Flexible Covariate Adjustments in Randomized Experiments

Christoph Rothe

Abstract

Linear regression adjustments for pre-treatment covariates are widely used in economics to lower the variance of treatment effect estimates when analyzing data from randomized experiments. This method is robust to misspecification, and delivers reliable confidence intervals even in relatively small samples. More flexible covariate adjustments, using nonlinear parametric or fully nonparametric methods, have the potential to improve efficiency. They are rather uncommon in practice, however, because they can introduce bias or require very large samples in order for asymptotic inference to be reliable. This paper shows that with a simple modification of the treatment effect estimator, it is possible to alleviate these issues substantially. For a large class of covariate adjustments, estimation and inference in randomized experiments is possible without sacrificing the robustness properties of linear regressions. Full efficiency can be achieved through nonparametric adjustments under minimal conditions, in particular without imposing high-order smoothness restrictions in settings with many covariates.

JEL Classification: C13, C14, C21

Keywords: *Treatment effects, semiparametric efficiency, randomized experiment, nonparametric estimation.*

This Version: April 1, 2020. This paper replaces an earlier working paper titled "The Value of Knowing the Propensity Score for Estimating Average Treatment Effects", first circulated in April, 2015. I thank Joshua Angrist, Miikka Rokkanen, Stefan Wager and attendants of the 2016 Econometric Society Winter Meeting in San Francisco for helpful comments. Author's contact information: Department of Economics, University of Mannheim, Email: rothe@vwl.uni-mannheim.de, Web: http://www.christophrothe.net.

1. INTRODUCTION

Randomized experiments have become increasingly popular in many areas of applied economics, as they allow for straightforward inference on causal effects. With a binary treatment, for instance, the difference in average outcomes among treated and untreated units constitutes an unbiased estimator of the average treatment effect. It is also easy to form a confidence interval based on this estimator that, while formally justified by asymptotic theory, is known to work well even with rather moderate sample sizes.

The "difference-in-means" estimator is not statistically efficient, however, if one observes pre-treatment covariates in addition to units' outcomes and treatment assignments. Such data are frequently available in economic applications, and often incorporated into a linear regression of outcomes on a treatment indicator as additional control variables (possibly interacted with the treatment indicator). Standard regression theory implies that proceeding like this improves the efficiency of the treatment effect estimator without introducing bias, irrespective of whether the linear model is correctly specified or not.[1] Moreover, it is straightforward to form confidence intervals through conventional robust standard errors.

While linear regression adjustments improve upon a "difference-in-means" analysis, they are generally not optimal, in the sense that the resulting estimator does not have the smallest asymptotic variance that a regular estimator could have in such a setup. More precise estimates can in principle be obtained by postulating more flexible nonlinear parametric specifications of the relationship between covariates and outcomes; and full statistical efficiency can also be achieved by adjusting for covariates nonparametrically, using methods like series regression or local linear smoothing.

These more involves types of covariate adjustments are not widely used in practice, however, since they lack the robustness properties of the linear regression approach. Nonlinear parametric adjustments can lead to biased treatment effect estimates under misspecification (e.g. Gail et al., 1984). Nonparametric adjustments can be sensitive to the choice of tuning parameters in finite samples, lead to estimators of the average treatment effect that are typically biased in finite samples, and are only guaranteed to lead to full efficiency under delicate regularity conditions that are generally considered unrealistic in settings with several

[1] Throughout the paper, the parameter of interest is the population-level average treatment effect, and sampling from the population is the source of uncertainty about its value. This framework is commonly used in economics. An alternative framework, used frequently in the statistics literature, considers the sample average treatment effect of the observed units as the parameter of interest, and random assignment of the treatment as the only source of uncertainty. Linear regression adjustments can be biased in the latter framework, and potentially increase the variance of the treatment effect estimator. See Freedman (2008) and Lin (2013) for further discussion.

covariates (e.g. Robins and Ritov, 1997).

In this paper, we show that such concerns can be substantially alleviated by choosing a treatment effect estimator that takes the form of a sample average of an empirical analogue of the underlying efficient influence function (EIF). This *EIF estimator* is very similar to doubly robust estimators of the average treatment effect in a model with unconfounded assignment (e.g. Robins et al., 1995; Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005; Farrell, 2015). The main difference is that it replaces the estimated propensity scores that appear in doubly robust estimators with their true values, which are determined by, and thus known to, the analyst in a randomized experiment.

We show that due to this simple modification the EIF estimator satisfies a robustness property with respect to the covariate-adjustment scheme that is substantially stronger than the Neyman orthogonality (or local robustness) condition considered, for example, in Belloni et al. (2017), Chernozhukov et al. (2018a) and Chernozhukov et al. (2018b). Due to this special robustness property, the EIF estimator is $\sqrt{n}$-consistent, essentially unbiased (in a sense made precise below), and asymptotically normal for a very large class of covariate adjustments that includes misspecified parametric and arbitrarily slowly converging nonparametric ones.[2] The EIF estimator is also fully efficient if the covariate-adjustment procedure chosen by the researcher consistently estimates the conditional expectation of outcomes given treatment status and covariates. Valid inference can be conducted, however, without the analyst having to take a stand on whether the covariate-adjustment procedure is consistent through a simple simple estimator of the EIF estimator's asymptotic variance. EIF estimation can thus make use of flexible covariate adjustments without sacrificing any of the robustness properties of linear regression adjustments.

To illustrate how our results for the EIF estimator differ from the ones typically obtained in the literature, consider the case of nonparametric covariate adjustments, for example via local linear or series regression. The EIF estimator then falls into the class of semiparametric two-step (STS) estimators, which are estimators of a finite-dimensional parameter that depend on a nonparametrically estimated nuisance function. Conditions for STS estimators to be $\sqrt{n}$-consistent and asymptotically normal typically include that the nonparametric component is consistent and converges with a rate of smaller order than $n^{-1/4}$ (Newey, 1994; Chen et al., 2003; Belloni et al., 2017; Chernozhukov et al., 2018a). In settings with several covariates, such rates can generally only be achieved under strong smoothness conditions on

[2]We focus on a traditional setup in which the number of covariates is fixed as the sample size increases. See Wager et al. (2016), and Wu and Gagnon-Bartsch (2017) for similar results in high-dimensional settings when machine learning algorithms are used to adjust for covariates.

the function that is being estimated nonparametrically. Such conditions formally justify the use of bias-reducing nonparametric estimators, such as higher-order local polynomial regression. However, in practice the properties of the resulting STS estimators are often poorly approximated by the corresponding asymptotic theory (e.g. Robins and Ritov, 1997).

Due to its special robustness property, the EIF estimator can achieve $\sqrt{n}$-consistency and asymptotically normality under much weaker conditions the nonparametrically estimated nuisance function. Specifically, it only requires that the stochastic part, but not the bias, of the nonparametric component converges to zero, and this rate can be arbitrary small. Allowing for a slowly converging and possibly inconsistent nonparametric component is possible in our setup because the structure of the efficient influence function removes the influence of first stage bias on the final estimator. Randomized assignment thus effectively ensures that efficient estimation can be carried out without any higher-order differentiability conditions.

The remainder of the paper is structured as follows. In the following section, we introduce the model and our proposed estimation procedure. In Section 3, we derive its theoretical properties under general conditions, and propose a simple method for inference. Sections 4 and 5 specifically consider parametric and nonparametric covariate adjustments, respectively. Section 6 presents some simulation results, and Section 7 contains an empirical illustration. Finally, Section 8 concludes. All proofs are collected in the Appendix.

2. SETUP

This section describes the model, reviews some results on efficiency bounds for treatment effect estimation, and introduces the proposed estimation procedure.

2.1. **Model.** In our model, the researcher observes an i.i.d. random sample $\{(Y_i, T_i, X_i)\}_{i=1}^n$ of n units from a large population. Here $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the outcome variable, $T_i \in \{0, 1\}$ is a treatment indicator, with $T_i = 1$ if unit i is treated and $T_i = 0$ otherwise, and $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of pre-treatment covariates. Each unit also has potential outcomes $Y_i(1)$ and $Y_i(0)$ with and without receiving the treatment, respectively, so that $Y_i = Y_i(T_i)$. The parameter of interest is the population average treatment effect

$$\tau = \mathbb{E}(Y_i(1) - Y_i(0)).$$

We consider a class of randomized experiments in which unit i is assigned to the treatment group independent of its potential outcomes with probability $\pi(X_i)$, where $\pi(x) = P(T_i = 1|X_i = x) \in (\epsilon, 1 - \epsilon)$ for some constant $\epsilon > 0$ is the propensity score function chosen by the

analyst.[3] With this structure, we have that

$$(Y_i(1), Y_i(0)) \perp T_i | X_i, \tag{2.1}$$

In practice, the choice of the propensity score function is typically guided by concerns about statistical efficiency (outcomes could be less variable among units with certain covariate values), data collection costs (the cost of sampling units could be related to their covariate values), and practical feasibility (implementing an experiment could be easier if the propensity score function is "simple", in the sense that it only varies with a few of the covariates, and only takes on a small number of distinct values). However, its exact form does not affect the analysis in this paper: the results in the following sections all hold irrespective of whether the propensity score function is constant, a simple step function of a only a few covariates, or a highly complex function of all covariates in the model.

2.2. **Efficiency Bounds.** The setup described above is formally equivalent to a treatment effect model with unconfounded assignment and a known propensity score. Such models have been studied extensively, and we can use existing results to determine the efficiency bound for estimating τ. Let $\mu(t,x) = \mathbb{E}(Y_i | T_i = t, X_i = x)$ and $\sigma^2(t,x) = \text{Var}(Y_i | T_i = t, X_i = x)$ be the conditional expectation and the conditional variance function, respectively, of Y_i given $T_i = t$ and $X_i = x$. Hahn (1998) shows that under Assumption (2.1) the asymptotic variance of any regular estimator of τ is bounded from below by

$$V_{\text{eff}} = \mathbb{E}\left(\frac{\sigma^2(1, X_i)}{\pi(X_i)} + \frac{\sigma^2(0, X_i)}{1 - \pi(X_i)} + (\mu(1, X_i) - \mu(0, X_i) - \tau)^2 \right),$$

and that any regular estimator of τ whose asymptotic variance achieves this efficiency bound is equal to $n^{-1} \sum_{i=1}^{n} \psi_i(\mu) + o_P(n^{-1/2})$, where

$$\psi_i(\mu) = \mu(1, X_i) - \mu(0, X_i) + \frac{T_i(Y_i - \mu(1, X_i))}{\pi(X_i)} - \frac{(1 - T_i)(Y_i - \mu(0, X_i))}{1 - \pi(X_i)}$$

is the efficient influence function (EIF) for estimating τ. Our notation $\psi_i(\mu)$ emphasizes the fact that the only component of this quantity that is not observed by the analyst is the conditional expectation function μ. Note that $V_{\text{eff}} = \text{Var}(\psi_i(\mu))$ by construction.

[3] Assuming that the propensity score is bounded away from zero and one ensures the existence of a regular estimator of τ (Khan and Tamer, 2010). Since the propensity score is chosen by the analyst, this condition is easy to fulfill.

2.3. **Some Existing Estimators.** A simple and widely used estimator of τ is the "difference-in-means" estimator $\hat{\tau}_{\text{unadj}}$, which is simply the unadjusted difference in weighted average outcomes among treated and untreated units, with unit i being weighted with its inverse treatment probability $1/(\pi(X_i)T_i+(1-\pi(X_i))(1-T_i))$. This estimator can of course also be obtained by running an weighted least squares regression of the outcome on the treatment indicator and an intercept:

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}_{\text{unadj}} T_i.$$

While the "difference-in-means" estimator is unbiased and consistent under general conditions, it is generally not efficient since it does not exploit the information contained in the covariates. Its variance is easily improved, however, by augmenting the regression with a linear term in the covariates, yielding the "linear regression adjusted" estimator $\hat{\tau}_{\text{reg}}$:

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}_{\text{reg}} T_i + \hat{\beta}' X_i.$$

Further improvements can be obtained by also including interactions between the covariates and the treatment indicator, yielding the estimator $\hat{\tau}_{\text{reg-int}}$:

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}_{\text{reg-int}} T_i + \hat{\beta}' X_i + \hat{\gamma}' X_i T_i.$$

It follows from standard regression theory that $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{reg-int}}$ are both unbiased and consistent for τ irrespective of whether the respective linear regression model constitutes a correct specification of the conditional expectation function $\mu(t,x)$ or not.[4] While the asymptotic variance of $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{reg-int}}$ generally does not achieve the efficient bound, robustness against misspecification makes these procedures attractive for empirical applications.

A more general class of estimators which incorporates the covariate data proceeds by postulating some specification of the conditional expectation function μ, and then generating some appropriate estimate $\hat{\mu}$. This could be a parametric specification of μ that is estimated by methods like (weighted) least squares or maximum likelihood, or a nonparametric specification estimated by methods like local linear or series regression. Irrespective of which procedure is used by the researcher, the resulting "covariate-adjusted" estimator of

[4]For illustration, consider the special case that the propensity score function is constant. Then T_i and X_i are stochastically independent, and it follows from standard "omitted variable bias" calculations that the expectation of $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{reg-int}}$ is the same as that of $\hat{\tau}_{\text{unadj}}$, and the latter is clearly an unbiased estimator of τ. We recall from footnote 1 that the general setup of papers that find a bias for linear regression adjustments is different from ours.

τ is given by

$$\hat{\tau}_{\text{adj}} = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\right).$$

This class of estimators contains $\hat{\tau}_{\text{unadj}}$, $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{reg-int}}$ as special cases when $\hat{\mu}$ is the least squares fit of the specification $\mu(t, x) = \beta_0 + \beta_1 t$, $\mu(t, x) = \beta_0 + \beta_1 t + \beta_2' x$, and $\mu(t, x) = \beta_0 + \beta_1 t + \beta_2' x + \beta_3 xt$, respectively. In general, however, there is no reason to expect $\hat{\tau}_{\text{adj}}$ to be unbiased or even consistent. Indeed, consistency of $\hat{\tau}_{\text{adj}}$ generally requires consistent estimation of the conditional expectation function μ. If the outcome is binary, for example, it would be natural to consider a Logit specification like $\mu(t, x) = \Lambda(\beta_0 + \beta_1 t + \beta_2' x)$, where Λ denotes the cumulative distribution function of the Logistic distribution, and estimate the unknown parameters by Maximum Likelihood. If the Logistic model is misspecified, however, resulting estimate $\Lambda(\widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2' x)$ is generally not consistent for μ, and the resulting treatment effect estimator $\hat{\tau}_{\text{adj}}$ is inconsistent as well.

Issues of misspecification can in principle be overcome by using some form of nonparametric regression to estimate μ, which yields consistent and efficient estimates of τ under appropriate regularity conditions. The problem is that versions of $\hat{\tau}_{\text{adj}}$ based on such estimates can still be severely biased in finite samples, and exhibit stochastic behavior that is not well approximated by conventional first-order asymptotic theory (Imbens et al., 2007), especially if the number of covariates is large. For these reasons, nonparametric covariate adjustments are hardly used in practice.

2.4. **The Efficient Influence Function Estimator.** In this paper, we study the properties of so-called EIF estimators of τ, which take the form of the sample average of an empirical analogue the efficient influence function:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{\mu}),$$

where again $\hat{\mu}$ is a generic estimate of the conditional expectation function μ. Such estimators can be interpreted as "corrected" versions of $\hat{\tau}_{\text{adj}}$, in the sense that

$$\hat{\tau} = \hat{\tau}_{\text{adj}} + \frac{1}{n}\sum_{i=1}^{n} \left(\frac{T_i(Y_i - \hat{\mu}(1, X_i))}{\pi(X_i)} - \frac{(1 - T_i)(Y_i - \hat{\mu}(0, X_i))}{1 - \pi(X_i)}\right).$$

EIF estimators are very similar to doubly robust estimators from the literature on treatment effect estimation under unconfounded assignment (e.g. Robins et al., 1995; Robins and

Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005; Farrell, 2015). The main difference is that for doubly robust estimation the known propensity score values are replaced with appropriate estimates (this is necessary under unconfoundedness to make the estimator feasible, since the propensity score function is unknown in such setups).

EIF estimators are also related to the class of estimators based on a moment condition that satisfies a so-called Neyman orthogonality condition with respect to an unknown nuisance function. Such estimators were studied, for example, by Belloni et al. (2017) and Chernozhukov et al. (2018a). In our context, Neyman orthogonality means that the expected EIF is first-order insensitive to changes in the conditional expectation function μ, in the sense that

$$\partial_m^1 \left[\mathbb{E}(\psi_i(m))\right]_{m=\mu} = 0,$$

where ∂_m^k is a kth order functional derivative operator with respect to m restricted to directions of possible deviations $\hat{\mu}$ from μ. Since our EIF clearly satisfies this condition, one could in principle use results from Belloni et al. (2017) or Chernozhukov et al. (2018a) to derive properties of $\hat{\tau}$.

In our context of a randomized experiment, however, the expected EIF satisfies a much stronger robustness property than Neyman orthogonality. Simple algebra shows that in fact $\mathbb{E}(\psi_i(m)) = \tau$ for *any* non-random function m, and thus not only the first but also *all* higher order functional derivatives vanish:

$$\partial_m^k \left[\mathbb{E}(\psi_i(m))\right]_{m=\mu} = 0 \text{ for all } k \in \mathbb{N}.$$

This stronger robustness property has important consequences for the types of results we are able to derive in this paper. In particular, it means that we can expect $\hat{\tau}$ to be $\sqrt{n}$-consistent for τ and asymptotically normal even if $\hat{\mu}$ converges arbitrarily slowly to some probability limit $\bar{\mu}$ that may even be different from the true function μ. In contrast, consistency of $\hat{\mu}$ for μ and a rate of $o(n^{-1/4})$ would be required if only Neyman orthogonality holds (Chernozhukov et al., 2018a).

With the stronger robustness property, it holds that $\sqrt{n}(\hat{\tau}-\tau) \approx (1/\sqrt{n})\sum_{i=1}^n(\psi_i(\bar{\mu})-\tau)$ in large samples with very high accuracy under rather weak restrictions on the estimator $\hat{\mu}$. This in turn implies that $\hat{\tau}$ is approximately unbiased for τ, $\sqrt{n}$-consistent and asymptotically normal with limiting variance $\text{Var}(\psi_i(\bar{\mu}))$, and fully efficient if $\hat{\mu}$ consistently estimates μ; i.e. if $\bar{\mu} = \mu$. Moreover, it means that the sample variance of the $\psi_i(\hat{\mu})$ is a natural estimate of $\text{Var}(\psi_i(\bar{\mu}))$; and that this estimate can be used to construct confidence intervals for τ that are valid irrespective of whether $\bar{\mu} = \mu$. We formalize this reasoning in the next section.

3. THEORETICAL RESULTS

In this section, we formally study the properties of the estimator $\hat{\tau}$ under general conditions on the properties of the estimator $\hat{\mu}$. We also propose methods to estimate the asymptotic variance, and to conduct inference.

3.1. **Regularity Conditions.** The following notation is helpful for presenting our regularity conditions. For any class of functions $\mathcal{M}$ over $\{0,1\} \times \mathcal{X}$, let $N_2(\epsilon, \mathcal{M})$ be the minimum number of ϵ-brackets with respect to the $L_2(P)$ norm needed to cover $\mathcal{M}$, where for two functions $u, l \in \mathcal{M}$ the set $\{f \in \mathcal{M} : l(t,x) \leq f(t,x) \leq u(t,x) \text{ for all } (t,x)\}$ is called an ϵ-bracket with respect to $L_2(P)$ if $\mathbb{E}((l(T_i, X_i) - u(T_i, X_i))^2) < \epsilon^2$. We also write $a(\eta) \lesssim b(\eta)$ for generic functions a and b if $a(\eta) \leq Cb(\eta)$ for some constant C not depending on η. All limits are taken as $n \to \infty$. We impose two "high level" assumptions about the estimator $\hat{\mu}$.

Assumption 1. *There exists a sequence μ_n of non-random functions, a non-random function $\bar{\mu}$, and sequences $a_n = o(1)$ and $b_n = o(1)$ of constants such that $\|\hat{\mu} - \mu_n\|_\infty = O_P(a_n)$ and $\|\mu_n - \bar{\mu}\|_\infty = O(b_n)$.*

The idea behind this assumption is to choose $\bar{\mu}$ as the probability limit of $\hat{\mu}$, and to define the function μ_n as the sum of $\bar{\mu}$ and the asymptotic bias of the respective estimator $\hat{\mu}$ with respect to $\bar{\mu}$. Put differently, the idea is to choose μ_n such that $\hat{\mu} - \mu_n$ is approximately mean zero. Proceeding like this allows, but does not require, $\hat{\mu}$ to be a consistent estimator of μ since is possible, but not necessary, that $\bar{\mu} = \mu$. With such choices of $\bar{\mu}$ and μ_n, Assumption 1 simply requires that the stochastic part and the bias of $\hat{\mu}$ converge to zero uniformly. It also denotes the corresponding rates by a_n and b_n, respectively. Uniform convergence results for nonparametric regression estimators are widely available in the literature; see e.g. Newey (1997) for series estimators and Masry (1996) for local polynomial regression. For parametric models, such results generally follow if $\hat{\mu}$ is not "too volatile" in the estimated parameter (Andrews, 1992).

Assumption 2. *There exists a sequence $\mathcal{M}_n$ of function classes such that $P(\hat{\mu} - \mu_n \in \mathcal{M}_n) = 1 + o(1)$ and $N_2(\epsilon, \mathcal{M}_n^*) \lesssim \exp(\epsilon^{-\alpha} c_n)$ for $\alpha \in (0,2)$, a sequence of constants $c_n = o(a_n^{\alpha-2})$, and all $\epsilon < a_n$, where $\mathcal{M}_n^* = \mathcal{M}_n \cap \{m \in \mathcal{M}_n : \|m - \mu_n\|_\infty \leq a_n\}$.*

This assumption states that the estimator $\hat{\mu}$ takes values in a function class whose entropy with bracketing, which is defined as the natural logarithm of the covering number, does not grow too quickly as the sample size increases. Entropy restrictions of this type are commonly found in the literature on semiparametric two-step estimation. They ensure that

the estimator $\hat{\mu}$ does not overfit the data by requiring that it takes values in a class whose elements cannot be "too complex" (see e.g. van der Vaart (1998) for further details on the interpretation of restrictions on covering numbers). The presence of the sequence c_n allows for function classes whose complexity increases with the sample size. For most nonparametric estimators, it is natural to take $\mathcal{M}_n$ as a smoothness class, such as that of functions with bounded partial derivatives up to a particular order. Alternatively, one could also take $\mathcal{M}_n$ as the sum of one potentially non-smooth function and other functions from a smoothness class. This allows dealing with settings where the bias of $\hat{\mu}$ is not a smooth function itself. For parametric models, the assumption again intuitively requires that $\hat{\mu}$ is not "too volatile" in the estimated parameter, in some appropriate sense.

We discuss explicit "low level" conditions under which Assumption 1–2 are satisfied for conventional parametric and nonparametric methods in Sections 4 and 5 below, respectively. For the moment, we would like to stress two important points. First, our two assumptions only restrict the rate a_n at which the stochastic component of $\hat{\mu}$ tends to zero, but not the rate b_n of the bias component. In our setup the estimator $\hat{\mu}$ can therefore not only be inconsistent for μ; it is also allowed to have an arbitrarily slowly vanishing asymptotic bias. Second, while our analysis is geared towards classical parametric and nonparametric estimation procedures for μ, our setup allows for all kinds of estimators $\hat{\mu}$. In particular, our assumptions in principle allow for estimators of μ based data dependent model specifications, or data dependent choices of tuning parameters, including those of the type used in modern machine learning methods. In order to satisfy our Assumptions 1–2, such data dependencies cannot be arbitrary, however, but have to be subject to sufficient "discipline". See Benkeser and Van Der Laan (2016) and Van Der Laan and Bibau (2017), for example, for conditions under which entropy and uniform convergence results hold for LASSO-type estimators.

3.2. **Treatment Effect Estimation.** Our main result regarding the properties of the EIF estimator is the following bound on the difference between $\sqrt{n}(\hat{\tau} - \tau)$ and its asymptotically linear representation, which is obtained using techniques from empirical process theory.

Theorem 1. *Suppose that Assumptions 1–2 hold. Then*

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\bar{\mu}) - \tau) + O_P(c_n^{1/2} a_n^{1-\alpha/2}) + O_P(b_n). \tag{3.1}$$

Moreover, the term of order $O_P(b_n)$ in the previous equation has mean zero.

The result in Theorem 1 is general in the sense that it is valid for *any* estimator $\hat{\mu}$,

including hypothetical or infeasible ones, as long as they satisfy the assumed high-level regularity conditions. If a specific estimator of μ is considered, the result of this theorem can potentially be improved by exploiting special features of the respective procedure. We illustrate this point in the context of local linear regression in Section 5. A direct implication of Theorem 1 is the following result.

Corollary 1. *Suppose that Assumptions 1–2 hold. Then (i)* $\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, Var(\psi_i(\bar{\mu})))$*; and (ii)* $Var(\psi_i(\bar{\mu})) = V_{\text{eff}}$ *if* $\bar{\mu} = \mu$*.*

The corollary shows that under our regularity conditions $\hat{\tau}$ is $\sqrt{n}$-consistent for τ, asymptotically normal, and asymptotically unbiased, irrespective of whether $\hat{\mu}$ consistently estimates μ. Moreover, it shows that the asymptotic variance of $\hat{\tau}$ reaches the efficiency bound V_{eff} if $\hat{\mu}$ consistently estimates μ, irrespective of the exact estimation procedure that is used to construct $\hat{\mu}$.

This result differs from other asymptotic normality results for estimators of a finite-dimensional parameter that depend on flexible, possibly nonparametric estimates of a nuisance function, as it does not require any restrictions on the rate b_n at which the bias of $\hat{\mu}$ tends to zero. While the details depend on the exact specification, generally speaking this means that we can satisfy Assumptions 1–2 by choosing an estimator $\hat{\mu}$ that sufficiently "over-smooths", or "over-regularizes", the data. This is because for all commonly used estimation procedures increased regularization speeds up the rate of convergence of the stochastic part and decreases the "complexity" of the estimated function.[5] On the other hand, $\sqrt{n}$-consistency and asymptotic normality of a generic estimator that depends on a flexibly estimated nuisance function often requires that the stochastic part *and* the bias of the estimated nuisance function are of smaller order than $n^{-1/4}$. As pointed out for example by Linton (1995) or Robins and Ritov (1997), asymptotic approximations to the distribution of some estimator that rely on the assumption of a very accurately estimated nuisance function can be fragile in practice. The fact that $\hat{\tau}$ does not require such conditions makes it attractive for empirical work.

3.3. **Variance Estimation and Confidence Intervals.** For empirical practice, it is important to obtain a consistent estimate of the asymptotic variance of $\hat{\tau}$, since this can be transformed into valid standard errors and confidence intervals for the parameter of interest.

[5]To see this, consider the case of classical nonparametric regression models. With local polynomial regression, for example, increasing the bandwidth decreases the variance and leads to a more regular estimate. Similarly, the variance of a series estimator decreases if a smaller number of series terms is chosen, and the complexity of the estimated function is being reduced.

Since $\text{Var}(\psi_i(\bar{\mu})) = \mathbb{E}((\psi_i(\bar{\mu}) - \tau)^2)$, the natural estimate of the asymptotic variance of $\hat{\tau}$ is

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}(\psi_i(\hat{\mu}) - \hat{\tau})^2,$$

irrespective of how $\hat{\mu}$ is obtained. This variance estimate is straightforward to compute, as it does not require estimating any additional nuisance parameters. It can be shown to be consistent under the conditions of Theorem 1. Together with the asymptotic normality result in Corollary 1, this then implies the validity of standard large sample methods for inference. In particular, it means that a confidence interval for τ with asymptotic coverage $1 - \alpha$ is given by

$$C_{1-\alpha} = \left(\hat{\tau} \pm q_{1-\alpha} \times \sqrt{\hat{V}/n}\right),$$

where $q_{1-\alpha} = \Phi^{-1}(1 - \alpha/2)$ with $\Phi^{-1}(\cdot)$ the standard normal quantile function is the critical value. The following corollary formally states these results.

Corollary 2. *Suppose that Assumptions 1–2 hold. Then (i)* $\hat{V} = Var(\psi_i(\bar{\mu})) + o_P(1)$*; (ii)* $P(\tau \in C_{1-\alpha}) = 1 - \alpha + o(1)$*; and (iii)* $P(\bar{\tau} \in C_{1-\alpha}) = o(1)$ *for all* $\bar{\tau} \neq \tau$*.*

3.4. **Leave-One-Out Estimation.** An interesting variant of the estimator $\hat{\tau}$ is one where for $t = 0, 1$ the estimate of $\mu(t, X_i)$ is computed without using the ith data point. Denoting the corresponding estimator by $\hat{\mu}_{(-i)}(t, X_i)$, the resulting "leave-one-out" EIF estimator is

$$\hat{\tau}_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\hat{\mu}_{(-i)}).$$

Leave-one-out estimation is well-known to reduce the risk of over-fitting the data, and to reduce bias, in various contexts (e.g. Powell et al., 1989). To show their usefulness for the present setup, assume that

$$\mathbb{E}(\hat{\mu}_{(-i)}(t, x)|Y_i, T_i, X_i) = \mu_n(t, x) \tag{3.2}$$

exists, is finite, and non-random, for $t = 0, 1$ and $i = 1, \ldots, n$. Then, by the law of iterated expectations and the linearity of the efficient influence function $\psi_i(\mu)$ in μ, we have that

$$\mathbb{E}(\hat{\tau}_{\text{loo}}) = \mathbb{E}(\psi_i(\mu_n)) = \tau,$$

which means that $\hat{\tau}_{\text{loo}}$ is exactly unbiased.[6] The following corollary shows that under a weak regularity condition, which basically states that one must be able to interpolate the estimates $\hat{\mu}_{(-i)}(t, X_i)$ with a sufficiently regular function, the conclusions of Theorem 1 hold for "leave-one-out" treatment effect estimation as well.

Corollary 3. *Suppose there exists a function $\hat{\mu}$ that satisfies Assumption 1–2, and is such that $\hat{\mu}(t, X_i) = \hat{\mu}_{(-i)}(t, X_i)$ for $t = 0, 1$ and $i = 1, \ldots, n$. Then the conclusion of Theorem 1 holds with $\hat{\tau}_{\text{loo}}$ replacing $\hat{\tau}$.*

3.5. **The Value of Knowing the Propensity Score.** We end this section by remarking that our results have some implications for our understanding of treatment effect estimation from observational (non-experimental) data. Recall that the framework used in this paper is formally equivalent to a canonical treatment effect setup with unconfounded assignment and a known propensity score (cf. Imbens, 2004). Knowledge of the propensity score does not affect the efficiency bound for estimating the average treatment effect under unconfoundedness, and efficient estimation is in principle possible if such knowledge is simply ignored (Hahn, 1998). On thus sometimes encounters the opinion that knowledge of the propensity score is not particularly valuable in such applications from a statistical point of view.

The results in this paper show, however, that knowledge of the propensity score makes it possible to construct estimators that achieve the efficiency bound under substantially weaker regularity conditions, and have superior finite-sample properties, relative to estimators that ignore such knowledge. These improvements can be interpreted as the value of knowing the propensity score in applications with unconfounded treatment assignment.

4. PARAMETRIC COVARIATE ADJUSTMENTS

In this section, we give primitive conditions under which our Assumptions 1–2 are satisfied when $\hat{\mu}$ is a parametric estimation procedure. Suppose that the researcher uses the specification $\mu(t, x) = m_\theta(t, x)$, where m_θ is a function that is known up $\theta \in \Theta \subset \mathbb{R}^s$ for some $s \in \mathbb{N}$. This specification may be correct or incorrect. That is, there may or may not be some $\theta \in \Theta$ such that $\mu = m_\theta$. Also put $\hat{\mu} = m_{\hat{\theta}}$, where $\hat{\theta} \in \Theta$ is some estimator. We then impose the following regularity condition.

Assumption 3. *(i) The function $m_\theta(t, x)$ is such that $|m_{\theta_1}(t, x) - m_{\theta_2}(t, x)| \leq h(t, x)\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \Theta$ and some function h with $\mathbb{E}(|h(T_i, X_i)|^2) < \infty$; (ii) there exists $\theta^* \in \Theta$ and*

[6] For many estimators the expectation in (3.2) does formally not exist. In this case we can obtain an analogous "conditional unbiasedness" result by assuming that the conditional expectation of $\hat{\mu}_{(-i)}(t, x)$ given all treatment assignments and covariates values exists.

a deterministic sequence θ_n^* *taking values in* Θ *such that* $\|\hat{\theta} - \theta_n^*\| = O_P(a_n)$ *and* $\|\theta_n^* - \theta^*\| = O(b_n)$ *for* $a_n = o(1)$ *and* $b_n = o(1)$.

The first part of this assumption is a continuity condition on the candidate functions for μ with respect to the unknown parameter; and the second part prescribes that $\hat{\theta}$ has a non-random probability limit, and that its stochastic and bias component vanish with rate a_n and b_n, respectively. This structure covers most parametric procedures that are typically used for estimating conditional expectation functions in practice. Examples include Ordinary Least Squares (OLS) estimates of linear regression models, and Maximum Likelihood (ML) estimates of Probit/Logit specifications in the case of a binary outcome variable. Note that the assumption does not require that $\hat{\mu} = m_{\hat{\theta}}$ is a consistent estimator of μ, but only that it converges to a fixed probability limit m_{θ^*}. The assumption also allows for parameter estimators with "irregular" rates of convergence, i.e. ones that differ from the usual rate of $n^{-1/2}$, as it only requires that $\hat{\theta}$ is consistent for some θ^*.

Corollary 4. *Suppose that Assumption 3 holds. Then Assumptions 1–2 are satisfied with* $\mathcal{M}_n = \{m_\theta(t,x) : \theta \in \Theta\}$, $\bar{\mu} = m_{\theta^*}$, $\mu_n = m_{\theta_n^*}$, $c_n = 1$, *and* $\alpha > 0$ *arbitrarily small.*

Through a minor variation of the proof of Theorem 1, one can show that its conclusion also holds under the conditions of Corollary 4 with $\alpha = 0$; see the appendix for details. This implies that the difference between $\sqrt{n}(\hat{\tau} - \tau)$ and its asymptotically linear representation is of the same order as the rate at which $\hat{\theta}$ approaches θ^*. That is, since $\|\hat{\theta} - \theta^*\| = O_P(a_n) + O(b_n)$ under our assumptions, it holds that

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\bar{\mu}) - \tau) + O_P(\|\hat{\theta} - \theta^*\|).$$

In the most widely used parametric models one can typically choose θ_n^* such that $\|\hat{\theta} - \theta_n^*\| = O_P(n^{-1/2})$ and $\|\theta_n^* - \theta^*\| = O(n^{-1})$ under standard conditions, but, as mentioned above, setups with slower converging parametric estimators exist.

5. NONPARAMETRIC COVARIATE ADJUSTMENTS

In this section, we give primitive conditions under which our Assumptions 1–2 are satisfied when $\hat{\mu}$ is obtained by classical nonparametric regression techniques. For concreteness, we focus on the case of local linear regression; but similar results can be obtained for other methods, such as series regression. We also illustrate how direct arguments can be used to derive results that improve upon the general finding of Theorem 1.

5.1. **Notation and Assumptions.** Local linear regression Fan and Gijbels (1996) is a class of kernel-based smoothers. It is well-known to have attractive bias properties relative to other kernel-based methods, such as the Nadaraya-Watson estimator. We use the following notation. Let $\mathcal{K}$ be a univariate probability density function, and put $K_h(b) = \prod_{j=1}^{d} \mathcal{K}(b_j/h)/h$ for any bandwidth $h \in \mathbb{R}_+$. Then the local linear estimator $\hat{\mu}(t,x)$ of $\mu(t,x)$ is given by the first component of

$$(\hat{\mu}(t,x), \hat{\beta}(t,x)) = \underset{(m,b)}{\operatorname{argmin}} \sum_{j=1}^{n} \left(Y_j - m - b'(X_j - x)\right)^2 K_h(X_j - x)\mathbf{1}\{T_j = t\}.$$

Note that we are using the same bandwidth for each component of the covariate vector X_i for notational convenience only, and that more general bandwidth choices are possible. The following assumption collects some regularity conditions that are standard in the literature on local linear regression.

Assumption 4. *(i) X_i is continuously distributed given $T_i = t$ with compact and convex support $\mathbb{X} \subset \mathbb{R}^d$ for $t = 0, 1$; (ii) the corresponding conditional density functions are bounded, have bounded first order derivatives, and are bounded away from zero, uniformly over the respective support; (iii) $\mu(t,\cdot)$ is twice continuously differentiable for $t = 0, 1$; (iv) $\sup_x \mathbb{E}(|Y_i|^{2+\delta}|T_i = t, X = x) < \infty$ for some constant $\delta > 0$ and $t = 0, 1$; (v) the kernel $\mathcal{K}$ is l times continuously differentiable, and such that $\int \mathcal{K}(u)du = 1$, $\int u\mathcal{K}(u)du = 0$, $\int |u^2\mathcal{K}(u)|du < \infty$, and $\mathcal{K}(u) = 0$ for u not contained in some compact set.*

5.2. **Application of General Results.** Let $\mathcal{M}(l, c_n)$ be the collection of all functions m defined over $\{0,1\} \times \mathbb{X}$ such that the partial derivatives of $m(t,\cdot)$ up to order l are uniformly bounded by c_n for $t = 0, 1$. We then have the following result.

Corollary 5. *Suppose that Assumption 4 holds with $l > \max\{1, d/2\}$, and that $h \propto n^{-\theta}$ with*

$$0 < \theta < \frac{3 - d/l}{(3 + 2l - d/l)d}.$$

Then Assumptions 1–2 are satisfied with $\mathcal{M}_n = \mathcal{M}(l, c_n)$, $\alpha = d/l$, $a_n = (nh^d/\log(n))^{-1/2}$, $b_n = h^2$, and $c_n = (nh^{d(1+2l)}/\log(n))^{-1/2}$.

The corollary implies that $\hat{\tau}$ reaches the efficiency bound with local linear covariate adjustments if the bandwidth h does not tend to zero too quickly. Since the bandwidth is allowed to vanish arbitrarily slowly, asymptotic efficiency can always be achieved (in the next subsection, we show that one can actually allow for a wider range of bandwidth values

than suggested by the corollary by exploiting the structure of the local linear regression estimator). Note that Corollary 5 only requires that the kernel function $\mathcal{K}$ has derivatives up to some order that increases with the number of covariates. This condition ensures that $\hat{\mu}$ is sufficiently often differentiable for taking values in $\mathcal{M}(l, c_n)$ with high probability. No higher-order differentiability conditions on μ are needed, as it is not necessary for our results that the bias of $\hat{\mu}$ vanishes quickly. This differs markedly from generic STS estimators, which typically require higher-order differentiability conditions on the respective nuisance function in settings with many covariates in order to justify the use of methods that control the magnitude of the asymptotic bias of the respective nonparametric estimate.

5.3. **Improved Results Using Direct Arguments.** It is possible to improve upon the result of Corollary 5 and Theorem 1 in the present context by using direct arguments that exploit the specific structure of the local linear estimator. Consider the leave-one-out (LOO) version of the treatment effect estimator, which uses an estimate of μ that uses every observation but the ith in order to estimate $\mu(t, X_i)$, for $t = 0, 1$. Specifically, let

$$\hat{\tau}_{\text{loo}} = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\hat{\mu}_{(-i)})$$

where, for $i = 1, \ldots, n$ and $t = 0, 1$, the estimator $\hat{\mu}_{(-i)}(t, X_i)$ is the first component of

$$(\hat{\mu}_{(-i)}(t, X_i), \hat{\beta}_{(-i)}(t, X_i)) = \underset{(m,b)}{\operatorname{argmin}} \sum_{j=1, j\neq i}^{n} \left(Y_j - m - b'(X_j - x)\right)^2 K_h(X_j - x) \mathbf{1}\{T_j = t\}.$$

Using results in Rothe and Firpo (2019), we then obtain the following result regarding the properties of the corresponding treatment effect estimator.

Corollary 6. *Suppose that Assumption 4 holds with $l = 2$. Then*

$$\begin{aligned} \sqrt{n}(\hat{\tau}_{\text{loo}} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\mu) - \tau) + O_P(h^2) + O_P(n^{-1/2} h^{-d/2}) \\ + O_P(\log(n)^{3/2} n^{-1} h^{-3d/2}). \end{aligned} \tag{5.1}$$

Moreover, if $h \propto n^{-\theta}$ with $0 < \theta < 2/(3d)$, then $\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V_{\text{eff}})$.

This result shows that the treatment effect estimator reaches the efficiency bound for a much wider range of bandwidths than those found using our general result in the previous subsection. Using a wider range of bandwidths is possible because the second and third

remainder term on the right-hand-side of (5.1) are substantially smaller than their counterpart under Theorem 1. [7] Using a "leave-one-out" estimator is of particular importance for obtaining this result. With a "leave-in" version of the local linear regression estimator, there would be an additional term of order $O_P(n^{-1/2}h^{-d})$ on the right-hand side of equation (5.1). Note that long as $d < 8$, the estimator $\hat{\tau}_{\text{loo}}$ is $\sqrt{n}$-consistent if we choose $h \propto n^{1/(4+d)}$. Such a choice minimizes the order of the integrated mean squared error of $\hat{\mu}_{(-i)}$, and hence a bandwidth satisfying this property can be estimated via cross-validation.

6. SIMULATIONS

In this section, we study the finite sample properties of the EIF estimator $\hat{\tau}$ through a Monte Carlo experiment, and compare them to those of other treatment effect estimators. Our aim is to illustrate that the theoretical results obtained above provide a realistic picture of the behavior of the EIF estimator in practice. For simplicity, we focus on the case of nonparametric covariate adjustments via local linear regression. We simulate the potential outcomes as $Y(1) = \lambda(X_1, \ldots, X_5) + 2\cos(\pi X_1 X_2) + \varepsilon_1$ and $Y(0) = \lambda(X_5, \ldots, X_1) + \varepsilon_0$. Here $\lambda(a) = \sin(\pi a_1 a_2) + (a_3 + a_4 - 1)^2 + a_5/2$, the covariates $X = (X_1, \ldots X_5)$ are independent random variables following uniform distributions on $[0, 1]$, and the error terms $(\varepsilon_1, \varepsilon_0)$ are independent of each other and the covariates, normally distributed with mean 0 and standard deviation $1/5$. We also simulate the treatment indicator T independently of covariates and potential outcomes as equal to one with probability $1/2$. These choices imply that $\tau \approx .589$ and $V_{\text{eff}} \approx 1.205$ for our data generating process.

We then consider the sample sizes $n = 100, 200, 400, 800$; and compare the performance of the following procedures: (i) a simple difference-in-means estimator, (ii) conventional linear regression adjustments, (iii) "direct" nonparametric covariate adjustments (i.e., the estimator $\hat{\tau}_{\text{adj}}$), and (iv) nonparametric covariate adjustments through the efficient influence function (i.e., the EIF estimator $\hat{\tau}$). For the last two estimators, nonparametric covariate adjustments are carried out via leave-one-out local linear regression.[8] We also consider two

[7] An inspection of the proof of Corollary 6 shows that the first two second-order terms on right-hand side of (5.1) have mean zero, which is in line with the argument in Section 3.4. The orders of magnitude of these two terms are the same as those of the asymptotic bias and the pointwise asymptotic standard deviation of $\hat{\mu}_{(-i)}$, respectively. The final term on the right-hand side of (5.1) is due to presence of the inverse of the estimated (local) second moments of the covariates in the explicit expression of the local linear regression estimator. We conjecture that the reminder rate given in Corollary 6 is actually not sharp. However, in view of Rothe and Firpo (2019), improving the rate would require lots of tedious calculations, and hence we do not investigate this issue any further here.

[8] We also considered the properties of estimators based on conventional "leave-in" local linear regression. This did not affect the empirical bias and variance properties of $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}$ in a meaningful way, but lead to

different methods for choosing the bandwidths, namely leave-one-out cross-validation and "fixed" bandwidths of the form $\text{sd}(X_j)n^{-1/(4+\dim(X))}$ for the jth covariate. Both choices are permissible for the EIF estimator under our theoretical results derived above. We remark that for the relatively small sample sizes we consider, adjusting for five continuous covariates in a completely nonparametric fashion would generally be considered impractical in the existing literature.

Table 1 presents the results of our simulation study based on 100,000 replications for each sample size under consideration. For each estimator, we report its bias scaled by $\sqrt{n}$, its variance scaled by n (so that it can easily be compared to the efficiency bound $V_{\text{eff}} \approx 1.205$), the average value of the respective variance estimator,[9] and the coverage probability of the corresponding confidence interval with nominal level 95%.

Both the difference-in-means estimator and linear regression adjustments perform as expected. These estimators are known to be exactly unbiased in our setup, so the non-zero empirical bias figures reported in Table 1 are due to simulation noise. Linear regression leads to minor reductions in variance relative to difference-in-means for the smaller sample sizes under consideration, and a roughly 10% reduction for the larger ones. In both cases, the resulting scaled variance still exceeds the efficiency bound. Linear regression confidence intervals show a minor deviation from nominal coverage for $n = 100$, but are correct up to simulation noise for the larger sample sizes.

Direct nonparametric covariate adjustments turn out to be substantially biased for both cross-validation and fixed bandwidths. With cross-validation bandwidths, the scaled bias is roughly constant over the various sample sizes under consideration, whereas for fixed bandwidths it increases with the sample size. Sampling variability is generally below that of linear regression adjustments, and improves substantially with the sample size. For $n = 800$, the variance of the estimator with cross-validation only exceeds the efficiency bound by about 6%. However, the corresponding variance estimator tends to be downward biased, which together with the bias issues leads to confidence intervals that undercover the parameter of interest. These problems illustrate why direct nonparametric covariate adjustments are rarely used in the context of randomized experiments.

In contrast, the modified treatment effect estimator based on the efficient influence func-

downward-biased estimates of the corresponding variance in smaller samples, and thus to under-coverage of the corresponding confidence intervals.

[9] We use the HC1 heteroscedasticity-robust variance estimator for linear regression adjustments, and the estimator $\hat{V}$ described above for both the "direct" nonparametric covariate adjustments and the EIF estimator. The latter choice is appropriate since both estimators have the same asymptotically linear representation.

Table 1: Simulation results

			$\sqrt{n}\times$Bias				$n\times$Variance					
n	DIM	LR	DNP-cv	EIF-cv	DNP-fix	EIF-fix	DIM	LR	DNP-cv	EIF-cv	DNP-fix	EIF-fix
100	.002	.015	.442	.022	.519	.007	2.074	2.015	1.783	1.545	2.297	1.764
200	.002	.007	.490	.021	.679	.007	2.066	1.929	1.501	1.377	1.822	1.499
400	.005	.013	.466	.001	.818	.004	2.049	1.882	1.352	1.291	1.548	1.367
800	.001	.006	.455	.001	.992	.003	2.045	1.849	1.279	1.260	1.399	1.297
			Avg Variance Estimate				CI Coverage Probability					
n	DIM	LR	DNP-cv	EIF-cv	DNP-fix	EIF-fix	DIM	LR	DNP-cv	EIF-cv	DNP-fix	EIF-fix
100	2.069	2.001	1.408	1.408	1.664	1.664	.938	.937	.905	.931	.888	.936
200	2.057	1.920	1.325	1.276	1.450	1.450	.945	.945	.919	.942	.889	.943
400	2.052	1.879	1.279	1.247	1.348	1.348	.947	.947	.928	.947	.881	.947
800	2.050	1.859	1.256	1.231	1.296	1.296	.949	.948	.931	.948	.856	.949

Results for difference-in-means estimator (DIM), linear regression adjustments (LR), "direct" nonparametric covariate adjustments with cross-validation and "fixed" bandwidth (DNP-cv and DNP-fix), and nonparametric covariate adjustments through the efficient influence function with cross-validation and "fixed" bandwidth (EIF-cv and EIF-fix); based on 100,000 replications.

Table 2: Summary Statistics

	Full Sample		Treated Only		Control Only	
	Mean	SD	Mean	SD	Mean	SD
Post-Program Earnings	16093	17070	16487	17390	15310	16392
Pre-Program Earnings	3233	4264	3205	4279	3287	4234
Age in Years	33.20	9.64	33.14	9.60	33.28	9.74
Years of Education	11.61	1.87	11.62	1.87	11.58	1.88
Sex (1 =male)	.53	.49	.53	.49	0.53	.49
Assignment (1 =treatment)	.66	.47				
Number of Observations	9,223		6,133		3,090	

tion performs very well with either cross-validation or fixed bandwidths. The reported bias figures are very close to those of the difference-in-means estimator and linear regression adjustments, which we know are exactly unbiased. The variance is well-below that of the other estimators we consider here for all sample sizes, and gets very close to the efficiency bound for $n = 800$. Moreover, with the exception of a small deviation for $n = 100$, the corresponding variance estimator accurately captures the finite-sample variance, and the resulting confidence intervals have excellent coverage.

Overall, the simulation results confirm that the theoretical results in this paper provide a realistic approximation to the EIF estimator's finite-sample properties even in settings with multiple continuously distributed covariates and relatively small sample sizes.

7. EMPIRICAL ILLUSTRATION

We illustrate the implementation of flexible covariate adjustments in randomized experiments by applying them to data from National Job Training Partnership Act (JTPA) Study. A detailed description the study's design and findings is given by Bloom et al. (1997). The study randomly assigned applicants into a treatment and a control group, with the treatment group being eligible to receive a mix of training, job-search assistance, and other services provided by the JTPA for a period of 18 months. The probability of being assigned to the treatment group was equal to 2/3, and independent of any covariates. The study collected background information on the applicants prior to random assignment, as well as data on applicants' earnings in the 30-month period following the assignment.

The sample that we use in this paper contains 9,223 individuals with non-missing data on the pre-assignment covariates sex, age, years of education, and pre-program earnings; and the primary outcome, which is earnings in the 30 months after program assignment. We

Table 3: Estimation results

	DIM	LR	EIF-cv	EIF-fix
Point Estimate	1,176	1,234	1,229	1,244
Standard Error	369	340	339	321

Estimated effect of JTPA eligibility on earnings in the 30 months after program assignment for difference-in-means estimator (DIM), linear regression adjustments (LR), and nonparametric covariate adjustments through the efficient influence function with cross-validation and "fixed" bandwidth (EIF-cv and EIF-fix); based on 9,223 observations.

present some descriptive statistics for these variables in Table 2. These show, among other things, that the covariates are well-balanced between the treatment and the control group. We then consider the effect of program assignment on post-assignment earnings, using the covariates to improve precision. We employ the same estimators as in our simulation study, except for the direct nonparametric adjustment estimators that did not perform well.[10]

Table 3 shows the empirical results for the estimators under consideration. The baseline difference-in-means estimator takes a value of \$1,176, with a standard error of \$369. Including linear regression adjustments produces a slightly larger point estimate of \$1,234, and reduces the standard error by about 8% to \$340. For the EIF estimator based on cross-validation, we obtain a similar point estimate and standard error of \$1,229 and \$339, respectively. the cross-validation algorithm produces bandwidths that are rather large in this setup due to the large variability in outcomes.[11] EIF estimation with a fixed bandwidth then produces a point estimate \$1,244 with standard error of \$321, which constitutes a 4% improvement over linear regression adjustments. While the gains from using a multivariate nonparametric covariate adjustment are thus modest in this setup, the analysis still shows that such adjustments are feasible in randomized experiments: they produce stable estimates that improve upon simple methods.

8. CONCLUSIONS

This paper shows that the scope for flexible covariate adjustments in randomized experiments is much bigger than generally considered in the empirical literature. By estimating average treatment effects through a sample analogue of the efficient influence function, one can

[10] We let age, years of education and pre-program earnings enter the local linear regression in the usual fashion, and smooth with respect to the binary variable "sex" as described in Racine and Li (2004).

[11] With large bandwidths local linear regression is almost identical to a global linear regression model, and hence the numerical results for EIF estimation are similar to those for linear regression adjustments.

improve upon linear regression adjustments in terms of efficiency without having to sacrifice any of their robustness properties. Moreover, fully efficient estimation is possible by using nonparametric covariate adjustments under conditions that are substantially weaker than those generally required in the literature on semiparametric two-step estimation.

A. PROOFS

A.1. **Proof of Theorem 1.** Let $\lambda_n(m) = n^{-1/2}\sum_{i=1}^n(\psi_i(\pi,m)-\mathbb{E}(\psi_i(\pi,m)))$ for any generic function $m(t,x)$ defined over $\{0,1\}\times\mathcal{X}$ such that $\mathbb{E}(\psi_i(\pi,m))$ exists and is finite. Simple algebra shows that $\mathbb{E}(\psi_i(\pi,m)) = \tau$ for *any* such generic function $m(t,x)$, and thus $\lambda_n(m) = n^{-1/2}\sum_{i=1}^n(\psi_i(\pi,m)-\tau)$. The first statement of the theorem follows from an application of the Central Limit Theorem to $\lambda_n(\bar{\mu})$ if $\lambda_n(\hat{\mu}) - \lambda_n(\bar{\mu}) = o_P(1)$. By linearity, we have that $\lambda_n(\hat{\mu}) - \lambda_n(\bar{\mu}) = \lambda_n(\hat{\mu}-\mu_n) + \lambda_n(\mu_n-\bar{\mu})$; and Assumption 2 implies that $\lambda_n(\mu_n-\bar{\mu}) = O_P(b_n) = o_P(1)$. Next, for any fixed $m^*\in\mathcal{M}_n^*$ and any $\epsilon>0$ it holds that

$$\begin{aligned} P(|\lambda_n(m^*)|>\epsilon) &\leq \frac{1}{\epsilon}\sup_{m\in\mathcal{M}_n^*}\mathbb{E}(|\lambda_n(m)|) \\ &\lesssim \frac{1}{\epsilon}\int_0^{a_n}\sqrt{\log(N_2(s,\mathcal{M}_n^*))}ds \\ &= \frac{a_n^{1-\alpha/2}c_n^{1/2}}{\epsilon}, \end{aligned}$$

using Markov's inequality, the maximal inequality in Corollary 19.35 in van der Vaart (1998), and our Assumption 2. Assumption 1 and 2 together also imply that $P(\hat{\mu}-\mu_n\in\mathcal{M}_n^*) = 1+o(1)$, and thus we find that $\lambda_n(\hat{\mu}-\mu_n) = O_P(a_n^{1-\alpha/2}c_n^{1/2}) = o_P(1)$, since $c_n = o(a_n^{\alpha-2})$ by Assumption 3. Taken together, we thus have that

$$\lambda_n(\hat{\mu}) - \lambda_n(\bar{\mu}) = O_P(a_n^{1-\alpha/2}c_n^{1/2}) + O_p(b_n),$$

as claimed. □

A.2. **Proof of Corollary 1.** The first statement follows from the Central Limit Theorem, since $O_P(a_n^{1-\alpha/2}c_n^{1/2}) + O_p(b_n) = o_P(1)$ under our assumptions. The second statement is obvious. □

A.3. **Proof of Corollary 2.** This result follows from standard arguments. □

A.4. **Proof of Corollary 3.** This result follows since $\hat{\mu}$ as defined in the Corollary satisfies the assumptions made for Theorem 1. □

A.5. **Proof of Corollary 4.** It follows from van der Vaart (1998, Example 19.7) that $N_2(\epsilon, \mathcal{M}_n) \lesssim \epsilon^{-\dim(\Theta)} \lesssim \exp(\epsilon^{-a})$ for all $a > 0$, which implies that Assumption 2 is satisfied under the conditions of the Corollary. From the Glivenko-Cantelli Theorem, it also follows that Assumption 1 is satisfied. However, it also follows from a slight modification of the proof of Theorem 1 that $\lambda_n(m_{\hat{\theta}}) - \lambda_n(m_{\theta^*}) = O_P(\|\hat{\theta} - \theta^*\|)$, as claimed in the comment after the Corollary. Specifically, since $N_2(\epsilon, \mathcal{M}_n)$ is actually of "smaller-than-exponential" order here, it follows that

$$P(|\lambda_n(m^*)| > \epsilon) \lesssim \frac{1}{\epsilon} \int_0^{a_n} \sqrt{\log(N_2(s, \mathcal{M}_n^*))} ds = \frac{a_n - a_n \log(a_n)}{\epsilon},$$

for any fixed $m^* \in \mathcal{M}_n^*$ and any $\epsilon > 0$. □

A.6. **Proof of Corollary 5.** For this proof, we use the notation that for $s = (s_1, \ldots, s_d)$ a vector of non-negative integers let $\partial_x^s m(t,x) = \partial_{x_1}^{s_1} \ldots \partial_{x_d}^{s_d} m(t,x)$ denotes the partial derivatives with respect to x of a generic function m. It then follows from Masry (1996) that we can choose a sequence of functions μ_n, equal to the sum of μ and an asymptotic bias term, such that

$$\|\hat{\mu} - \mu_n\|_\infty = O_P\left(\left(\frac{\log(n)}{nh^d}\right)^{1/2}\right) \quad \text{and} \quad \|\mu_n - \mu\|_\infty = O(h^2).$$

Hence Assumption 1 is satisfied. Moreover, van der Vaart (1998, Example 19.9) shows that $N_2(\epsilon, \mathcal{M}_n) \lesssim \exp(\epsilon^{-d/l} c_n)$. Differentiability of the kernel function then implies that $\hat{\mu}$ is continuously differentiable up to order l; and by arguing as in Masry (1996) and Portier and Segers (2018) one can also show that for all s with $|s| \equiv \sum_{j=1}^d s_d \leq l$ we have

$$\|\partial_x^s \hat{\mu} - \partial_x^s \mu_n\|_\infty = O_P\left(\left(\frac{\log(n)}{nh^{d+2|s|}}\right)^{1/2}\right).$$

It then follows from simple algebra that Assumption 2 is satisfied given the restrictions on the bandwidth.

A.7. **Proof of Corollary 6.** This result follows from exactly the same arguments as those used in the proof of Lemma 3 in Rothe and Firpo (2019). □

REFERENCES

Andrews, D. (1992): "Generic uniform convergence," *Econometric theory*, 8, 241–257.

Bang, H. and J. Robins (2005): "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

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

Benkeser, D. and M. Van Der Laan (2016): "The highly adaptive lasso estimator," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 689–696.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos (1997): "The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act study," *Journal of Human Resources*, 549–576.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a): "Double/debiased machine learning for treatment and structural parameters," .

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2018b): "Locally robust semiparametric estimation," *Working Paper*.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Farrell, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189, 1–23.

Freedman, D. A. (2008): "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 40, 180–193.

Gail, M. H., S. Wieand, and S. Piantadosi (1984): "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates," *Biometrika*, 71, 431–444.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Imbens, G. (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G., W. Newey, and G. Ridder (2007): "Mean-square-error calculations for average treatment effects," *Working Paper*.

Khan, S. and E. Tamer (2010): "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, 78, 2021–2042.

Lin, W. (2013): "Agnostic notes on regression adjustments to experimental data: Reexamining Freedmans critique," *Annals of Applied Statistics*, 7, 295–318.

Linton, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, 63, 1079–1112.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571–599.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Portier, F. and J. Segers (2018): "On the weak convergence of the empirical conditional copula under a simplifying assumption," *Journal of Multivariate Analysis*, 166, 160–181.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica*, 57, 1403–1430.

Racine, J. and Q. Li (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119, 99–130.

Robins, J. and Y. Ritov (1997): "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

Robins, J. M. and A. Rotnitzky (2001): "Comment on "Inference for semiparametric models: some questions and an answer" by P. Bickel and J. Kwon," *Statistica Sinica*, 11, 920–936.

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

Rothe, C. and S. Firpo (2019): "Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically," *Econometric Theory*, 35, 1048–1087.

Van Der Laan, M. and A. Bibau (2017): "Uniform Consistency of the Highly Adaptive Lasso Estimator of Infinite Dimensional Parameters," *Working Paper*.

Van der Laan, M. and J. Robins (2003): *Unified methods for censored longitudinal data and causality*, Springer.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016): "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.

Wu, E. and J. Gagnon-Bartsch (2017): "The LOOP Estimator: Adjusting for Covariates in Randomized Experiments," *Working Paper*.