doi:10.1017/S0003055422001022

Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression

MARC RATKOVIC *Princeton University, United States*

Valid inference in an observational study requires a correct control specification, but a correct specification is never known. I introduce a method that constructs a control vector from the observed data that, when included in a linear regression, adjusts for several forms of bias. These include nonlinearities and interactions in the background covariates, biases induced by heterogeneous treatment effects, and specific forms of interference. The first is new to political science; the latter two are original contributions. I incorporate random effects, a set of diagnostics, and robust standard errors. With additional assumptions, the estimates allow for causal inference on both binary and continuous treatment variables. In total, the model provides a flexible means to adjust for biases commonly encountered in our data, makes minimal assumptions, returns efficient estimates, and can be implemented through publicly available software.

INTRODUCTION

The standard linear regression is the field's most commonly encountered quantitative tool, used to estimate effect sizes, adjust for background covariates, and conduct inference. At the same time, the method requires a set of assumptions that have long been acknowledged as problematic (e.g., Achen 2002; Leamer 1983; Lenz and Sahn 2021; Sami 2016). The fear that quantitative inference will reflect these assumptions rather than the design of the study and the data has led our field to explore alternatives including estimation via machine learning (e.g., Beck and Jackman 1998; Beck, King, and Zheng 2000; Grimmer, Messing, and Westwood 2017; Hill and Jones 2014) and identification using the analytic tools of causal inference (e.g., Acharya, Backwell, and Sen 2016; Imai et al. 2011).

I integrate these two literatures tightly, formally, and practically, with a method and associated software that can improve the reliability of quantitative inference in political science and the broader social sciences. In doing so, I make two contributions. First, I introduce to political science the concepts and strategies necessary to integrate machine learning with the standard linear regression model (Athey, Tibshirani, and Wager 2019; Chernozhukov et al. 2018). Second, I extend this class of models to address two forms of bias of concern to political scientists. Specifically, I adjust for a bias induced by unmodeled treatment effect heterogeneity, highlighted by Aronow and Samii (2016). In correcting this bias, and under additional assumptions on the data, the proposed method allows for causal effect estimation whether the treatment variable is continuous or binary. I also adjust for biases induced by exogenous interference, which occurs when an observation's outcome or treatment is affected by the characteristics of other observations (Manski 1993).

The goal is to allow for valid inference that does not rely on a researcher-selected control specification. The proposed method, as with several in this literature, uses a machine learning method to adjust for background variables while returning a linear regression coefficient and standard error for the treatment variable of theoretical interest. Following the double machine learning approach of Chernozhukov et al. (2018), my method implements a split-sample strategy. This consists of, first, using a machine learning method on one part of the data to construct a control vector that can adjust for nonlinearities and heterogeneities in the background covariates as well as the two biases described above. Then, on the remainder of the data, this control vector is included in a linear regression of the outcome on the treatment. The split-sample strategy serves as a crucial guard against overfitting. By alternating which subsample is used for constructing control variables from the background covariates and which subsample is used for inference and then aggregating the separate estimates, the efficiency lost by splitting the sample can be regained. I illustrate this on experimental data, showing that the proposed method generates point estimates and standard errors no different than those from a full-sample linear regression model.

My primary audience is the applied researcher currently using a linear regression for inference but who may be unsettled by the underlying assumptions. I develop the method first as a tool for descriptive inference, generating a slope coefficient and a standard error on a variable of theoretical interest but relying on machine learning to adjust for background covariates. I then discuss the assumptions necessary to interpret the coefficient as a causal estimate. In order to encourage adoption of the proposed method, software for implementing the proposed method and the diagnostics

Marc Ratkovic, Assistant Professor, Department of Politics, Princeton University, United States, ratkovic@princeton.edu.

Received: March 17, 2019; revised: June 04, 2020; accepted: September 13, 2022.

described in this manuscript are available on the Comprehensive R Archive Network in the package PLCE.

IMPLICATIONS AND APPLICATIONS OF THE PROPOSED METHOD

Quantitative inference in an observational setting requires a properly specified model, meaning the control variables must be observed and entered correctly by the researcher in order to recover an unbiased estimate of the effect of interest. A correct specification, of course, is never known, raising concerns over "model-dependent" inference (King and Zeng 2006).

Contrary advice on how to specify controls in a linear regression remains unresolved. This advice ranges from including all the relevant covariates but none of the irrelevant ones (King, Keohane, and Verba 1994, secs. 5.2–5.3), which states rather than resolves the issue; including at most three variables (Achen 2002); or at least not all of them (Achen 2005); or maybe none of them (Lenz and Sahn 2021). Others have advocated for adopting machine learning methods including neural nets (Beck, King, and Zeng 2000), smoothing splines (Beck and Jackman 1998), nonparametric regression (Hainmueller and Hazlett 2013), tree-based methods (Hill and Jones 2014; Montgomery and Olivella 2018), or an average of methods (Grimmer, Messing, and Westwood 2017).

None of this advice has found wide use. The advice on the linear regression is largely untenable, given that researhers normally have a reasonable idea of which background covariates to include but cannot guarantee that an additive, linear control specification is correct. Machine learning methods offer several important uses, including prediction (Hill and Jones 2014) and uncovering nonlinearities and heterogeneities (Beck, King, and Zeng 2000; Imai and Ratkovic 2013). Estimating these sorts of conditional effects and complex models are useful in problems that involve prediction or discovery. For problems of inference, where the researcher desires a confidence interval or p-value on a regression coefficient, these methods will generally lead to invalid inference, a point I develop below and illustrate through a simulation study.

Providing a reliable and flexible means of controlling for background covariates and clarifying when and whether the estimated effect admits a causal interpretation is essential to the accumulation of knowledge in our field. I provide such a strategy here.

Turning Toward Machine Learning

Conducting valid inference with a linear regression coefficient without specifying how the control variables enter the model has long been studied in the fields of econometrics and statistics (see, e.g., Bickel et al. 1998; Newey 1994; Robinson 1988; van der Vaart 1998, esp. chap. 25). Recent methods have brought these theoretical results to widespread attention by combining machine learning methods to adjust for background covariates with a linear regression for the variable of interest, particularly the double machine learning approach of Chernozhukov et al. (2018) and the generalized random forest of Athey, Tibshirani, and Wager (2019). I work in this same area, introducing the main concepts to political science.

Although well-developed in cognate fields, political methodologists have put forth several additional critiques of linear regression left unaddressed by these aforementioned works. The first critique comes from King (1990) in the then-nascent subfield of political methodology. In a piece both historical and forward-looking, he argued that unmodeled geographic interference was a first-order concern of the field. More generally, political interactions are often such that interference and interaction across observations is the norm. Most quantitative analyses simply ignore interference. Existing methods that do address it rely on strong modeling assumptions requiring, for example, that interference is being driven by known covariates, like ideology (Hall and Thompson 2018), or that observations only affect similar or nearby observations either geographically or over a known network (Aronow and Samii 2017; Ripley 1988; Sobel 2006; Ward and O'Loughlin 2002), and these models only allow for moderation by covariates specified by the researcher. I extend on these approaches, offering the first method that learns and adjusts for general patterns of exogenous interference.

The second critique emphasizes the limits on using a regression for causal inference in observational studies (see, e.g., Angrist and Pischke, 2010, sec 3.3.1). From this approach, I adopt three concerns. The first is a careful attention to modeling the treatment variable. The second is precision in defining the parameter of interest as an aggregate of observation-level effects. Aronow and Samii (2016) show that a correlation between treatment effect heterogeneity and variance in the treatment assignment will cause the linear regression coefficient to be biased in estimating the causal effect—even if the background covariates are included properly. I offer the first method that explicitly adjusts for this bias. Third, I provide below a set of assumptions that will allow for a causal interpretation of the estimate returned by the proposed method.

Practical Considerations of the Proposed Method

The major critique of the linear regression, that its assumptions are untenable, is hardly new. Despite this critique, the linear regression has several positive attributes worthy of preservation. First is its transparency and ease of use. The method, its diagnostics, assumptions, and theoretical properties are well-understood and implemented in commonly available software, and they allow for easy inference. Coefficients and standard errors can be used to generate confidence intervals and p-values, and a statistically significant result provides a necessary piece of evidence that a hypothesized association is present in the data. Importantly, the proposed method maintains these advantages.

I illustrate these points in a simulation study designed to highlight blind spots of existing methods. I then

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

reanalyze experimental data from Mattes and Weeks (2019), showing that if the linear regression model is in fact correct, my method neither uncovers spurious relationships in the data nor comes at the cost of inflated standard errors. In the second application, I illustrate how to use the method with a continuous treatment. Enos (2016) was forced to dichotomize a continuous treatment, distance from public housing projects, in order to estimate the causal effect of racial threat. To show his results were not model-dependent, he presented results from dichotomizing the variable at 10 different distances. The proposed method handles this situation more naturally, allowing a single estimate of the effect of distance from housing projects on voter turnout.

ANATOMY OF A LINEAR REGRESSION

My central focus is in improving estimation and inference on the *marginal effect*, which is the average effect of a one unit move in a variable of theoretical interest t_i on the predicted value of an outcome y_i, after adjusting for background covariates $\mathbf{x}_i$.[1] I will denote the marginal effect as θ.

Estimation of the marginal effect is generally done with a linear regression,

$$y_i = \theta t_i + \mathbf{x}_i^{\mathrm{T}}\gamma + e_i; \quad \mathbb{E}(e_i|\mathbf{x}_i, t_i) = 0, \tag{1}$$

where the marginal effect θ is the *target parameter*, meaning the parameter on which the researcher wishes to conduct inference.

I will refer to terms constructed from the background covariates $\mathbf{x}_i$ and entered into the linear regression as *control variables*. For example, when including a square term of the third variable x_{3i} in the linear regression, the background covariate vector is $\mathbf{x}_i$ but the control vector is now $\left[\mathbf{x}_i^{\mathrm{T}}, x_{3i}^2\right]^{\mathrm{T}}$. I will reserve γ for slope parameters on control vectors.

Inference on a parameter is *valid* if its point estimate and standard error can be used to construct confidence intervals and p-values with the expected theoretical properties. Formally, $\hat{\theta}$ and its estimated standard deviation $\hat{\sigma}_{\hat{\theta}}$ allow for valid inference if, for any θ,

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}\right) \rightsquigarrow \mathbb{N}(0, 1). \tag{2}$$

The *limiting distribution* of a statistic is the distribution to which its sampling distribution converges (see Wooldridge 2013, app. C12), so in the previous display the limiting distribution of the z-statistic on the left is a standard normal distribution.

The remaining elements of the model, the control specification $(\mathbf{x}_i^{\mathrm{T}}\gamma)$ and the distribution of the error term, are the *nuisance components*, meaning they are not of direct interest but need to be properly adjusted for in order to allow valid inference on θ. The component with the control variables is *specified* in that its precise functional form is assumed by the researcher.

Heteroskedasticity-consistent (colloquially, "robust") standard errors allow valid inference on θ without requiring the error distribution to be normal, or even equivariant.[2] In this sense, the error distribution is *unspecified.* This insight proves critical to the proposed method: valid inference in a statistical model is possible even when components of the model are unspecified.[3]

In statistical parlance, the linear regression model fit with heteroskedasticity-consistent standard errors is an example of a *semiparametric model*, as it combines both a specified component $(\theta t_i + \boldsymbol{x}_i^{\mathrm{T}}\gamma)$ and an unspecified component, the error distribution.

This chain of reasoning then begs the question, can even less be specified? And, does the estimated coefficient admit a causal interpretation? I turn to the first question next and then address the second in the subsequent section.

MOVING BEYOND LINEAR REGRESSION

In moving beyond linear regression, I use a machine learning method to construct a control vector that can be included in a linear regression of the outcome on the treatment. This vector will allow for valid inference on θ even in the presence of unspecified nonlinearities and interactions in the background covariates. This section relies on the development of double machine learning in Chernozhukov et al. (2018) and the textbook treatment of van der Vaart (1998). The presentation remains largely informal, with technical details available in Appendix A of the Supplementary Materials. I then extend this approach in the next section.

The Partially Linear Model

Rather than entering the background covariates in an additive, linear fashion, they could enter through unspecified functions, f, g:[4]

$$y_i = \theta t_i + f(\boldsymbol{x}_i) + e_i; \ \mathbb{E}(e_i|t_i, \boldsymbol{x}_i) = 0. \tag{3}$$

$$t_i = g(\boldsymbol{x}_i) + v_i; \quad \mathbb{E}(v_i|t_i, \boldsymbol{x}_i) = \mathbb{E}(e_i v_i|\boldsymbol{x}_i) = 0. \tag{4}$$

[1] The marginal effect is sometimes referred to as the *average partial effect.*

[2] For more on the use and misuse of heteroskedasticity-consistent standard errors, Freedman (2006) notes that they are not useful if the model is misspecified; King and Roberts (2015) propose using disagreement between analytic and heteroskedasticity-consistent as a model diagnostic but note Aronow's (2016) critique of this approach as overreliant on modeling assumptions. My view aligns most closely with Aronow (2016) and derives from the general approach in van der Vaart (1998).

[3] Unspecified does not mean arbitrary. Heteroskedasticity-consistent standard errors require that the residuals be mean zero given the covariates and treatment and that the estimated residuals follow the central limit theorem; see White (1980, Assumptions 2 and 3). This includes distributions commonly encountered in observational data while excluding fat-tailed distributions like the Cauchy.

[4] Although the covariates can enter the model nonlinearly, the estimate will still be linear in the sense of being additive in the outcome variable (Wooldridge 2013, sec. 2.4.).

The resulting model is termed the *partially linear model*, as it is still linear in the treatment variable but the researcher need not assume a particular control specification. This model subsumes the additive, linear specification, but the functions f, g also allow for nonlinearities and interactions in the covariates.

Semiparametric Efficiency

With linear regression, where the researcher assumes a control specification, the least squares estimates are the uniformly minimum variance unbiased estimate (e.g., Wooldridge 2013, sec. 2.3). This efficiency result does not immediately apply to the partially linear model, as a particular form for f, g is not assumed in advance but instead learned from the data. We must instead rely on a different conceptualization of efficiency: semiparametric efficiency.

An estimate of θ in the partially linear model is semiparametrically efficient if, first, it allows for valid inference on θ and, second, its variance is asymptotically indistinguishable from an estimator constructed from the true, but unknown, nuisance functions f, g. Establishing this property proceeds in two broad steps.[5] The first step involves constructing an estimate of θ assuming the true functions f, g were known. This estimate is *infeasible*, as it is constructed from unknown functions. The second step then involves providing assumptions and an estimation strategy such that the feasible estimate constructed from the estimated functions $\hat{f}, \hat{g}$ shares the same limiting distribution as the infeasible estimate constructed from f, g.

For the first step, consider the *reduced form* model that combines the two models in Equations 3 and 4,

$$y_i = \theta t_i + [f(\boldsymbol{x}_i), g(\boldsymbol{x}_i)]\gamma + e_i. \tag{5}$$

If f, g were known, θ could be estimated efficiently using least squares.[6] The estimate, $\hat{\theta}$ will be efficient and allow for valid inference on θ.

Following Stein (1956), we would not expect any feasible estimator to outperform this infeasible estimator, so its limiting distribution is termed the *semiparametric efficiency bound.* With this bound established, I now turn to generating a feasible estimate that shares a limiting distribution with this infeasible estimate.

Double Machine Learning for Semiparametrically Efficient Estimation

Estimated functions $\hat{f}, \hat{g}$, presumably estimated using a machine learning method, can be used to construct and enter control variables into a linear regression as

$$y_i = \theta t_i + \left[\hat{f}(\mathbf{x}_i), \hat{g}(\mathbf{x}_i)\right]\gamma + e_i. \tag{6}$$

Semiparametric efficiency can be established by characterizing and eliminating the gap between the infeasible model in Equation 5 and the feasible model in Equation 6. I do so by introducing *approximation error* terms,

$$\varDelta_{\hat{f},i} = \hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i); \;\; \Delta_{\hat{g},i} = \hat{g}(\mathbf{x}_i) - g(\mathbf{x}_i), \tag{7}$$

that capture the distance between the true functions f, g and their estimates $\hat{f}, \hat{g}$ at each $\mathbf{x}_i$.

Given these approximation errors, Equation 6 can be rewritten in the familiar form of a *measurement error* problem (Wooldridge 2013, chap. 9.4), where the estimated functions $\hat{f}, \hat{g}$ can be thought of as "mismeasuring" the true functions f, g:

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma_1 + \left[\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i), \hat{g}(\mathbf{x}_i) - g(\mathbf{x}_i)\right]\gamma_2 + e_i. \tag{8}$$

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma_1 + \left[\varDelta_{\hat{f},i}, \varDelta_{\hat{g},i}\right]\gamma_2 + e_i. \tag{9}$$

Establishing semiparametric efficiency of a feasible estimator, then, consists of establishing a set of assumptions and an estimation strategy that leaves the approximation error terms asymptotically negligible.

There are two pathways by which the approximation error terms may bias an estimate. The first is if the approximation errors do not tend toward zero. Eliminating this bias requires that the approximation errors vanish asymptotically, specifically at an $n^{1/4}$ rate.[7] Though seemingly technical, this assumption is actually liberating. Many modern machine learning methods that are used in political science provably achieve this rate (Chernozhukov et al. 2018), including random forests (Hill and Jones 2014; Montgomery and Olivella 2018), neural networks (Beck, King, and Zeng 2000), and sparse regression models (Ratkovic and Tingley 2017). This assumption allows the researcher to condense all the background covariates into a control vector constructed from $\hat{f}, \hat{g}$, where these functions are estimated via a flexible machine learning method. Any nonlinearities and interactions in the background

[5] Appendix A in the Supplementary Materials contains a complete, self-contained technical discussion.

[6] There are often multiple and asymptotically equivalent ways to estimate θ (see, e.g., Robins et al. 2007). In their double machine learning algorithm, Chernozhukov et al. (2018) propose regressing $y_i - f(\mathbf{x}_i)$ on $t_i - g(\mathbf{x}_i)$, whereas I instead estimate θ from the reduced form model. The two are asymptotically equivalent, but I favor the reduced form approach because it more easily incorporates intuitions and diagnostics from linear regression.

[7] Formally, this requires

$$n^{1/4}\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\Delta_{\hat{f},i}^2} \overset{u}{\rightarrow} 0; \;\; n^{1/4}\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\varDelta_{\hat{g},i}^2} \overset{u}{\rightarrow} 0, \tag{10}$$

where $\overset{\mathrm{u}}{\rightarrow}$ denotes converges uniformly, which is the notion of convergence needed for complex, nonparametric functions. I provide full details in Appendix A of the Supplementary Materials.

TABLE 1. The Double Machine Learning Algorithm of Chernozhukov et al. (2018)

Algorithm 1: The Double Machine Learning Algorithm

Outcome, treatment, and covariates$\{y_i, t_i, \mathbf{x}_i\}_{i=1}^n$

```
Result: for r in 1 to R do
    Split the sample in half, generating $\mathcal{S}_1, \mathcal{S}_2$.
    for j in 1 to 2 do
        Estimate $f, g$ in subsample $\mathcal{S}_j$ using a machine learning method.
        Regress $y_i - \widehat{f}(\mathbf{x}_i)$ on $t_i - \widehat{g}(\mathbf{x}_i)$ using data from the other subsample.
    end
    Aggregate the point estimate, $\widehat{\theta}$ and standard error, $\widehat{\sigma}_{\widehat{\theta}}$, over splits.
end
```

Note:The double machine learning algorithm combines machine learning, to learn how the covariates enter the model, with a regression for the coefficient of interest. Each step is done on a separate subsample of the data (*sample-splitting*), the roles of the two subsamples are swapped (*cross-fitting*), and the estimate results from aggregating over multiple cross-fit estimates (*repeated cross-fitting*). The proposed method builds on these strategies while adjusting for several forms of bias ignored by the double machine learning strategy.

covariates are then learned from the data rather than specified by the researcher.

Eliminating the second pathway requires that any correlation between the approximation errors $\Delta_{\hat{f},i}$, $\Delta_{\hat{g},i}$ and the error terms e_i,v_i tend toward zero.[8] Doing so requires addressing a subtle aspect of the approximation error: the estimates $\hat{f}$,$\hat{g}$ are themselves functions of e_i,v_i, as they are estimates constructed from a single observed sample. Even under the previous assumption on the convergence rate of $\hat{f}$,$\hat{g}$, this bias term may persist.

The most elegant, and direct, way to eliminate this bias is to employ a *split-sample* strategy, as shown in Table 1.[9] First, the data are split in half into subsamples denoted $\mathcal{S}_1$ and $\mathcal{S}_2$ of size n_1 and n_2 such that $n_1 + n_2 = n$. Data from $\mathcal{S}_1$ are used to learn $\hat{f}$,$\hat{g}$ and data from $\mathcal{S}_2$ to conduct inference on θ. Because the nuisance functions are learned on data wholly separate from that on which inference is conducted, this bias term tends toward zero. The resultant estimate is semiparametrically efficient, under the conditions given in in 4.3.

Sample-splitting raises real efficiency concerns, as it uses only half the data for inference and thereby inflates standard errors by $\sqrt{2}\approx 1.4$. In order to restore efficiency, double machine learning implements a *cross-fitting* strategy, whereby the roles of the subsamples $\mathcal{S}_1$,$\mathcal{S}_2$ are swapped and the estimates combined. *Repeated cross-fitting* consists of aggregating estimates over multiple cross-fits, allowing all the data to be used in estimation and returning results that are not sensitive to how the data is split. A description of the algorithm appears in Table 1.

Constructing Covariates and Second-Order Semiparametric Efficiency

My first advance over the double machine learning strategy of Chernozhukov et al. (2018) is constructing a set of covariates that will further refine the estimates of the nuisance functions. Doing so gives more assurance that the method will, in fact, adjust for the true nuisance functions f, g.

In order to do so, consider the approximation

$$\hat{f}(\mathbf{x}_i)\approx f(\mathbf{x}_i) + U_{f,i}^{\mathrm{T}}\gamma_f; \quad \hat{g}(\mathbf{x}_i)\approx g(\mathbf{x}_i) + U_{g,i}^{\mathrm{T}}\gamma_g \tag{12}$$

or, equivalently,

$$\Delta_{\hat{f},i}\approx U_{f,i}^{\mathrm{T}}\gamma_f; \quad \Delta_{\hat{g},i}\approx U_{g,i}^{\mathrm{T}}\gamma_g \tag{13}$$

for some vector of parameters γ_f,γ_g.

These new vectors of control variables $U_{f,i}$,$U_{g,i}$ capture the fluctuations of the estimated functions $\hat{f}$,$\hat{g}$ around the true values, f, g. The expected fluctuation of an estimate around its true value is measured by its standard error (Wooldridge 2013, sec. 2.5), so I construct these control variables from the variance matrix of the estimates themselves. Denoting as $\hat{f}(\mathbf{X})$, $\hat{g}(\mathbf{X})$ the vectors of estimated nuisance component $\hat{f}$, $\hat{g}$, I first construct the variance matrices $\widehat{\mathrm{Var}}\left(\hat{f}(\mathbf{X})\right)$ and $\widehat{\mathrm{Var}}(\hat{g}(\mathbf{X}))$. In order to summarize these matrices, I construct the control vectors $\hat{U}_{\hat{f},i}$,$\hat{U}_{\hat{g},i}$ from principal components of the square root of the variance matrices.[10]

[8] Again, details appear in Appendix A, but valid inference will require that the terms

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\Delta_{\hat{f},i}e_i\right\}\xrightarrow{u}0; \quad \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\Delta_{\hat{f},i}v_i\right\}\xrightarrow{u}0 \tag{11}$$

as well as corresponding terms with $\Delta_{\hat{g},i}$.

[9] See Fong and Tyler (2021) and Ratkovic (2021) for contemporary works in political science exploring a split-sample strategy.

[10] For technical and implementation details, including how this approach integrates with the split-sample strategy, see Appendices A and D–E in the Supplementary Materials.

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

Including these constructed covariates as control variables offers advantages both practical and theoretical. As a practical matter, augmenting the control set $\hat{f}, \hat{g}$ with the constructed control vectors $\hat{U}_{\hat{f},i}$,$\hat{U}_{\hat{g},i}$ helps guard against misspecification or chance error in the estimates $\hat{f}$,$\hat{g}$, adding an extra layer of accuracy to the estimate and making it more likely that the method will properly adjust for f .

As a theoretical matter, the method is an example of a *second-order semiparametrically efficient* estimator. Double machine learning is *first-order semiparametrically efficient* because it only adjusts for the conditional means $\hat{f}, \hat{g}$. By including the estimated controls $\hat{f}, \hat{g}$ but also principal components $U_{\hat{f},i}$, $U_{\hat{g},i}$ constructed from the variance (the second moment, see Wooldridge 2013, app. D.7), the estimates gain an extra order of efficiency and return a *second-order semiparametrically efficient* estimate.[11] The theoretical gain is that second-order efficiency requires only an $n^{1/8}$ order of convergence on the nuisance terms rather than $n^{1/4}$. Although seemingly technical, this simply means that the method allows valid inference on θ while demanding less accuracy from the machine learning method estimating the nuisance terms. At the most intuitive level, including these additional control vectors makes it more likely that the nuisance terms will be captured, with the sample-splitting guarding against overfitting.

IMPROVING ON THE PARTIALLY LINEAR MODEL

Double machine learning addresses a particular issue—namely learning how the background covariates enter the model. Several issues of interest to political scientists remain unaddressed. I turn to these next, which comprise my central contributions.

Adjusting for Treatment Effect Heterogeneity Bias

Aronow and Samii (2016) show that the linear regression estimate of a coefficient on the treatment variable is biased for the marginal effect. The bias emerges through insufficient care in modeling the treatment variable and heterogeneity in the treatment effect, and the authors highlight this bias as a critical difference between a linear regression estimate and a causal estimate.

To see this bias, denote as θ_i the effect of the treatment on the outcome for observation i such that the marginal effect is defined as $\theta = \mathbb{E}(\theta_i)$. To simplify matters, presume the true functions $f\,g$ are known, allowing a regression to isolate as-if random fluctuations of e_i, v_i. Incorporating the heterogeneity in θ_i into the partially linear model gives

$$y_i = t_i\theta_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma + e_i \tag{14}$$

$$= t_i\theta + t_i(\theta_i{-}\theta) + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma + e_i. \tag{15}$$

The unmodeled effect heterogeneity introduces an omitted variable, $t_i(\theta_i{-}\theta)$, which gives a bias of[12]

$$\mathbb{E}\left(\hat{\theta}{-}\theta\right) = \frac{\mathbb{E}\{\mathrm{Cov}(t_i, t_i(\theta_i{-}\theta)|\mathbf{x}_i)\}}{\mathbb{E}\{\mathrm{Var}(t_i|\mathbf{x}_i)\}} \tag{16}$$

$$= \frac{\mathbb{E}\left(v_i^2(\theta_i{-}\theta)\right)}{\mathbb{E}\left(v_i^2\right)}, \tag{17}$$

which I will refer to as *treatment effect heterogeneity bias.*

Inspection reveals that either one of two conditions are sufficient to guarantee that the treatment heterogeneity variance bias is zero. The first occurs when there is no treatment effect heterogeneity ($\theta_i = \theta$ for all observations), and the second when there is no treatment assignment heteroskedasticity—$\mathbb{E}\left(v_i^2\right)$ is constant across observations. As observational studies rarely justify either assumption (see Samii 2016, for a more complete discussion), researchers are left with a gap between the marginal effect θ and the parameter estimated by the partially linear model.

I will address this form of bias through modeling the random component in the treatment assignment. As with modeling the conditional means through unspecified nuisance functions, I will introduce an additional function that will capture heteroskedasticity in the treatment variable.

Interference and Group-Level Effects

The proposed method also adjusts for group-level effects and interference. For the first, researchers commonly encounter data with some known grouping, say at the state, province, or country level. To accommodate these studies, I incorporate random effect estimation into the model. The proposed method also adjusts for interference, where observations may be affected by observations that are similar in some respects ("homophily") or different in some respects ("heterophily"). For example, observations that are geographically proximal may behave similarly (Ripley 1988; Ward and O'Loughlin 2002), actors may be connected via a social network (Aronow and Samii 2017; Sobel 2006), or social actors may react to ideologues on the other end of the political divide (Hall and Thompson 2018). In each setting, some part of an observation's outcome may be attributable to the characteristics of other observations.

Existing approaches require a priori knowledge over what variables drive the interference as well as how the

[11] For more on second- and higher-order efficiency, see van der Vaart (2014), Li et al. (2011), Robins et al. (2008), and Dalalyan, Golubev, and Tsybakov (2006, esp. eq. [4]). Although this theoretical literature is developed, I am the first to incorporate these ideas into software and allow their use in an applied setting.

[12] See Wooldridge (2013, eq. [5.4]).

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

interference affects both the treatment variable and the outcome. Instead, I use a machine learning method to learn the type of interference in the data: what variables are driving interference and in what manner.

The problem involves two components: a measure of proximity and an interferent. The proximity measure addresses which variables are driving how close two observations are.[13] In the spatial setting, for example, these may be latitude and longitude. Alternatively, observations closer in age may behave similarly (homophily) or observations with different education levels may behave similarly (heterophily). The strength of the interference is governed by a bandwidth parameter, which characterizes the radius of influence of proximal observations on a given observation. For example, with a larger bandwidth, interference may be measurable between people within a ten-year age range, but for a narrower bandwidth, it may only be discernible within a three-year range. The interferent is the variable that affects other observations. For example, the treatment level of a given observation may be driven in part by the income level (the interferent) of other observations with a similar age (the proximity measure).

The method learns and adjusts for two types of interference: that driven entirely by covariates and the effect of one observations' treatment on other observations' outcomes. For example, if the interference among observations is driven entirely by exogenous covariates, such as age or geography, the method allows for valid inference on θ. Similarly, if there are spillovers such that one observation's treatment affects another's outcome, as with, say, vaccination, the method can adjust for this form of spillover (e.g., Hudgens and Halloran 2008).[14]

The proposed method does not adjust for what Manski (1993) terms *endogenous interference*, which occurs when an observation's outcome is driven by the behavior of some group that includes itself. This form of interference places the outcome variable on both the left-hand and right-hand sides of the model, inducing a simultaneity bias (see Wooldridge 2013, chap. 16). Similarly, the method cannot adjust for the simultaneity bias in the treatment variable. The third form of interference not accounted for is when an observation's treatment is affected by its own or others' outcomes, a form of posttreatment bias (Acharya, Blackwell, and Sen 2016).

RELATION TO CAUSAL EFFECT ESTIMATION

I have developed the method so far as a tool for descriptive inference, estimating a slope term on a treatment variable of interest. If the data and design allow, the researcher may be interested in a causal interpretation of her estimate.

Generating a valid causal effect estimate of the marginal effect requires two steps beyond the descriptive analysis. First, the estimate must be consistent for a parameter constructed from an average of observation-level causal effects. Correcting for the treatment effect heterogeneity bias described in the previous section accomplishes this. Doing so allows for estimation of causal effects regardless of whether the treatment variable is binary or continuous. Second, the data must meet conditions that allow identification of the causal estimate. I discuss the assumptions here, with a formal presentation in Appendix B of the Supplementary Materials.

First, a *stable value assumption* requires a single version of each level of the treatment.[15] Most existing studies include a noninterference assumption in this assumption, which I am able to avoid due to the modeling of interference described in the previous section. Second, a *positivity assumption* requires that the treatment assignment be nondeterministic for every observation. These first two assumptions are standard. The first is a matter of design and conceptual clarity, whereas the accompanying software implements a diagnostic for the second; see Appendix C of the Supplementary Materials.

The third assumption, the *ignorability assumption*, assumes that the observed observation-level covariates are sufficient to break confounding (Sekhon 2009, 495) such that treatment assignment can be considered as-if random for observations with the same observation-level covariate profile. Implicit in this assumption is the absence of interference. The proposed method relaxes this assumption, allowing for valid inference in the presence of interference.

Even after adjusting for interference, simultaneity bias can still invalidate the ignorability assumption, as discussed in the previous section. This bias occurs when an observation's treatment is affected by other observations' treatment level or when there is a direct effect of any outcome on the treatment. Although the proposed method's associated software implements a diagnostic to assess the sensitivity of the estimates to these assumptions (see Appendix C of the Supplementary Materials), their plausibility must be established through substantive knowledge by the researcher.

These assumptions clarify the nature of the estimand. By assuming the covariates adjust for indirect effects that may be coming from other observations, the proposed method estimates an average direct effect of the treatment on the outcome. Because the proposed method adjusts for other observations' treatments at their observed level, it recovers an average controlled direct effect. The estimated causal effect is then the average effect of a one-unit move of a treatment on the

[13] Manski (1993; 2013) refers to this as the *reference group.*

[14] In this situation, the proposed method estimates what is termed the "direct effect" of the treatment, as the method adjusts for indirect effects that come from other observations.

[15] The issue is one of conceptual clarity and must be handled by the researcher. For example, taking as the treatment variable "attends college" ignores both the quality of the schools and multiple versions of the control condition, i.e., the many paths one may take in not attending college. For more, see Imbens and Rubin (2015, secs. 1.2, 1.6.2).

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

outcome, given all observations' covariates and fixing their treatments at the realized value.

THE PROPOSED MODEL

The proposed model expands the partially linear model to include exogenous interference, heteroskedasticity in the treatment assignment mechanism, and random effects. I refer to it as the partially linear causal effect (PLCE) model as, under the causal assumptions given above, it returns a causal estimate of the treatment on the outcome.

The treatment and outcome models for the proposed method are

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y\left(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}, \mathbf{h}_y\right) + a_{j[i]} + e_i, \qquad (18)$$

and

$$t_i = g_1(\mathbf{x}_i) + g_2(\mathbf{x}_i, \mathbf{X}_{-i})\tilde{v}_i + \phi_t(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}, \mathbf{h}_t) + b_{j[i]} + v_i, \qquad (19)$$

with the following conditions on the error terms:

$$a_j \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_a^2\right); \quad b_j \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_b^2\right) \qquad (20)$$

$$\mathbb{E}(e_i|\mathbf{x}_i, \mathbf{X}_{-i}, t_i, \mathbf{t}_{-i}) = \mathbb{E}(v_i|\mathbf{x}_i, \mathbf{X}_{-i}) = \mathbb{E}(\tilde{v}_i|\mathbf{x}_i, \mathbf{X}_{-i}) = 0, \qquad (21)$$

$$\mathbb{E}(e_i v_i|\mathbf{x}_i, \mathbf{X}_{-i}) = \mathbb{E}(e_i \tilde{v}_i|\mathbf{x}_i, \mathbf{X}_{-i}) = \mathbb{E}(v_i \tilde{v}_i|\mathbf{x}_i, \mathbf{X}_{-i}) = 0, \qquad (22)$$

and

$$\mathbb{E}\left(e_i^4|\mathbf{x}_i, \mathbf{X}_{-i}, t_i, \mathbf{t}_{-i}\right) > 0; \quad \mathbb{E}\left(v_i^4|\mathbf{x}_i, \mathbf{X}_{-i}\right) > 0. \qquad (23)$$

Moving through the components of the model, θ is the parameter of interest. The first set of nuisance functions $(f(\mathbf{x}_i), g_1(\mathbf{x}_i))$ are inherited from the partially linear model. The pure error terms e_i, v_i also follow directly from the partially linear model.

The interference components are denoted $\varphi_y\left(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}, \mathbf{h}_y\right), \varphi_t(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{h}_t)$. The vector of bandwidth parameters are denoted $\mathbf{h}_t, \mathbf{h}_y$, which will govern the radius for which one observation affects others. Note that either the treatment or the covariates from one observation can affect another's outcome, but the only interference allowed in the treatment model comes from the background covariates, as described in the previous section.

The treatment variable has two error components. The term v_i is "pure noise" in that its variance is not a function of covariates. The term $\tilde{v}_i$ is noise associated with heteroskedasticity in the treatment variable. The component $g_2(\mathbf{x}_i, \mathbf{X}_{-i})\tilde{v}_i$ will adjust for treatment effect heterogeneity bias. The term $\tilde{v}_i$ is the error component in the treatment associated with the function $g_2(\mathbf{x}_i, \mathbf{X}_{-i})$, which drives any systematic heteroskedasticity in the treatment variable.

The conditions on the error terms are also standard. The terms $a_{j[i]}, b_{j[i]}$ are random effects with observation i in group $j[i]$ (Gelman and Hill 2007), and Condition 20 assumes the random effects are realizations from a common normal distribution. Equation 21 assumes no omitted variables that may bias inference on θ. This conditional independence assumption is standard in the semiparametric literature (see, e.g., Chernozhukov et al. 2018; Donald and Newey 1994; Robinson 1988). Equation 22 ensures that the error terms are all uncorrelated. Any correlation between e_i and either v_i or $\tilde{v}_i$ would induce simultaneity bias. The absence of correlation between v_i and $\tilde{v}_i$ fully isolates the heteroskedasticity in the treatment variable in order to eliminate treatment effect heterogeneity bias. The final assumptions in Expression 23 require that there be a random component in the outcome variable and the treatment for each observation but that they not vary too wildly as to preclude inference on θ. The right-side element of the last display implies the positivity assumption from the previous section.

Equation 18 and 19 can be combined into the infeasible reduced-form equation

$$y_i = \theta t_i + \Big[f(\mathbf{x}_i), \phi_y\left(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}, \mathbf{h}_y\right), g_1(\mathbf{x}_i), g_2(\mathbf{x}_i, \mathbf{X}_{-i})\tilde{v}_i, \phi_t(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{h}_t)\Big]^{\mathrm{T}} \gamma_{PLCE} + c_{j[i]} + e_i. \qquad (24)$$

where the random effect combines those from the treatment and outcome model, $c_{j[i]} = a_{j[i]} + b_{j[i]}$. The next section adapts the repeated cross-fitting strategy to the proposed model in order to construct a semiparametrically efficient estimate of θ.

Formal Assumptions

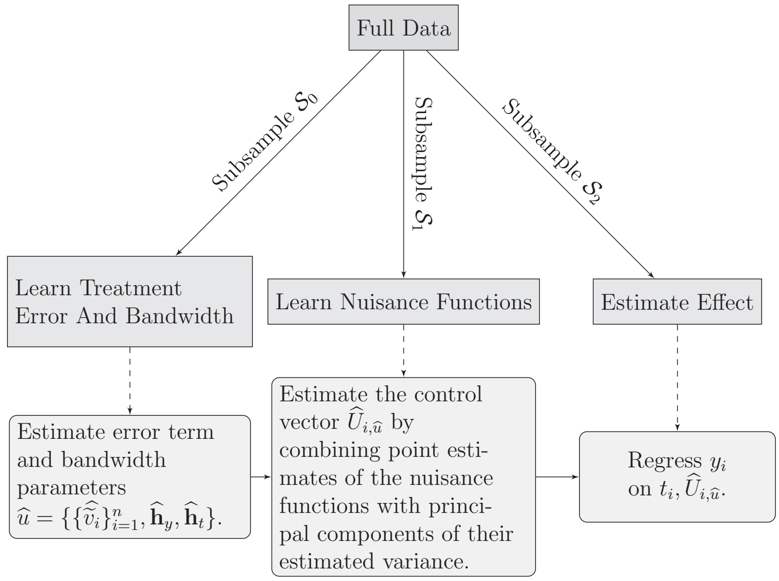
The following assumption will allow a semiparametrically efficient estimate of the marginal effect in the PLCE model.

Assumption 1 (PLCE Assumptions)

1. *Population Model. The population model is given in Equations 18 and 19, and all random components satisfy the conditions in Equations 20–23.*
2. *Efficient Infeasible Estimate. Were all nuisance functions known, the least squares estimate from the reduced form model in Equation 24 would be efficient and allow for valid inference on θ.*
3. *Representation. There exists a finite dimensional control vector $U_{u,i}$ that allows for valid and efficient inference on θ.*
4. *Approximation Error. All nuisance components are estimated such that the approximation errors converge uniformly at the rate $n^{1/8}$.*
5. *Estimation Strategy. The split-sample strategy of Figure 1 is employed.*

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

FIGURE 1. The Estimation Strategy


Full Data
Subsample $\mathbb{S}_0$
Subsample $\mathcal{S}_1$
Subsample $\mathcal{S}_2$
Learn Treatment Error And Bandwidth
Learn Nuisance Functions
Estimate Effect
Estimate error term and bandwidth parameters $\widehat{u} = \{\{\widehat{\widehat{v_i}}\}_{i=1}^n, \widehat{\mathbf{h}}_y, \widehat{\mathbf{h}}_t\}$.
Estimate the control vector $\widehat{U}_{i,\widehat{u}}$ by combining point estimates of the nuisance functions with principal components of their estimated variance.
Regress y_i on $t_i, \widehat{U}_{i,\widehat{u}}$.

Note: To estimate the components in the model, the data are split into thirds. Data in the first subsample, $\mathbb{S}_0$, are used to estimate the interference bandwidths and treatment residuals. The second subsample, $\mathcal{S}_1$, given the estimates from the first, is used to construct the covariates that will adjust for all the biases in the model. The third, $\mathcal{S}_2$, estimates a linear regression with the constructed covariates as controls.

The first assumption requires that the structure of the model and conditions on the error terms are correct. The second assumption serves two purposes. First, it requires that the standard least squares assumptions (see, e.g., Wooldridge 2013, Assumptions MLR 1–5 in chap. 3) hold for the infeasible, reduced form model. This requires no unobserved confounders or unmodeled interference.[16] Second, it establishes the semiparametric efficiency bound, which is the limiting distribution of the infeasible estimate $\hat{\theta}$ from this model.

The third assumption structures the control vector, $U_{u,i}$. This vector contains all estimates of each of the nuisance functions in Equation 24, producing a first-order semiparametrically efficient estimate. This assumption guarantees that, by including the second-order covariates discussed earlier, least squares can be still be used to estimate θ.[17]

The fourth and fifth assumptions are analogous to those implemented in the double machine learning strategy (Chernozhukov et al. 2018). Including the constructed covariates relaxes the accuracy required of the approximation error from $n^{1/4}$ to $n^{1/8}$, and the importance of the repeated cross-fitting strategy in eliminating biases between approximation errors and the error terms e_i, v_i motivates a cross-fitting strategy.

Scope Conditions and Discussion of Assumptions

The assumption that $U_{i,u}$ is finite dimensional is the primary constraint on the model. Effectively, this assumption allows all nuisance functions to condense into a single control vector, allowing valid inference with a linear regression in subsample $\mathcal{S}_2$. This assumption compares favorably to many in the literature. Belloni, Chernozhukov, and Hansen (2014) make a "sparsity assumption," that the conditional mean can be well approximated by a subset of functions of the covariates.[18] I relax this assumption, as the estimated principal components may be an average of a large number of covariates and functions of covariates.

The use of principal components is a form of "sufficient dimension reduction" (Hsing and Ren 2009; Li 2018), where I assume that the covariates and nonlinear functions of the covariates can be reduced to a set that fully captures any systematic variance in the outcome.[19] I am able to sidestep the analytic issues in characterizing the covariance function of the observations analytically (see, e.g., Wahba 1990) by instead

[16] Crucial to the split-sample strategy is that the observations are conditionally independent, meaning a valid marginal effect estimate can be recovered on any randomly generated split. This requires that this aspect is not broken by unmodeled interference. Intuitively, all of the interference is condensed into the functions ϕ_y, ϕ_t such that, after conditioning on these, observations are independent.

[17] With added assumptions, the dimensionality of these covariates could grow on the order of $\sqrt{n}$, though I save this for further work (see, e.g., Cattaneo, Jansson, and Newey 2018; Chernozhukov et al. 2018).

[18] The authors rely on an "approximate sparsity" assumption where the model is sparse up to an error tending toward zero in sample size.

[19] See Appendix D of the Supplementary Materials for a discussion of how the software implements nonlinearities and interactions.

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

taking principal components of the variance matrix. The sample-splitting strategy is also original.

Restricting $U_{i,u}$ to be finite dimensional means that the proposed method cannot accommodate data where the dimensionality of the variance grows in sample size. To give two examples, in the panel setting, the method can handle random effects for each unit but not arbitrary nonparametric functions per unit. Second, the proposed method can account for interference but only if the dimensionality of the interference does not grow in the sample size. This assumption is in line with those made by other works on interference (Savje, Aronow, and Hudgens 2021).

In not requiring distributional assumptions on the treatment variable, the proposed method pushes past a causal inference literature that is most developed with a binary treatment. Many of the problems I address have been resolved in the binary treatment setting (Robins, Rotnitzky, and Zhao 1994; van der Laan and Rose 2011) or where the treatment density is assumed (Fong, Hazlett, and Imai 2018). Nonparametric estimates of inverse density weights are inherently unstable, so I do not pursue this approach but see Kennedy et al. (2017). Rather, the proposed method mean-adjusts for confounding by constructing a set of control variables. I show below, through simulation and empirical examples, that the method generates reliable estimates.

Estimation Strategy: Three-Fold Split Sample

Heuristically, two sets of nuisance components enter the model. The first are used to construct nuisance functions: the treatment error $\tilde{v}_i$ that interacts with g_2 and the bandwidth parameters $\mathbf{h}_y, \mathbf{h}_t$ that parameterize the interference terms. I denote these parameters as the set $u = \left\{\{\tilde{v}_i\}_{i=1}^n, \mathbf{h}_y, \mathbf{h}_t\right\}$. The second set are those that, given the first set, enter additively into the model. These consist of the functions f, g_1 but also the functions g_2,ϕ_y,ϕ_t. If u were known, estimating these terms would collapse into the double machine learning of Chernozhukov et al. (2018). Because u is not known, it must also be estimated in a separate step, necessitating a third split of the data.

I defer precise implementation details to Appendices D and E in the Supplementary Materials, but more important than particular implementation choices is the general strategy for estimating the nuisance components such that the approximation errors do not bias inference on θ. I outline this strategy here.

The proposed method begins by splitting the data into three subsamples, $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{S}_2$, each containing a third of the data. Then, in subsample $\mathcal{S}_0$, all of the components in the models in Equations 18 and 19 are estimated, but only those marked below are retained:

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y(\mathbf{x}_i,\mathbf{X}_{-i},\mathbf{t}_{-i}, \underbrace{\mathbf{h}_y}_{\mathcal{S}_0}) + a_{j[i]} + e_i, \tag{25}$$

$$t_i = g_1(\mathbf{x}_i) + g_2(\mathbf{x}_i, \mathbf{X}_{-i}) \underbrace{\tilde{v}_i}_{\mathcal{S}_0} + \phi_t(\mathbf{x}_i,\mathbf{X}_{-i},\mathbf{t}_{-i}, \underbrace{\mathbf{h}_t}_{\mathcal{S}_0}) + b_{j[i]} + v_i. \tag{26}$$

These retained components, $\hat{\mathbf{h}}_y, \hat{\mathbf{h}}_t$ and a model for estimating the error terms $\left\{\widehat{\tilde{v}}_i\right\}_{i=1}^n$, are then carried to subsample $\mathcal{S}_1$.

Data in subsample $\mathcal{S}_1$ are used to evaluate the bandwidth parameters and error term using the values from the previous subsample and, given these, to estimate the terms marked below:

$$y_i = \theta t_i + \underbrace{f}_{\mathcal{S}_1}(\mathbf{x}_i) + \underbrace{\phi_y}_{\mathcal{S}_1}\left(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}, \hat{\mathbf{h}}_y\right) + \underbrace{a_{j[i]}}_{\mathcal{S}_1} + e_i. \tag{27}$$

$$t_i = \underbrace{g_1}_{\mathcal{S}_1}(\mathbf{x}_i) + \underbrace{g_2}_{\mathcal{S}_1}(\mathbf{x}_i, \mathbf{X}_{-i})\widehat{\tilde{v}}_i + \phi_t\left(\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}, \hat{\mathbf{h}}_t\right) + \underbrace{b_{j[i]}}_{\mathcal{S}_1} + v_i. \tag{28}$$

Having estimated all nuisance terms, including the random effects, the feasible control variable $\hat{U}_{\hat{u},i}$ is now constructed. This variable consists of two sets of covariates. The first is the point estimates of all of the nuisance components estimated from $\mathcal{S}_0$ and $\mathcal{S}_1$ but evaluated on $\mathcal{S}_2$. It also includes the second-order terms, also estimated on subsample $\mathcal{S}_1$ but evaluated on subsample $\mathcal{S}_2$. This control vector is then entered into the reduced form model

$$y_i = \theta t_i + \hat{U}_{\hat{u},i}^{\mathrm{T}}\gamma + e_i, \tag{29}$$

which generates an estimate $\hat{\theta}$ and its standard error.

Estimation is done via a cross-fitting strategy, where the roles of each subsample in generating the estimate are swapped, this cross-fitting is repeated multiple times, and the results aggregated. Complete details appear in Appendix E of the Supplementary Materials.

I now turn to illustrate the performance of the proposed method in two simulation studies.

ILLUSTRATIVE SIMULATIONS

The simulations assess performance across three dimensions: treatment effect heterogeneity bias, random effects, and interference, generating eight different simulation settings. In each, a standard normal covariate x_{i1} is drawn along with error terms v_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 all covariates 0.5, but only the first is used to generate the treatment and the outcome.

The simulations were designed to highlight my theoretical expectations in the simplest possible setting. In each setting, the marginal effect is in-truth 1, the systematic component is driven entirely by the first covariate, and all covariates, random effects, and the error terms are normally distributed. Table 2

TABLE 2. Simulation Specifications

Model Specifications		
Baseline:	$y_i = t_i + x_{i1}^2 + \epsilon_i$	$t_i = x_{i1} + v_i; v_i \overset{i.i.d.}{\sim} N(0, 1)$
Treatment Effect Heterogeneity:	$y_i = t_i \times x_{i1}^2 + \epsilon_i$	$t_i = x_{i1} + v_i; v_i \overset{i.i.d.}{\sim} N\left(0, \frac{x_{i1+1}^2}{2}\right)$
Random Effects:	$y_i = \cdot + a_{j[i]}$	$t_i = \cdot + a_{j[i]}; a_j \overset{i.i.d.}{\sim} N(0, 1); \#j = 50$
Interference:	$y_i = \cdot + \psi_{t,i}$	$t_i = \cdot + \psi_{x,i}$

Constructing Interference Terms

$$\text{Interference Covariates}: \psi_{t,i} = \sum\nolimits_{i' \neq i} p_{i,i'} \times t_{i'};\ \psi_{x,i} = \sum\nolimits_{i' \neq i} p_{i,i'} \times x_{i'1}^2,$$

$$\text{where } p_{i,i'} = \frac{e^{-\left(x_{i1} - x_{i'1}\right)^2}}{\sum_{i' \neq i} e^{-\left(x_{i1} - x_{i'1}\right)^2}}.$$

Note: The first simulation begins with the baseline additive model, and the second adds treatment effect heterogeneity bias by introducing a correlation between effect heterogeneity and treatment assignment heteroskedasticity. In the third, a fifty-leveled random effect is included as a confounder. The final specification adds an interference term, with the precise construction of the term at the bottom. The residual term ε_i follows a standard normal.

provides details. The first model is additive, noninteractive, and equivariant in all errors, serving as a baseline. The second model induces treatment effect heterogeneity bias by including an interaction between the treatment and squared covariate along with heteroskedasticity in the treatment residual. The third adds a fifty-leveled, standard normally distributed random effect as a confounder, and the fourth adds an interaction term. Note that summations are over all other observations such that the outcome is a function of other observations' treatment level and the treatment is a function of other observations' squared covariate.

The covariates are then transformed as

$$\mathbf{x}_i^* = \left[x_{i1} - \frac{1}{2}x_{i2},\ x_{i2} - \frac{1}{2}x_{i1},\ x_{i3},\ x_{i4},\ x_{i5}\right],$$

and each method is given the outcome, treatment, transformed covariates, and indicator variables for the random effects regardless of whether the random effects are in the true data-generating process. I report results for $n = 1{,}000$, with additional sample sizes in Appendix F of the Supplementary Materials.[20]

Along with the proposed method (PLCE), I implement four different machine learning methods. Kernel regularized least squares (KRLS; Hainmueller and Hazlett 2013) fits a single, nonparametric regression, takes the partial derivative of the fitted model with respect to the treatment variable, and returns the average of these values as the marginal effect. The covariate balancing propensity score for continuous treatments (CBPS; Fong, Hazlett, and Imai 2018)[21] generates a set of weights that eliminate the effect of confounders under the assumption that the treatment distribution is normal and equivariant. I also include the double machine learning (DML) algorithm of Chernozhukov et al. (2018) with random forests used to learn $\hat{f}, \hat{g}$ and the generalized random forest (GRF) of Athey, Tibshirani, and Wager (2019), which is similar to DML but uses a particular random forest algorithm tuned for efficient inference on a marginal effect. Ordinary least squares (OLS) is included for comparison.

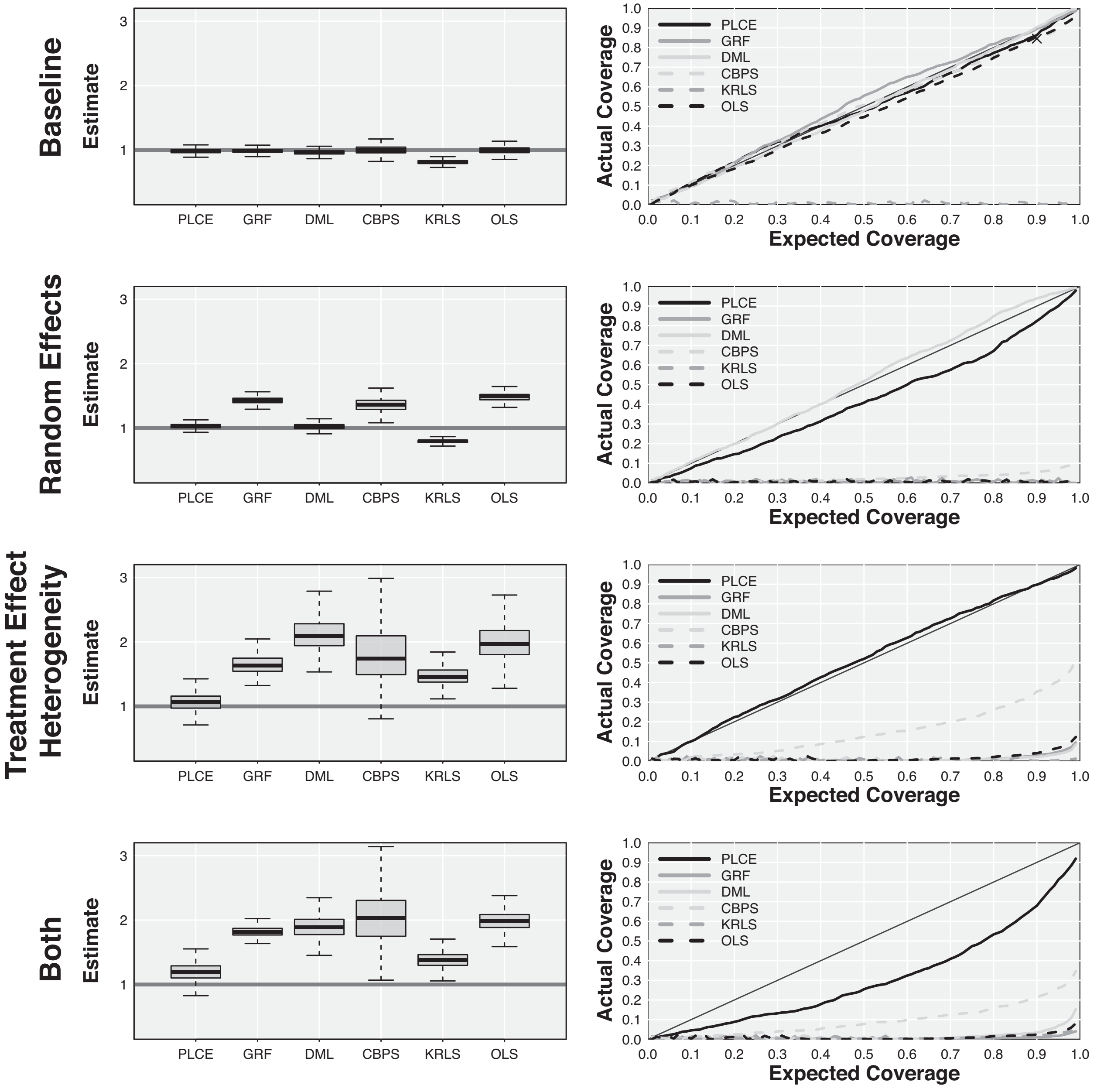
The KRLS approach is closest to the proposed method in that both implement a nonparametric regression model. Thus, KRLS should handle nonlinearities well, but because it does not engage in a split-sample strategy, I expect undercoverage with its confidence intervals. The DML and GRF methods do engage in a split-sample strategy, but, like KRLS, they were not designed to handle random effects. I expect all three to perform poorly. Ordinary least squares should handle the random effects well, as they are simply entered as covariates in the model, but this should be particularly susceptible to treatment effect heterogeneity bias. None of the methods were constructed to adjust for interference. The proposed

[20] At smaller sample sizes, the method performs similarly in terms of point estimation, and $n = 250$ the confidence intervals are valid but a bit conservative, while for $n = 500$ and above, the results appear similar to the results in the body.

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

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

FIGURE 2. Results for Simulations without Interference



Note: The first column shows the distribution of point estimates, where the true value is 1, in gray. The second column shows the coverage rates: expected coverage is on the *x*-axis and actual coverage is on the *y*-axis. If a curve falls below the 45° line, the confidence intervals are too narrow and thus invalid. If the curve falls above the 45° line, the confidence intervals are valid but wide. The proposed method (PLCE) is compared with GRF, DML, CBPS, KRLS, and OLS. The proposed method, PLCE, is the only one to perform well across all settings.

method, PLCE, should do a reasonable job across all settings, as it was designed to handle random effects and adjust for both interference and treatment effect heterogeneity bias.

Results for the Setting without Interference

The results for the simulations without interference are in Figure 2. The first column shows the distribution of point estimates, with the true value of 1 in gray. The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis.[22] For example, consider in the top right plot the point marked "×" at (0.90,0.85), which is on the CBPS curve. Here, I constructed a 90% confidence interval of

[22] The "coverage rate" is the proportion of samples for which the constructed confidence interval contains the true value of 1 (see, e.g., Wooldridge 2013, sec. 4.3.).

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

the form $\left[\hat{\theta}-1.64\hat{\sigma}_{\hat{\theta}}, \hat{\theta}+1.64\hat{\sigma}_{\hat{\theta}}\right]$ and measured the proportion of simulations where the confidence interval contains the true value of 1. In this case, for CBPS, this value is 0.85, so the 90% confidence interval is invalid, albeit only slightly too narrow. More generally, if a curve falls below the 45° line, the confidence intervals are too narrow and thus invalid. If the curve falls above the 45° line, the confidence intervals are valid but wide.

The simulation settings increase in complexity going down the rows. The first row of figures contains contains the baseline model, the second, the model with group indicators added, the third, the baseline model with treatment effect heterogeneity, and the fourth, both treatment effect heterogeneity model and random effects.

Starting in the first row, every method performs well in the baseline model, though KRLS exhibits undercoverage. In the second row, with random effects added, only the proposed method and OLS provide unbiased estimates and valid inference. In the third row, the proposed method and KRLS are unbiased with valid intervals. In the final row, with both random effects and treatment effect heterogeneity, every method shows discernible bias, but the proposed method and KRLS have the lowest bias and the fewest misleading confidence errors.

Several machine learning methods fail to provide unbiased estimation in the presence of random effects or a simple interaction between the treatment effect and treatment residual. Across all settings, the proposed method is the only one that that allows for valid inference.

Results for the Setting with Interference

Figure 3 presents results from the simulations in the presence of interference. All methods save least squares return accurate point estimates in the simplest setting, with the proposed method, DML, and CBPS providing narrow but reasonable confidence intervals. Coverage from GRF, although valid in the setting without interference, is now near zero. In the remaining rows, the point estimates are reasonable, particularly for the proposed method but also KRLS. The effect of interference shows up in the coverage rates. In the bottom three settings, coverage rates are near zero for all methods. Only the proposed method provides both reliable point estimates and confidence intervals across each of the settings.

EMPIRICAL APPLICATIONS

I illustrate the proposed method using data from two recent studies. First, I reanalyze experimental data to illustrate that the proposed method returns estimates and standard errors similar to those from a linear regression when the linear regression is the correct thing to do. Second, I show how the method can estimate a treatment effect with a continuous treatment variable. I use data from a study where the researcher was forced to dichotomize a continuous treatment in order to estimate a causal effect.

Maintaining Efficiency

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.

I focus on how the estimated causal effect of conciliation varies between hawks and doves, as reported in Table 2 of the original work. The results appear in Table 3. For the two estimated effects, the proposed method returns results quite similar to those obtained with least squares. Importantly, the standard errors are comparable across the methods, suggesting no efficiency loss when employing the proposed method in a situation where least squares is known to be unbiased and efficient.

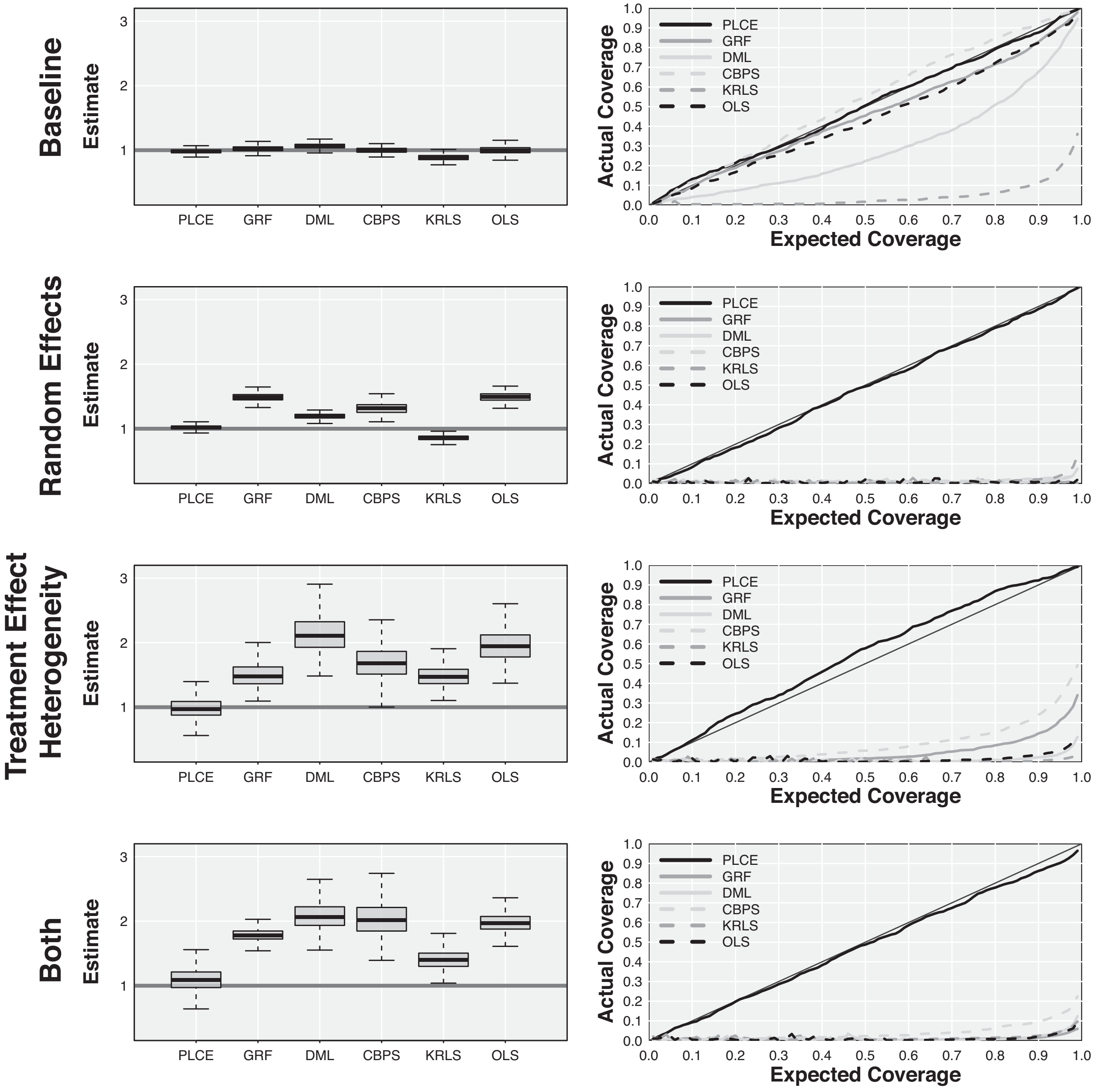
Estimating a Causal Effect in the Presence of a Continuous Treatment

I next reanalyze data from a recent study that estimated the causal effect of racial threat on voter turnout (Enos 2016). The author operationalizes racial threat by distance to a public housing project, a continuous measure, and measures its effect on 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 analysis that, unfortunately, requires a binary treatment. To accommodate the method, 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. However, 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. The proposed method allows estimation of the average causal effect of distance on the outcome.

I conduct four separate analyses. For the first, I estimate the causal effect of distance on change in turnout for white residents within one kilometer of a demolished housing project. The treatment variable is distance to the housing project, and the control variables consist of turnout in the previous two elections (1996, 1998), 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. I also include a random effect for

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

FIGURE 3. Results for Simulations with Interference



Note: The first column shows the distribution of point estimates, where the true value is 1, in gray. The second column shows the coverage rates: expected coverage is on the *x*-axis and actual coverage is on the *y*-axis. If a curve falls below the 45°, the confidence intervals are too narrow and thus invalid. If the curve falls above the 45° line, the confidence intervals are valid but wide. The proposed method (PLCE) is compared with GRF, DML, CBPS, KRLS, and OLS. The proposed method, PLCE, is the only one to perform well across all settings.

identifying the housing project nearest to each individual.[23] I next generate three matched samples for further analysis.[24] The first contains Black voters within one kilometer of a demolished housing project. As argued in the original piece (11), this group will not face racial threat, so it provides a measure of any trend in turnout absent racial threat. The next two samples consist of white and Black voters, but both are further than one kilometer from any housing project, either demolished or not. The latter two groups serve as

[23] See the supplementary materials of Enos (2016) for more details.

[24] I estimate distance as a function of all covariates for white residents within one kilometer of a demolished project using a random forest. I 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.

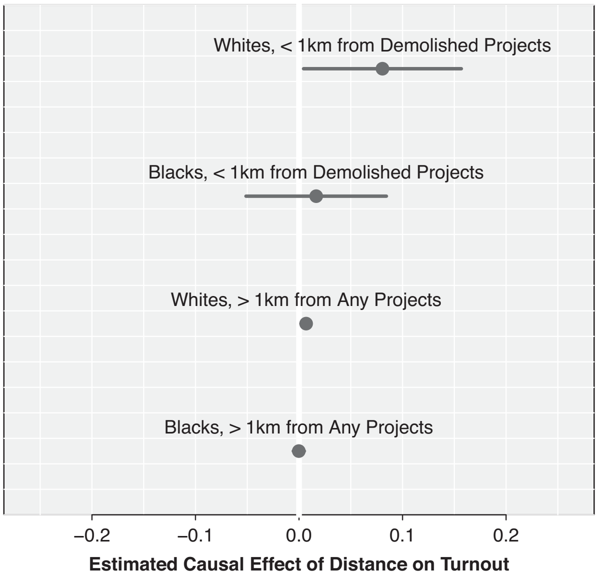
https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

TABLE 3. Comparing PLCE and a Linear Regression in an Experimental Setting

	Hawks			Doves		
	PLCE	Diff-in-Mean	OLS	PLCE	Diff-in-Mean	OLS
Hawks	11.83	11.98	11.97	36.03	35.43	35.19
SE	3.56	3.80	3.80	2.71	3.12	2.85

Note: Across all settings, the PLCE estimates perform comparably to least squares on these experimental data. The repeated cross-fitting strategy does not inflate standard errors in this setting.

FIGURE 4. Causal Effect Estimate of Racial Threat



Note: Revisiting the study by Enos (2016), I find a statistically and substantively relevant effect of racial threat on white voters (*top row*) and, as predicted by theory, not for Black voters near the housing projects (second row) or for Black and white voters further than one kilometer from any projects.

placebo groups, as they are sufficiently far from a demolished project that any threat should be muted.

Figure 4 presents the effect estimates. I estimate that living adjacent to a public housing unit, rather than one kilometer away, causes a decrease in turnout of about 8.06 percentage points for white residents ($SE = 0.0389$, $z = 2.07$), an effect in line with the results from the original analysis (see Figure 1 there). The estimated effect for Black voters near housing projects of 0.017 ($SE = 0.035$, $z = 0.48$) is not significant. The bottom two lines consider distal Blacks and whites, providing a placebo test. I find no effect of distance on turnout. Along with not relying on a user-specified control set, the proposed method allows for causal effect estimation with a continuous treatment variable. I find results of a magnitude similar to those from the original study but without needing to transform the data so that it is amenable to a framework that generally relies on a binary treatment.

CONCLUSION

Testing intuitions and hypotheses against the data in a way that does not rely on strong assumptions is essential to a reliable accumulation of knowledge. Doing so builds faith that the results and theory are driven by actual trends in the data and not a particular set of choices made by the researcher. To this end, I have introduced to political science a framework, taken from the field of semiparametric inference, for conducting valid inference while allowing machine learning methods to construct a control vector that can account for a wide range of commonly encountered biases. Essential to this approach is a sample-splitting strategy, where the same data is never used to both construct the control vector and conduct inference. I have extended this literature, allowing for inference that is robust to both heterogeneities in the treatment effect and particular patterns of interference among observations. The method extends causal inference, as well, accommodating continuous treatment variables. The accompanying software allows these analyses to be done in a line or two of code and allows for several diagnostics.

Ultimately, my goal is to allow for more believable, less assumption-driven inference. I move the field in this direction, where machine learning can be incorporated into workaday research as a means of controlling for background covariates, freeing the researcher to develop and test theories with some confidence that the results are not driven by her ability to specify every element of a statistical model.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0003055422001022.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/SWHIKY.

ACKNOWLEDGMENTS

I would like to thank Soichiro Yamauchi and Max Goplerud for work on developing the software; Dustin

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press

Tingley, Max Goplerud, John Londregan, Scott de Marchi, Brandon Stewart, Kevin Munger, Curtis Signorino, Christopher Lucas, Matt Blackwell, Dean Knox, Neal Beck, Cyrus Samii, Matias Cattaneo, Rod Little, Walter Mebane, and Jonathan Katz, for helpful comments; Camille DeJarnett for excellent research assistance; and Stefan Wager for guidance in implementing his software.

CONFLICT OF INTEREST

The author declares no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The author affirms that this research did not involve human subjects.

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): 512–29.

Achen, Christopher. 2002. “Toward a New Political Methodology: Microfoundations and ART.” *Annual Review of Political Science* 5: 423–50.

Achen, Christopher H. 2005. “Let’s Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong.” *Conflict Management and Peace Science* 22 (4): 327–39.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics.” *Journal of Economic Perspectives* 24 (2): 3–30.

Aronow, Peter M. 2016. “A Note on “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It.”” Working Paper. http://arxiv.org/abs/1609.01774.

Aronow, Peter, and Cyrus Samii. 2016. “Does Regression Produce Representative Estimates of Causal Effects?” *American Journal of Political Science* 60 (1): 250–67.

Aronow, Peter, and Cyrus Samii. 2017. “Estimating Average Causal Effects under General Interference.” *Annals of Applied Statistics* 11 (4): 1912–47.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. “Generalized Random Forests.” *Annals of Statistics* 47 (2): 1148–78.

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* 42 (2): 596–627.

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

Bickel, Peter J., Chris A. J. Klaassen, Ya’acov Ritov, and Jon A. Wellner. 1998. *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

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, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. “Double/Debiased Machine Learning for Treatment and Structural Parameters.” *The Econometrics Journal* 21 (1): C1–68.

Dalalyan, A. S., G. K. Golubev, and A. B. Tsybakov. 2006. “Penalized Maximum Likelihood and Semiparametric Second-Order Efficiency.” *The Annals of Statistics* 34 (1): 169–201.

Donald, Stephen G., and Whitney K. Newey. 1994. “Series Estimation of Semilinear Models.” *Journal of Multivariate Analysis* 50 (1): 30–40.

Enos, Ryan D. 2016. “What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior.” *American Journal of Political Science* 60 (1): 123–42.

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): 156–77.

Fong, Christian, and Matthew Tyler. 2021. “Machine Learning Predictions as Regression Covariates.” *Political Analysis* 29 (4): 467–84.

Freedman, David A. 2006. “On The So-Called ‘Huber Sandwich Estimator’ and ‘Robust Standard Errors.’” *The American Statistician* 60 (4): 299–302.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge 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): 143–68.

Hall, Andrew B., and Daniel M. Thompson. 2018. “Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in US Elections.” *American Political Science Review* 112 (3): 509–24.

Hill, Daniel, and Zachary Jones. 2014. “An Empirical Evaluation of Explanations for State Repression.” *American Political Science Review* 108 (3): 661–87.

Hsing, Tailen, and Haobo Ren. 2009. “An RKHS Formulation of the Inverse Regression Dimension-Reduction Problem.” *Annals of Statistics* 37 (2): 726–55.

Hudgens, Michael G., and Elizabeth Halloran. 2008. “Toward Causal Inference with Interference.” *Journal of the American Statistical Association* 103 (482): 832–42.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. “Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies.” *American Political Science Review* 105 (4): 765–89.

Imai, Kosuke, and Marc Ratkovic. 2013. “Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation.” *The Annals of Applied Statistics* 7 (1): 443–70.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Kennedy, Edward, Zongming Ma, Matthew McHugh, and Dylan Small. 2017. “Nonparametric Methods for Doubly Robust Estimation of Continuous Treatment Effects.” *Journal of the Royal Statistical Society, Series B* 79 (4): 1229–45.

King, Gary. 1990. “On Political Methodology.” *Political Analysis* 2:1–29.

King, Gary, and Langche Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14 (2): 131–59.

King, Gary, and Margaret E. Roberts. 2015. “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do about It.” *Political Analysis* 23 (2): 159–79.

King, Gary, Robert Keohane, and Sidney Verba. 1994. *Designing Social Inquiry*. Princeton, NJ: 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 S., and Alexander Sahn. 2021. “Achieving Statistical Significance with Control Variables and without Transparency.” *Political Analysis* 29 (3): 356–69.

Li, Bing. 2018. *Sufficient Dimension Reduction: Methods and Applications with R*. Boca Raton, FL: Chapman and Hall/CRC.

Li, Lingling, Eric Tchetgen Tchetgen, Aad van der Vaart, and James M. Robins. 2011. “Higher Order Inference on a Treatment

Effect under Low Regularity Conditions." *Statistics & Probability Letters* 81 (7): 821–28.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60 (3): 531–42.

Manski, Charles F. 2013. "Identification of Treatment Response with Social Interactions." *The Econometrics Journal* 16 (1): S1–23.

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

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–44.

Newey, Whitney. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator." *Econometric Theory* 10 (2): 233–53.

Ratkovic, Marc. 2021. Subgroup Analysis: Pitfalls, Promise, and Honesty. In *Advances in Experimental Political Science*, eds. James N. Druckman and Donald P. Green, 271–88. Cambridge: Cambridge University Press.

Ratkovic, Marc. 2022. "Replication Data for: Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/SWHIKY.

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

Ripley, Brian D. 1988. *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Robins, James, Lingling Li, Eric Tchetgen, and Aad van der Vaart. 2008. *Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals*. Beachwood, OH: Institute of Mathematical Statistics.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. "Comment: Performance of Double-Robust Estimators When 'Inverse Probability' Weights Are Highly Variable." *Statistical Science* 22 (4): 544–59.

Robinson, Peter. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–54.

Samii, Cyrus. 2016. "Causal Empricism in Quantitative Research." *Journal of Politics* 78 (3): 941–55.

Savje, Fredrik, Peter M. Aronow, and Michael G. Hudgens. 2021. "Average Treatment Effects in the Presence of Unknown Interference." *Annals of Statistics* 49 (2): 673–701.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Sobel, Michael E. 2006. "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference." *Journal of the American Statistical Association* 101 (476): 1398–407.

Stein, Charles. 1956. "Efficient Nonparametric Testing and Estimation." In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, ed. Jerzy Neyman, 187–95. Oakland: University of California Press.

van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning Causal: Inference for Observational and Experimental Data*. New York: Springer.

van der Vaart, Aad. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, Aad. 2014. "Higher Order Tangent Spaces and Influence Functions." *Statistical Science* 29 (4): 679–86.

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

Ward, Michael, and John O'Loughlin. 2002. "Special Issue on Spatial Methods in Political Science." *Political Analysis* 10 (3): 211–16.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–38.

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

https://doi.org/10.1017/S0003055422001022 Published online by Cambridge University Press