulation offers two major advantages. First, given our estimates of the covariates from sample $\mathcal{S}_1$, our estimate of θ can be calculated from a simple regression. Second, the familiar "robust" standard errors are consistent (e.g. Long and Ervin, 2000), making the standard error calculations straightforward. We utilize cross-fitting, so we rotate each of the subsamples as the estimation sample, averaging point estimates and standard errors.

C.3 Proof of Lemma 1.

To reduce notation, we suppress notation indicating that all calculations are done on split sample $\mathcal{S}_1$, and, for the treatment model, the basis representation are evaluated using estimated errors from sample $\mathcal{S}_0$. We denote as $\widehat{\epsilon}_j \in \{\widehat{\epsilon}, \widehat{u}\}$ with i^{th} element $\widehat{\epsilon}_{j,i}$

Since the conditional means are in a reproducing kernel Hilbert space, the estimates $\widehat{c}_j$

are sample minimizers of

$$\mathcal{L}_{K,\lambda_j}(\widetilde{c}_j) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - B_{j,i}^\top \widetilde{c}_f)^2 + \frac{1}{n}\lambda_j \widetilde{c}_j^\top B_j^\top K_j B_j \widetilde{c}_j \tag{79}$$

$$\widehat{c}_j = \underset{\widetilde{c}_j}{\operatorname{argmin}}\, \mathcal{L}_{K_j,\lambda_j}(\widetilde{c}_j) \tag{80}$$

for some K_j, λ_j. The estimating equations give

$$\partial_{\widetilde{c}_j}\mathcal{L}_{K_j,\lambda_j}(\widetilde{c}_j)\big|_{\widetilde{c}_j=\widehat{c}_j} = 0 \Rightarrow \tag{81}$$

$$\frac{1}{n}\sum_{i=1}^{n} B_{j,i}\widehat{\epsilon}_{j,i} = \frac{1}{n}\lambda_j B_j^\top K_j B_j \widehat{c}_j \tag{82}$$

and taking the variance of both sides

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} B_{j,i}\widehat{\epsilon}_{j,i}\right) = \frac{\lambda_j^2}{n^2} B_j^\top K_j B_j \mathrm{Var}(\widehat{c}_j) B_j^\top K_j B_j \tag{83}$$

Noting that $\frac{1}{n}B_j^\top B_j$ converges to a Gram matrix, since the bases are bounded, a nondegenerate solution requires $\lambda_j = O_p(1/\sqrt{n})$

Let $U_{j,k}$ be an arbitrary unit vector in the span of K_j. Right-multiplying by $B_j^\top U_{j,k}$ and rearranging, with the inverse with respect to the range of K_j, gives

$$\left\{\frac{1}{n^2} B_f^\top K_j B_j \mathrm{Var}(\widehat{c}_j) B_j^\top K_j B_j\right\}^{-} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} B_{j,i}\widehat{\epsilon}_{j,i}\right) B_j^\top U_{j,k} = N\lambda^2 B_j^\top U_{j,k}, \tag{84}$$

showing that $U_{j,k}$ is a singular vector of the matrix on the lefthand side.

We now show that $span(U_j) = span\{\mathrm{Var}(\widehat{c}_j)^{-} B_f^\top \mathrm{Var}(\widehat{\epsilon})\}$. First, denote V_1 as an arbitrary

vector in $span\{\text{Var}(\widehat{c}_j)^- B_j^\top \text{Var}(\widehat{\epsilon}_j)\}$. V_1 is, therefore, in the span of the lefthandside of Equation 84, and hence in the span of U_j. Next, denote V_2 an arbitrary element of $span(U_j)$. V_2 is therefore additive in the singular vectors $U_{j,k}$, and therefore satisfies equation 84. Therefore, the two spaces are equal.

C.4 Proof of Lemma 2.

First, the bootstrap variance estimate is valid by the uniform convergence of $\widehat{c}_j$ and van der Vaart (1998, ch. 23.2). The second claim follows through three steps. First, the error attributable to $\widehat{u}_i$ in $B_{i,j,\widehat{u}_i}$ is $o_p(n^{-1/4})$ uniformly. Second, the projection matrix is a continuous function of $\widehat{c}_j^2$, so it converges to zero faster than $\widehat{c}_j$. Third, since the two errors come from different samples, they are conditionally independent. Therefore, the total error rate is the larger of the two, which is $o_p(n^{-1/4})$. Combined, the result gives uniform convergence in the projection matrices.

D Estimation Details

We implement a tensor-product smoothing spline; see James et al. (2013) for an accessible introduction. Specifically, we include in all conditional means the original covariates in X_i. We then include all degree 3 and 5 B-splines for each covariate as well as all two-way interactions among the spline terms. We use a marginal correlation screen to reduce this large number of bases down to the $100 \times (1 + n^{1/5})$ with the largest correlation, with a minimum of 50 and maximum of 400. For estimation, we implement a sparse Bayesian regression model (Ratkovic and Tingley, 2017).

For the sample splitting, we split in thirds at random. For splitting with random effects,

we split to ensure there is approximate balance of splits within each level of the grouping of the random effect.

E Simulation Study

We have designed a simulations study to illustrate how existing methods respond in the face of treatment effect heterogeneity and to treatment variance bias. In each setting, we draw a single covariate X_i and error terms u_i and ϵ_i, each standard normal, with the covariate standardized so that $\frac{1}{n}\sum_{i=1}^{n} X_i = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = 1$. Four additional normal noise covariates are included, with pairwise correlations among the covariates 0.5, but only the first is used to generate the treatment and the outcome.

In order to consider the impact of treatment effect heterogeneity, we vary whether the treatment interacts with the covariate X_i: The two models are given below:

	The Additive Model	The Interactive Model
Outcome Model:	$Y_i = T_i + X_i^2 + \epsilon_i$	$Y_i = T_i \times X_i^2 + \epsilon_i$
Treatment Model:	$T_i = X_i + v_i$	$T_i = X_i + v_i$

In the first model, X_i^2 enters additively, so the treatment effect is homogenous. In the second, it enters interactively, inducing a simple, but nonlinear, heterogeneity. In both models, the sample average causal effect is 1, due to the homogenous effect in the additive model and the scaling of X_i in the interactive model.

We then allow for two models of the allow for two specifications of treatment variance,

$$\text{No Treatment Heterogeneity: } \epsilon_i \sim \mathcal{N}(0,1) \;\; u_i \sim \mathcal{N}(0,1)$$

$$\text{Treatment Heterogeneity: } \epsilon_i \sim \mathcal{N}\left(0, \frac{X_i^2+1}{2}\right) \;\; u_i \sim \mathcal{N}\left(0, \frac{X_i^2+1}{2}\right)$$

Because we have scaled X_i^2 to have mean one, the errors have the same variance, $\mathbb{E}(v_i^2) = \mathbb{E}(\epsilon_i^2) = 1$ across both settings. We vary the sample size $n \in \{250, 500, 750, 1000\}$

We first consider the accuracy of each method's estimates in Figure 2, considering the setting with $n = 500$.[26] The first row contains results with no treatment heterogeneity; the second row contains results with treatment heterogeneity. The columns present the results from the model with (left) and without (right) treatment effect heterogeneity. PLCE is the leftmost boxplot, in red, with the additional methods to the right: Double Machine Learning (DML), Generalized Random Forests (GRF), the CBPS for continuous treatments (CBPS),[27] and least squares (OLS). The true value is 1, in both sample and population, and is denoted with a horizontal line.

Starting in the top left corner, we consider the model with homogenous effects and homoskedastic errors. As expected, every method performs well in this situation, though DML and KRLS appear to have some visible downwards bias. The top right figure presents results with effect heterogeneity but homoskedastic errors. The introduction of this heterogeneity, even with equivariant errors, induces a visible bias in all methods except OLS and PLCE.

[26] Results are qualitatively at other sample size settings, except for GRF in the lower righthand corner, which does much worse at $n = 250$ and much better at $n = 1000$.

[27] In this simulation, we use parametric CBPS, so that we can recover standard error estimates. So as not to handicap the method, we give it both the covariates and their square terms, so the true generative model is being balanced.

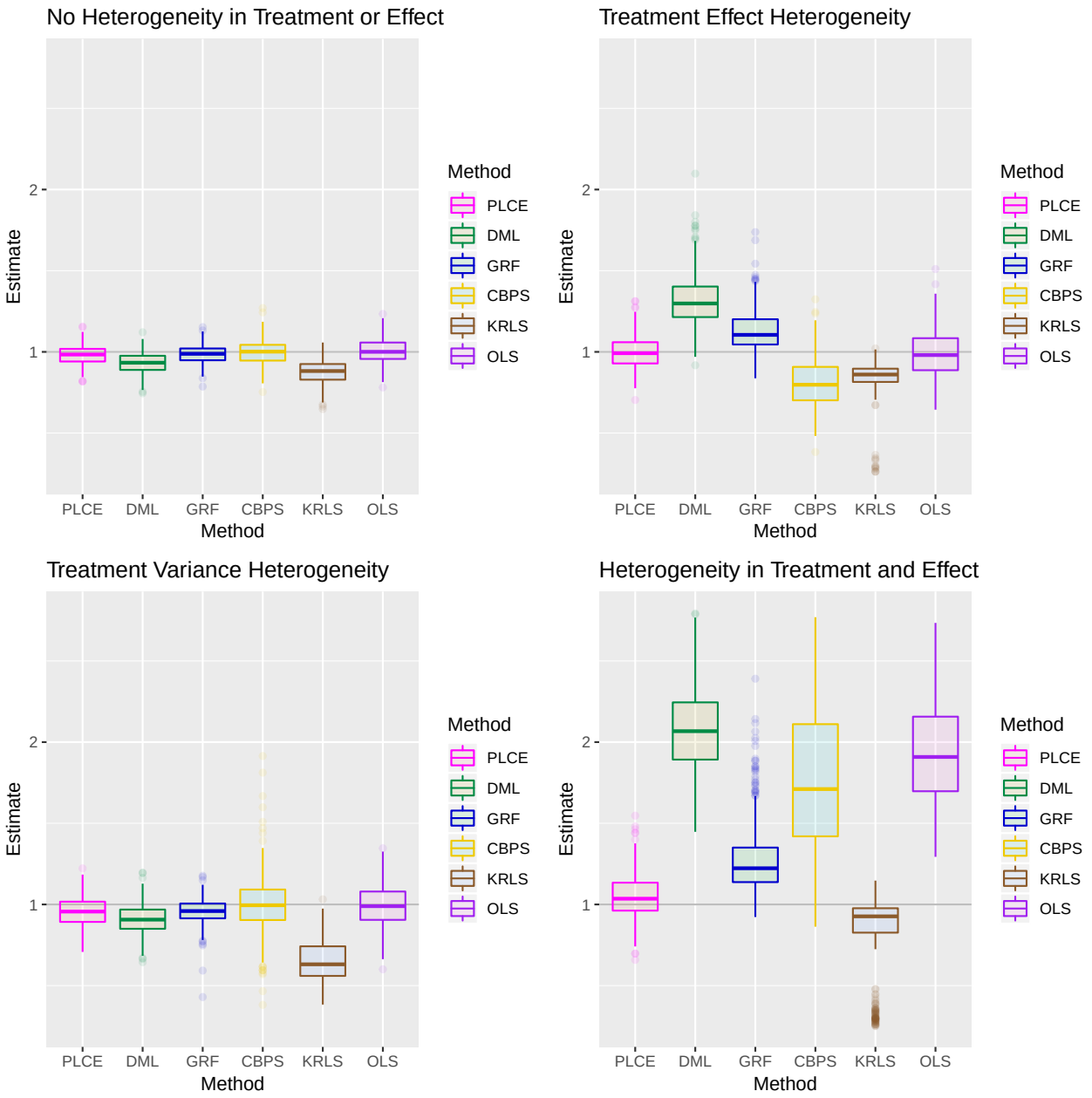

Figure 2: Distributions of causal estimates ($\theta_{\text{true}} = 1$).

We move next to the bottom row, with treatment heterogeneity. Starting in the left column, we need not worry about treatment variance bias with a homogenous treatment effect over the sample. Only KRLS and DML again evidence a downward bias. This homogeneity assumption, though, is almost certainly hard to justify from a substantive standpoint. The

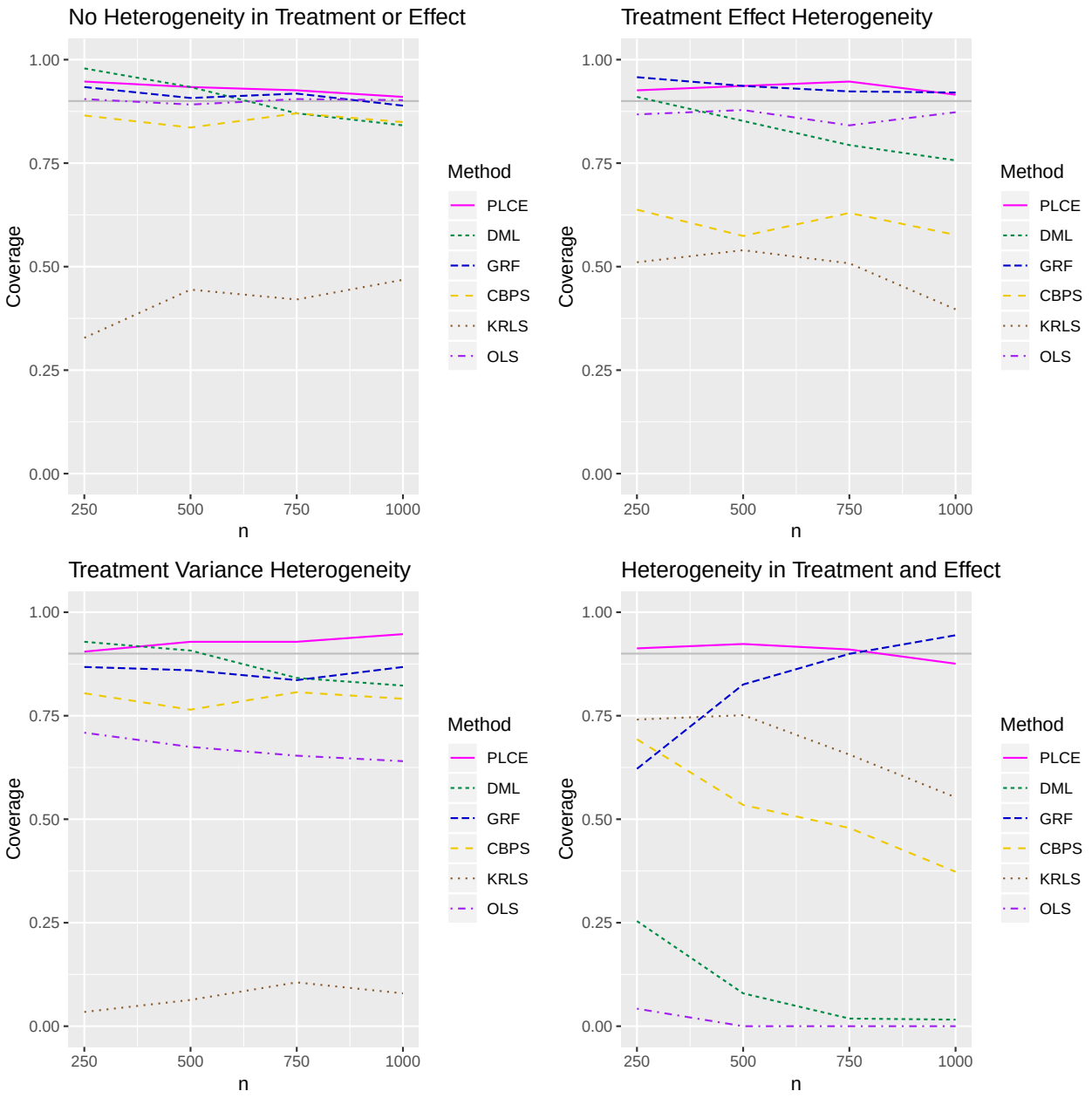

Figure 3: Coverage of 90% confidence intervals averaged over 1,000 simulations.

figure on the bottom-right illustrates the interplay of both forms of heterogeneity. Now, DML, CBPS, and, as expected, OLS display a strong positive bias. KRLS still displays its attenuation bias, while PLCE does well and GRF exhibits little bias.

We see that both GRF and PLCE perform well in terms of the accuracy of the point

estimates. Point estimation, though, is not our only concern. We are equally concerned with whether the method allows for valid inference. To assess validity, for each simulation, we constructed 90% confidence intervals centered on the point estimate and 1.64 standard errors above and below. The "coverage rate" is the proportion of simulations for which the constructed confidence interval contains the true value of 1. If the confidence interval is valid, the coverage rate is above 90%.

Results for confidence intervals are in Figure 3, with the simulation the same as in the previous figure. The gray line is at the nominal rate of 90%; methods placing below above the line are conservative and those below lead to invalid inference. All methods save KRLS are valid in the absence of treatment effect heterogeneity and treatment variance heterogeneity. In the presence of treatment effect heterogeneity, GRF, PLCE, and OLS return valid confidence intervals. With treatment variance heterogeneity, all methods except PLCE return invalid confidence intervals, with GRF falling consistently below 90%. In the presence of both forms of heterogeneity, only PLCE and GRF perform well, with coverage rates often below 50%.

In sum, across the settings for this simulation, only PLCE returns reasonably accurate point estimates and valid confidence intervals in the presence of treatment heterogeneity and treatment effect heterogeneity.

E.1 Efficiency Concerns

While theoretical results guarantee that the cross-fitting estimates are consistent and asymptotically efficient, we understand the hesitant researcher who may prefer some finite-sample assurances. We report estimates on the standard errors from the simulations above in Ta-

n	PLCE	OLS
250	0.103	0.110
500	0.074	0.078
750	0.060	0.064
1000	0.051	0.055

Table 2: **Efficiency of Least Squares versus PLCE.** The table reports average standard errors for PLCE and OLS from simulation setting 1 above, when the model is in-truth linear and errors are homoskedastic. Standard errors from PLCE are comparable to those from OLS.

ble 2. The lefthand table reports on standard errors for PLCE and OLS from simulation setting 1 above, when the model is in-truth linear and errors are homoskedastic. Standard errors from PLCE are comparable to, and actually slightly less than, those from OLS. We, of course, cannot guarantee smaller standard errors than OLS, but we can suggest that the standard errors from PLCE should not be too much larger if the linear model is in-truth correct.