The Japanese Economic Review
https://doi.org/10.1007/s42973-023-00138-1

SPECIAL ISSUE: ARTICLE
Statistical Decision Theory and Treatment Choice

Inference on optimal treatment assignments

Timothy B. Armstrong[1] · Shu Shen[2]

Received: 16 May 2023 / Revised: 10 July 2023 / Accepted: 26 July 2023

Abstract
We consider inference on optimal treatment assignments. Our methods allow inference on the treatment assignment rule that would be optimal given knowledge of the population treatment effect in a general setting. The procedure uses multiple hypothesis testing methods to determine a subset of the population for which assignment to treatment can be determined to be optimal after conditioning on all available information, with a prespecified level of confidence. A Monte Carlo study confirms that the inference procedure has good small sample behavior. We apply the method to study Project STAR and the optimal assignment of a small class intervention based on school and teacher characteristics.

Keywords Optimal treatment assignment · Set inference · Multiple testing

1 Introduction

In recent decades, there has been increasing recognition in both academic and public circles that social experiments or social programs, as costly as they are, should be rigorously evaluated to learn lessons from past experience and to better guide future policy decisions. While recent literature has considered the problem of treatment decision rules given experimental or observational data (see, among others, Manski, 2004; Dehejia, 2005; Hirano and Porter, 2009; Stoye, 2009; Chamberlain, 2011; Tetenov, 2012; Bhattacharya and Dupas, 2012), the problem of constructing confidence statements for the optimal treatment assignment has received little attention. The goal of this paper is to formulate this problem and propose a solution. This allows researchers to quantify how strong the evidence is in favor of treating certain individuals.

✉ Timothy B. Armstrong
timothy.armstrong@usc.edu

Shu Shen
shushen@ucdavis.edu

[1] University of Southern California, Los Angeles, CA, USA

[2] University of California, Davis, CA, USA

Published online: 27 September 2023

To understand the importance of confidence statements for optimal treatment assignments, consider the case where a policy-maker wants to design a social program that gives some selected individuals a treatment intervention (say, reduced class size). The effect of the treatment on the response outcome (say, student test score) is expected to be heterogeneous and varies along certain observed variables (say, teacher experience). A natural goal of the policy maker is to assign treatment only to those with treatment effect expected to be above some prespecified threshold such as zero or the cost of the treatment. The expected treatment effects of different individuals are unknown, but, if data from a previous experimental intervention are available, the policy-maker can make an informed guess about who should be treated, say, by selecting only individuals with values of observed variables linked to an estimated conditional average treatment effect (conditional on individuals' observed characteristics) exceeding the prespecified threshold. The literature on statistical treatment rules has formulated the notion of an "informed guess" and proposed solutions in terms of statistical decision theory. The contribution of this paper is to develop methods that accompany the treatment assignment rule with a confidence statement quantifying the strength of the evidence in favor of providing treatment to certain selected individuals.

We formulate the problem of inference on the optimal treatment assignment as one of reporting a subset of individuals for which treatment can be determined to be optimal conditional on observables while controlling the probability that this set contains any individual for whom treatment should not be recommended conditional on the available information. Our procedures recognize the equivalence of this problem with the problem of multiple hypothesis testing. We propose to select the individuals for whom it can be determined that the population optimal assignment gives treatment by testing multiple hypotheses regarding the conditional average treatment effect for each individual based on the value of the conditioning variable, while controlling the probability of false rejection of any single hypothesis.

The proposed inference procedure for optimal treatment assignment is useful in policy analysis and program evaluation studies. In this paper, we apply the inference method to study the assignment of small class in Project STAR. With a 5% significance level, the method determines that the population optimal treatment assignment rule assigns less experienced teachers in poor schools to teach small classes. The proposed inference method also finds evidence for treatment effect heterogeneity among students with different observed characteristics.

The problem of optimal statistical decision rules for treatment assignment has been considered by Manski (2004), Dehejia (2005), Hirano and Porter (2009), Stoye (2009), Chamberlain (2011), Tetenov (2012), Bhattacharya and Dupas (2012), and others. Additional papers that consider this problem and were circulated after the initial draft of the present paper include Kitagawa and Tetenov (2018), Mbakop and Tabord-Meehan (2021) and Athey and Wager (2021). In this literature, individuals are assigned to different treatments by a social planner who maximizes social welfare or minimizes the risk associated with different statistical decision rules for treatment based on noisy data. As discussed above, our goal is distinct from and complementary to the goal of this literature: we seek to formulate and solve the problem of confidence statements for the (population) optimal treatment assignment,

which can be reported along with a "point estimate" given by the solution to the statistical decision problem formulated and solved in the literature described above. We emphasize that our methods are intended as confidence statements for the treatment assignment that would be optimal given knowledge of the joint distribution of variables in the population, not as a statistical treatment assignment rule that should be implemented given the data at hand (which is the problem formulated by the papers cited above). Rather, we recommend that results based on our methods be reported so that readers can quantify the statistical evidence in favor of treating each individual. We provide further discussion of situations where our confidence region is of interest in Appendix A.

While confidence regions are often interpreted as a measure of statistical precision, they do not, in general, provide a statement about the performance of any given estimator. The same distinction arises in our setting: our confidence regions are for the population optimal treatment assignment rule; they do not provide a statement about the performance of any particular statistical decision rule proposed in the literature described above. The problem of reporting and interpreting guarantees on the performance of statistical decision rules has been considered by Manski and Tetenov (2016) and Manski and Tetenov (2019).

While we are not aware of other papers that consider inference on the treatment assignment rule that would be optimal in the population, Luedtke and Laan (2016) consider inference on expected welfare under the population optimal treatment rule and Bhattacharya and Dupas (2012) derive confidence intervals for the expected welfare associated with certain statistical treatment rules. In contrast, we focus on inference on the population optimal treatment rule itself. These two methods achieve different goals. Our methods for inference on the optimal treatment rule can be used to answer questions about how optimal treatment assignment varies along observed covariates. On the other hand, our methods do not attempt to quantify the increase in welfare from a given treatment rule, which is the goal of estimates and confidence intervals for average welfare.

This paper is closely related to Anderson (2008) and to Lee and Shaikh (2014). Those papers use finite sample randomization tests to construct subsets of a discrete conditioning variable for which treatment can be determined to have some effect on the corresponding subpopulation. Our problem is formulated differently from theirs. Our goal of finding correct inference on optimal treatment assignment rule leads us to report only those values of covariates for which treatment increases the average outcome (rather than, say, increasing the variance or decreasing the average outcome). This, and our desire to allow for continuous covariates, leads us to an asymptotic formulation of the corresponding multiple testing problem. In short, while we both use the idea of multiple hypothesis testing for set construction, our multiple hypotheses are different, leading to different test statistics and critical values.

The method we use to construct confidence statements on optimal treatment decision rules is related to the recent literature on set inference, including Chernozhukov et al. (2007) and Romano and Shaikh (2010). Indeed, the complement of our treatment set can be considered a setwise confidence region in the sense of Chernozhukov et al. (2007), and our solution in terms of multiple hypothesis testing can be considered a confidence region for this set that extends the methods of

Romano and Shaikh (2010) to different test statistics. In addition, our paper uses step-down methods for multiple testing considered by Holm (1979) and Romano and Wolf (2005) and applied to other set inference problems by Romano and Shaikh (2010). In the case of continuous covariates, we use results from the literature on uniform confidence bands (see Neumann and Polzehl, 1998; Claeskens, 2003; Chernozhukov et al., 2013; Kwon, 2022). In particular, we use results from Chernozhukov et al. (2013), who are interested in testing a single null hypothesis involving many values of the covariate. Our testing formulation is different from theirs as our formulation leads us to the multiple hypothesis testing problem of determining which values of the covariates lead to rejection; the step-down method gains precision in our context, but would be irrelevant in Chernozhukov et al. (2013).

The phrase "optimal treatment assignment" is also used in the experimental design literature, where treatment assignments are designed to minimize the asymptotic variance bound or risk of treatment effect estimators (see Hahn et al., 2011). In contrast to this literature, which considers the design phase of the experiment, we take data from the initial experiment as given and focus on implications for future policy.

Our proposed inference procedure on optimal treatment assignments is also related to the test for treatment effect heterogeneity considered by Crump et al. (2008). In fact, it not only tests the null hypothesis that the treatment effect does not vary along an observed variable, but also solves the additional problem of determining which values of the variable cause this null to be rejected. Thus, our paper extends the body of knowledge on treatment effect heterogeneity by providing a procedure to determine for which values of the conditioning variable the conditional average treatment effect differs from the average over the entire population.

Monte Carlo experiments show that our proposed inference procedures have good size and power properties in small samples. The method properly controls the probability of including wrong individuals to the confidence region and successfully selects a large portion of the true treatment beneficiaries. The step-down method in multiple testing improves the power of the inference procedure given a sample size, meaning that it helps to include more individuals into the confidence region while properly controlling its type I error. The size and power properties of the proposed inference procedure are also compared with a "folk wisdom" method based on pointwise confidence bands of the conditional average treatment effect. We show that the latter method often generates nonempty treatment sets in cases where no treatment effect is actually present.

The remainder of the paper is organized as follows: Section 2 formulates the problem of constructing confidence statements for treatment assignment rules. Section 3 links the problem of statistical inference to multiple hypothesis testing and proposes an inference method that derives the treatment assignment rule with statistical precision controlled for. Section 4 conducts several Monte Carlo experiments that study the small sample behavior of the proposed inference method. Section 5 applies the method to Project Star. Section 6 concludes. Appendix A discusses situations where our confidence region is of interest. Appendix B discusses an extension to two-sided confidence regions. Appendix C derives some of the

properties of our confidence region in terms of average welfare when used as a statistical treatment rule.

2 Setup

To describe the problem in more detail, we introduce some notation. For each individual i, there is a potential outcome $Y_i(1)$ with treatment, a potential outcome $Y_i(0)$ with no treatment, and a vector of variables X_i observed before a treatment is assigned. Let $D_i \in \{0, 1\}$ be an indicator for treatment. The goal of a policy-maker is to decide which individuals should be assigned to the treatment group so as to maximize the expectation of some social objective function. We take the social objective function, without loss of generality, to be the realized outcome itself.[1]

Let $t(x) \equiv E(Y_i(1) - Y_i(0)|X_i = x)$ be the conditional average treatment effect. Then the population optimal treatment policy is to treat only those individuals with a covariate $X_i = x$ such that the conditional average treatment effect $t(x)$ is positive. In other words, the treatment rule that would be optimal given knowledge of the distribution of potential outcomes in the population and the covariate X_i of each individual would assign treatment only to individuals with covariate X_i taking values included in the set

$$\mathcal{X}_+ \equiv \{x|t(x) > 0\}.$$

While the ideas in this paper are more general, for the sake of concreteness, we formulate our results in the context of i.i.d. data from an earlier policy intervention with randomized experimental data or observational data in which an unconfoundedness assumption holds. Formally, we observe n observations of data $\{(X_i, D_i, Y_i)\}_{i=1}^n$ where realized outcome $Y_i \equiv Y_i(D_i)$ and $D_i \in \{0, 1\}$ is an indicator for treatment and X_i is a vector of pre-treatment observables. The data are assumed to satisfy the following unconfoundedness assumption.

Assumption 1

$$E(Y_i(j)|D_i = j, X_i = x) = E(Y_i(j)|X_i = x), \qquad j = 0, 1.$$

Assumption 1 is restrictive only if the policy intervention is non-experimental. It is also called the selection on observables assumption as it requires that the observational data behave as if the treatment is randomized conditional on the covariate X_i. Assumption 1 is a standard assumption in the treatment effect literature. Under the assumption, the expected outcomes for both the treatment and the control group in the sample give the same expected outcomes as if both potential outcome variables were observed for all individuals.

If the data we observe is from an initial trial period of the policy intervention with a random sample from the same population, Assumption 1 is enough for us to

[1] This is without loss of generality because costs can be incorporated into the set-up by being subtracted from the treatment.

perform inference on the positive treatment set $\mathcal{X}_+$. However, if the policy maker is deciding on a treatment policy in a new location, or for a population that differs systematically from the original sample in some other way, one must make additional assumptions (see Hotz et al., 2005). In general, one needs to assume that the conditional average treatment effect is the same for whatever new population under consideration for treatment in order to directly apply estimates and confidence regions from the original sample.

We propose to formulate the problem of forming a confidence statement of the true population optimal treatment rule $\mathcal{X}_+$ as one of reporting a treatment set $\hat{\mathcal{X}}_+$ for which we can be reasonably confident that treatment is, on average, beneficial to individuals with any value of the covariate x that is included in the set. Given a pre-specified significance level α, we seek a set $\hat{\mathcal{X}}_+$ that satisfies

$$\liminf_n P\left(\hat{\mathcal{X}}_+ \subseteq \mathcal{X}_+\right) \geq 1 - \alpha, \tag{1}$$

or a treatment group that, with more than probability $(1 - \alpha)$, consists only of individuals who are expected to benefit from the treatment. Therefore, $\hat{\mathcal{X}}_+$ is defined as a set that is contained in the true optimal treatment set $\mathcal{X}_+$, rather than a set containing $\mathcal{X}_+$. This definition of $\hat{\mathcal{X}}_+$ corresponds to the goal of reporting a sub-population for which there is overwhelming evidence that the conditional average treatment effect is positive. As discussed in the introduction, this goal need not be taken as a policy prescription: a researcher may recommend a policy based on a more liberal criterion while reporting a set satisfying (1) as a set of individuals for whom evidence for treatment is particularly strong. We propose methods to derive the set $\hat{\mathcal{X}}_+$ by noticing that a set that satisfies (1) is also the solution to a multiple hypothesis testing problem with an infinite number of null hypotheses $H_x : t(x) \leq 0$ for all $x \in \tilde{\mathcal{X}}$, where $\tilde{\mathcal{X}}$ is the set of values of X_i under consideration. The multiple hypothesis testing problem controls the familywise error rate (FWER), or the probability of rejecting a single x for which H_x is true. With this interpretation, $\hat{\mathcal{X}}_+$ gives a subset of the population for which we can reject the null that the conditional average treatment effect is non-positive given the value of X_i while controlling the probability of assigning to treatment even a single individual for which the conditional average treatment effect (conditional on X_i) is negative. The next section describes in detail the proposed inference method for deriving the set $\hat{\mathcal{X}}_+$. In any case, the role of $Y_i(0)$ and $Y_i(1)$ can be reversed to obtain a confidence region that contains $\mathcal{X}_+$ with $1 - \alpha$ probability. We give a formulation of two-sided confidence sets in Appendix B.

3 Inference procedures

Let $\hat{t}(x)$ be an estimate of the conditional average treatment effect $t(x)$ and $\hat{\sigma}(x)$ an estimate of the standard deviation of $\hat{t}(x)$. Let $\tilde{\mathcal{X}}$ be a subset of the support of the X_i under consideration. For any set $\mathcal{X} \subseteq \tilde{\mathcal{X}}$, let the critical value $\hat{c}_{u,\alpha}(\mathcal{X})$ satisfy

$$\liminf_n P\left(\sup_{x\in\mathcal{X}}\frac{\hat{t}(x)-t(x)}{\hat{\sigma}(x)}\le\hat{c}_{u,\alpha}(\mathcal{X})\right)\ge 1-\alpha. \tag{2}$$

The critical value $\hat{c}_{u,\alpha}(\mathcal{X})$ can be obtained for different estimators $\hat{t}(x)$ using classical central limit theorems (if $\mathcal{X}$ is discrete), or, for continuously distributed X_i, results on uniform confidence intervals for conditional means such as those contained in Neumann and Polzehl (1998), Claeskens (2003), Chernozhukov et al. (2013) or Kwon (2022) as we describe later. For some of the results, we will require that these critical values be non-decreasing in $\mathcal{X}$ in the sense that

$$\mathcal{X}_a\subseteq\mathcal{X}_b\Longrightarrow\hat{c}_{u,\alpha}(\mathcal{X}_a)\le\hat{c}_{u,\alpha}(\mathcal{X}_b). \tag{3}$$

Given the critical value, we can obtain a set $\hat{\mathcal{X}}_+^1$ that satisfies (1). Let

$$\hat{\mathcal{X}}_+^1\equiv\left\{x\in\tilde{\mathcal{X}}\middle|\hat{t}(x)/\hat{\sigma}(x)>\hat{c}_{u,\alpha}(\tilde{\mathcal{X}})\right\}.$$

Clearly $\hat{\mathcal{X}}_+^1$ satisfies (1), since the event in (2) implies the event in (1).

However, we can make an improvement on inference using a step-down procedure (see Holm, 1979; Romano and Wolf, 2005). That is, we can find a set $\hat{\mathcal{X}}_+$ that includes $\hat{\mathcal{X}}_+^1$ but also satisfies (1). The procedure is as follows. Let $\hat{\mathcal{X}}_+^1$ be defined as above. For $k>1$, let $\hat{\mathcal{X}}_+^k$ be given by

$$\hat{\mathcal{X}}_+^k=\left\{x\in\tilde{\mathcal{X}}\middle|\hat{t}(x)/\hat{\sigma}(x)>\hat{c}_{u,\alpha}\left(\tilde{\mathcal{X}}\backslash\hat{\mathcal{X}}_+^{k-1}\right)\right\}.$$

Note that $\hat{\mathcal{X}}_+^{k-1}\subseteq\hat{\mathcal{X}}_+^k$, so the set of rejected hypotheses expands with each step.

Whenever $\hat{\mathcal{X}}_+^k=\hat{\mathcal{X}}_+^{k-1}$, or when the two sets are close enough to some desired level of precision, we stop and take $\hat{\mathcal{X}}_+=\hat{\mathcal{X}}_+^k$ to be our set.

Theorem 1 *Let* (2) *and* (3) *hold. Then* $\hat{\mathcal{X}}_+^k$ *satisfies* (1) *for each k.*

Proof On the event that $\hat{\mathcal{X}}_+\not\subseteq\mathcal{X}_+$, let $\hat{j}$ be the first j for which $\hat{\mathcal{X}}_+^{\hat{j}}\not\subseteq\mathcal{X}_+$. Since $\hat{\mathcal{X}}_+^{\hat{j}-1}\subseteq\mathcal{X}_+$ (where $\hat{\mathcal{X}}_+^0$ is defined to be the empty set), this means that

$$\sup_{x\in\tilde{\mathcal{X}}\backslash\mathcal{X}_+}\frac{\hat{t}(x)-t(x)}{\hat{\sigma}(x)}\ge\sup_{x\in\tilde{\mathcal{X}}\backslash\mathcal{X}_+}\hat{t}(x)/\hat{\sigma}(x)>\hat{c}_{u,\alpha}\left(\tilde{\mathcal{X}}\backslash\hat{\mathcal{X}}_+^{\hat{j}-1}\right)\ge\hat{c}_{u,\alpha}\left(\tilde{\mathcal{X}}\backslash\mathcal{X}_+\right).$$

Thus, for $\mathcal{X}=\tilde{\mathcal{X}}\backslash\mathcal{X}_+$, we have that, on the event that $\hat{\mathcal{X}}_+\not\subseteq\mathcal{X}_+$, the event in (2) will not hold. Since the probability of this is asymptotically no greater than α, it follows that $P(\hat{\mathcal{X}}_+\not\subseteq\mathcal{X}_+)$ is asymptotically no greater than α, giving the result. □

Next we provide critical values that satisfy (2) for different estimators $\hat{t}(x)$ depending whether the covariate X_i is discrete or continuous. The inference procedure described below for the discrete covariate case parallels results described in Lee and Shaikh (2014) while the procedure for the continuous covariates case uses

results from the literature on uniform confidence bands and is new to the treatment effect literature.

3.1 Discrete covariates

Suppose that the support of X_i is discrete and takes on a finite number of values. We write

$$\tilde{\mathcal{X}} = \{x_1, \ldots, x_\ell\}$$

for the set $\tilde{\mathcal{X}}$ of values of the covariate under consideration, which we may take to be the entire support of X_i. In this setting, we may estimate the treatment effect $\hat{t}(x)$ with the sample analog. Let $N_{0,x} = \sum_{i=1}^n 1(D_i = 0, X_i = x)$ be the number of observations for which $X_i = x$ and $D_i = 0$, and let $N_{1,x} = \sum_{i=1}^n 1(D_i = 1, X_i = x)$ be the number of observations for which $X_i = x$ and $D_i = 1$. Let

$$\hat{t}(x_j) = \frac{1}{N_{1,x_j}} \sum_{1 \le i \le n, D_i = 1, X_i = x_j} Y_i - \frac{1}{N_{0,x_j}} \sum_{1 \le i \le n, D_i = 0, X_i = x_j} Y_i$$

We estimate the variance using

$$\hat{\sigma}^2(x_j) = \frac{1}{N_{1,x_j}} \sum_{1 \le i \le n, D_i = 1, X_i = x_j} \left(Y_i - \frac{1}{N_{1,x_j}} \sum_{1 \le i \le n, D_i = 1, X_i = x_j} Y_i \right)^2 / N_{1,x_j}$$
$$+ \frac{1}{N_{0,x_j}} \sum_{1 \le i \le n, D_i = 0, X_i = x_j} \left(Y_i - \frac{1}{N_{0,x_j}} \sum_{1 \le i \le n, D_i = 0, X_i = x_j} Y_i \right)^2 / N_{0,x_j}.$$

Under an i.i.d. sampling scheme, $\{(\hat{t}(x_j) - t(x_j))/\hat{\sigma}(x_j)\}_{j=1}^{\ell}$ converge in distribution jointly to ℓ independent standard normal variables. Thus, one can choose $\hat{c}_{u_\alpha}(\mathcal{X})$ to be the $1 - \alpha$ quantile of the maximum of $|\mathcal{X}|$ independent normal random variables where $|\mathcal{X}|$ is the number of elements in $\mathcal{X}$. Some simple calculations show that this gives

$$\hat{c}_{u,\alpha}(\mathcal{X}) = \Phi^{-1}\left((1 - \alpha)^{1/|\mathcal{X}|} \right) \tag{4}$$

where Φ is the cdf of a standard normal variable. For ease of calculation, we can also use a conservative Bonferroni procedure, which uses Bonferroni's inequality to bound the distribution of $|\mathcal{X}|$ variables with standard normal distributions regardless of their dependence structure. The Bonferroni critical value is given by

$$\hat{c}_{u,\alpha}(\mathcal{X}) = \Phi^{-1}(1 - \alpha/|\mathcal{X}|). \tag{5}$$

The Bonferroni critical values will be robust to correlation across the covariates (although $\hat{\sigma}$ would have to be adjusted to take into account serial correlation across the outcomes for a given x).

Both of these critical values will be valid as long as we observe i.i.d. data with finite variance where the probability of observing each treatment group is strictly positive for each covariate.

Theorem 2 *Suppose that the data are i.i.d. and* $P(D_i = d, X_i = x_j)$ *is strictly positive and* Y_i *has finite variance conditional on* $D_i = d, X_i = x_j$ *for* $d = 0, 1$ *and* $j = 1, \ldots, \ell$ *, and that Assumption*1 *holds. Then the critical values defined in* (4) *and* (5) *both satisfy* (2) *and* (3)*.*

3.2 Continuous covariates

For the case of a continuous conditioning variable, we can use results from the literature on uniform confidence bands for conditional means to obtain estimates and critical values that satisfy (2) (see, among others, Neumann and Polzehl, 1998; Claeskens, 2003; Chernozhukov et al., 2013; Kwon, 2022). For convenience, we describe the procedure here for multiplier bootstrap confidence bands based on local linear estimates, specialized to our case.

Let $m_1(x) = E(Y_i(1)|X_i = x)$ and $m_0(x) = E(Y_i(0)|X_i = x)$ be the average of potential outcomes with and without the treatment intervention given a fixed value of the covariate X_i. Under Assumption 1,

$$m_j(x) = E(Y_i(j)|X_i = x) = E(Y_i(j)|X_i = x, D_i = j) = E(Y_i|X_i = x, D_i = j), \quad j = 0, 1.$$

Let $X_i = (X_{i1} \ldots X_{id})$ and $x = (x_1 \ldots x_d)$. For a kernel function K and a sequence of bandwidths $h_1 \to 0$, define the local linear estimate $\hat{m}_1(x)$ of $m_1(x)$ to be the intercept term a for the coefficients a and $\{b_j\}_{j=1}^d$ that minimize

$$\sum_{1 \le i \le n, D_i = 1} \left[Y_i - a - \sum_{j=1}^{d} b_j (X_{i,j} - x_j) \right]^2 K((X_i - x)/h_1)$$

Similarly, define $\hat{m}_0(x)$ to be the corresponding estimate of $m_0(x)$ for the control group with $D_i = 0$ and h_0 the corresponding sequence of bandwidths. Let $\hat{\varepsilon_i} = Y_i - D_i \hat{m}_1(X_i) - (1 - D_i)\hat{m}_0(X_i)$ be the residual for individual i. Then define the standard error $s_1(x)$ of estimator $\hat{m}_1(x)$ as

$$s_1^2(x) = \frac{\sum_{1 \le i \le n, D_i = 1} [\hat{\varepsilon_i} K((X_i - x)/h_1)]^2}{\left[\sum_{1 \le i \le n, D_i = 1} K((X_i - x)/h_1) \right]^2}$$

and similarly define $s_0(x)$ for $\hat{m}_0(x)$.

Let n_1 and n_0 denote the sample sizes for the treatment and control group respectively. Let the estimator for the conditional average treatment effect be $\hat{t}(x) = \hat{m}_1(x) - \hat{m}_0(x)$ and its standard error $\hat{\sigma}(x) = \sqrt{s_1^2(x) + s_0^2(x)}$. To obtain the asymptotic properties of $\hat{t}(x)$, we use the following smoothness assumptions and assumptions on kernel function and bandwidths, which specialize the regularity conditions given in Chernozhukov et al. (2013) to our case.

Assumption 2

1. The observations $\{(X_i, D_i, Y_i)\}_{i=1}^n$ are i.i.d. and $P(D_i = 1|X_i = x)$ is bounded away from zero and one.
2. $m_0(x)$ and $m_1(x)$ are twice continuously differentiable and $\mathcal{X}$ is convex.
3. $X_i|D_i = d$ has a conditional density that is bounded from above and below away from zero on $\mathcal{X}$ for $d \in \{0, 1\}$.
4. Y_i is bounded by a nonrandom constant with probability one.
5. $(Y_i - m_d(x))|X_i = x, D_i = d$ has a conditional density that is bounded from above and from below away from zero uniformly over $x \in \mathcal{X}$ and $d \in \{0, 1\}$.
6. The kernel K has compact support and two continuous derivatives, and satisfies that $\int uK(u)\,du = 0$ and $\int K(u)\,du = 1$.
7. The bandwidth for the control group, h_0, satisfies the following asymptotic relations as $n \to \infty$: $nh_0^{d+2} \to \infty$, $nh_0^{d+4} \to 0$ and $n^{-1}h_0^{-2d} \to 0$ at polynomial rates. In addition, the same conditions hold for the bandwidth h_1 for the treated group.

Part 7 of Assumption 2 incorporates an undersmoothing assumption as well as assumptions on the bandwidth needed for technical reasons in the proofs of the results in Chernozhukov et al. (2013). The undersmoothing assumption leads to confidence bands that are suboptimal in rate, which may lead to a loss in power for our multiple testing procedure.

To approximate the supremum of this distribution over a non-degenerate set, we follow Neumann and Polzehl (1998) and Chernozhukov et al. (2013) and approximate $\hat{m}_1$ and $\hat{m}_0$ by simulating and using the following multiplier processes

$$\hat{m}_1^*(x) \equiv \frac{\sum_{1 \le i \le n, D_i = 1} \eta_i \hat{\varepsilon}_i K((X_i - x)/h_1)}{\sum_{1 \le i \le n, D_i = 1} K((X_i - x)/h_1)}$$

and

$$\hat{m}_0^*(x) \equiv \frac{\sum_{1 \le i \le n, D_i = 0} \eta_i \hat{\varepsilon}_i K((X_i - x)/h_0)}{\sum_{1 \le i \le n, D_i = 0} K((X_i - x)/h_0)}$$

where $\eta_1, \ldots, \eta_n$ are i.i.d. standard normal variables drawn independently of the data. To form critical values $\hat{c}_{u,\alpha}(\mathcal{X})$, we simulate S replications of n i.i.d. standard normal variables $\eta_1, \ldots, \eta_n$ that are drawn independently across observations and bootstrap replications. For each bootstrap replication, we form the test statistic

$$\sup_{x \in \mathcal{X}} \frac{\hat{t}^*(x)}{\hat{\sigma}(x)} = \sup_{x \in \mathcal{X}} \frac{\hat{m}_1^*(x) - \hat{m}_0^*(x)}{\hat{\sigma}(x)}. \tag{6}$$

The critical value $\hat{c}_{u,\alpha}(\mathcal{X})$ is taken to be the $1 - \alpha$ quantile of the empirical distribution of these S simulated replications.

To avoid issues with estimation at the boundary, we place some restrictions on the set $\tilde{\mathcal{X}}$ of values of the covariate under consideration. In practice, one can choose $\tilde{\mathcal{X}}$ in

the following theorem to be any set such that the kernel functions $x \mapsto K((x-\tilde{x})/h_0)$ and $x \mapsto K((x-\tilde{x})/h_1)$ are contained entirely in the support of X_i for all $\tilde{x} \in \tilde{\mathcal{X}}$.

Theorem 3 *Let $\tilde{\mathcal{X}}$ be any set such that, for some $\varepsilon > 0$, $\{\tilde{x} | \|\tilde{x} - x\| \leq \varepsilon \text{ for some } x \in \tilde{\mathcal{X}}\} \subseteq \text{supp}(X)$, where $\text{supp}(X)$ denotes the support of the X_i's. Suppose Assumptions 1 and 2 hold. Then the multiplier bootstrap critical value $\hat{c}_{u,\alpha}(\mathcal{X})$ defined above satisfies (2) and (3) for any $\mathcal{X} \subseteq \tilde{\mathcal{X}}$.*

Proof The critical value satisfies (2) by the arguments in Example 7 of Chernozhukov et al. (2013, pp. 7–9 of the supplementary appendix). The conditions in that example hold for the treated and untreated observations conditional on a probability one set of sequences of D_i. The strong approximations to $\hat{m}_0(x)$ and $\hat{m}_1(x)$ and uniform consistency results for $s_1(x)$ and $s_2(x)$ then give the corresponding approximation for $(\hat{m}_1(x) - \hat{m}_0(x))/\hat{\sigma}(x)$. The critical value satisfies Condition (3) by construction. □

The multiplier processes $m_1^*(x)$ and $m_0^*(x)$ and standard errors $s_1(x)$ and $s_0(x)$ given above follow Chernozhukov et al. (2013), who use a Nadaraya–Watson (local constant) estimator with an equivalent kernel as an asymptotic approximation to the local polynomial estimator (see Fan and Gijbels, 1996, Section 3.2.2 for a definition and discussion of equivalent kernels). The formula for the equivalent kernel given in Chernozhukov et al. (2013) requires restricting attention to points on the interior of the support of X_i, which leads to the additional conditions[2] on the set $\tilde{\mathcal{X}}$ used in Theorem 3. For local linear estimators on the interior of the support of X_i, this equivalent kernel is the same as the original kernel, which leads to form of the multiplier processes given above.

3.3 Extension: testing for treatment effect heterogeneity

The inference procedure described above can be easily modified to test for treatment effect heterogeneity. Here we focus on the continuous covariate case since the testing problem in the discrete covariate case is well-studied in the multiple comparison literature. Let t be the (unconditional) average treatment effect. The null hypothesis of treatment effect heterogeneity is

$$H_0 : t(x) = t \quad \forall x.$$

Let $\mathcal{X}_{+-} = \{x | t(x) \neq t\}$ and $\hat{\mathcal{X}}_{+-}$ be a set that satisfies

$$\liminf_n P\left(\hat{\mathcal{X}}_{+-} \subseteq \mathcal{X}_{+-}\right) \geq 1 - \alpha.$$

The probability that $\hat{\mathcal{X}}_{+-}$ includes some value(s) of x such that $t(x) = t$ cannot exceed the significance level α. Then the decision rule of the test is to reject H_0 if the set $\hat{\mathcal{X}}_{+-}$ is nontrivial.

[2] It appears that such conditions are an unstated assumption needed for the results in Kong et al. (2010) used in Example 7 of Chernozhukov et al. (2013).

The set $\hat{\mathcal{X}}_{+-}$ is in fact more informative than simply testing the null hypothesis of no treatment effect heterogeneity. It also helps researchers determine for which values of the conditioning covariate X_i the conditional average treatment effect differs from its average over the entire population. The set $\hat{\mathcal{X}}_{+-}$ can be obtained using a method similar to that described in the previous section for set $\hat{\mathcal{X}}_{+}$. Let t denote the unconditional average treatment effect, and $\hat{c}_{\text{het},\alpha}(\mathcal{X})$ the critical value of this test for treatment effect heterogeneity. It satisfies

$$\liminf_n P\left(\sup_{x\in\mathcal{X}}\left|\frac{\hat{t}(x)-\hat{t}-(t(x)-t)}{\hat{\sigma}(x)}\right| \le \hat{c}_{\text{het},\alpha}(\mathcal{X})\right) \ge 1-\alpha,$$

where $\hat{t}$ is a $\sqrt{n}$-consistent estimator of t. Let $\hat{\mathcal{X}}^1_{+-} \equiv \left\{x\in\tilde{\mathcal{X}}\middle||(\hat{t}(x)-\hat{t})/\hat{\sigma}(x)|>\hat{c}_{\text{het},\alpha}(\tilde{\mathcal{X}})\right\}$. For $k>1$, let $\hat{\mathcal{X}}^k_{+-} = \left\{x\in\tilde{\mathcal{X}}\middle||(\hat{t}(x)-\hat{t})/\hat{\sigma}(x)|>\hat{c}_{\text{het},\alpha}(\tilde{\mathcal{X}}\backslash\hat{\mathcal{X}}^{k-1}_{+-})\right\}$. When $\hat{\mathcal{X}}^k_{+-}=\hat{\mathcal{X}}^{k-1}_{+-}$, or when the two sets are close enough to some desired level of precision, stop and take $\hat{\mathcal{X}}_{+-}=\hat{\mathcal{X}}^k_{+-}$. In practice, $\hat{c}_{\text{het},\alpha}(\mathcal{X})$ could be set as the $1-\alpha$ quantile of the empirical distribution of the multiplier bootstrap statistic $\sup_{x\in\mathcal{X}}\left|\frac{\hat{t}^*(x)-\hat{t}^*}{\hat{\sigma}(x)}\right|$, where $\hat{t}^*(x)$ is the multiplier process defined earlier and $\hat{t}^*$ is the estimator for t in the simulated dataset.

4 Monte Carlos

In this section, we investigate the small sample behavior of our proposed inference procedure for optimal treatment assignment. We consider three data generating processes (DGPs) for the conditioning variable X_i, the outcome Y_i and the treatment indicator D_i.

DGP 1: $X_i\sim U(0,1)$, $e_i\sim N(0,1/9)$, $v_i\sim U(0,1)$, $D_i=1(0.1X_i+v_i>0.55)$, $Y_i=5(X_i-0.5)^2+5(X_i-0.5)^2D_i+e_i$;
DGP 2: $X_i\sim U(0,1)$, $e_i\sim N(0,1/9)$, $v_i\sim U(0,1)$, $D_i=1(0.1X_i+v_i>0.55)$, $Y_i=0.5\sin(5X_i+1)+0.5\sin(5X_i+1)D_i+e_i$;
DGP 3: $X_i\sim U(0,1)$, $e_i\sim N(0,1/9)$, $v_i\sim U(0,1)$, $D_i=1(0.1X_i+v_i>0.55)$, $Y_i=10(X_i-1/2)^2+e_i$.

The unconfoundedness assumption is satisfied in all three DGPs. The conditional average treatment effect $t(x)$ is the difference between the conditional mean $m_1(x)=E(Y_i|X_i=x,D_i=1)$ and $m_0(x)=E(Y_i|X_i=x,D_i=0)$. In the first DGP, $t(x)$ always lies above zero except for one tangent point. In the second DGP, $t(x)=0.5\sin(5x+1)$ is positive in some parts of the X_i support and negative in the other parts. In the third DGP, $t(x)$ is uniformly zero.

For each DGP, datasets are generated with three different sample sizes and repeated 500 times. The conditional mean $m_0(x)$ and $m_1(x)$ are estimated using local

linear estimation with Epanechnikov kernel and bandwidths chosen by following rule of thumb:

$$h_l = \hat{h}_{l,ROT} \times \hat{s}_l \times n_l^{1/5-1/4.75} \quad l = 0, 1,$$

where $\hat{s}_l$ is the standard deviation of X_i in the subsample with $D_i = l$, and $n_l^{1/5-1/4.75}$ is used to ensures under-smoothing, $l = 0, 1$. $\hat{h}_{l,ROT}$ minimizes the weighted Mean Integrated Square Error (MISE) of the local linear estimator with studentized X_i values and is given by Fan and Gijbels (1996):

$$\hat{h}_{l,ROT} = 1.719 \left[\frac{\tilde{\sigma}_l^2 \int w(x)dx}{n_l^{-1} \sum_{i=1}^{n_l} \left\{ \tilde{m}_l^{(2)}(X_i) \right\}^2 w(X_i)} \right]^{1/5} n_l^{-1/5}.$$

In the formula, $\tilde{m}_l^{(2)}$ is the second-order derivative of the quartic parametric fit of $m_l(x)$ with studentized X_i and $\tilde{\sigma}_l^2$ is the sample average of squared residuals from the parametric fit. $w(.)$ is a weighting function, which is set to 1 in this section. The computation is carried out using the np package in R (see Hayfield and Racine, 2008). To avoid the boundary issue, the local linear estimator $\hat{t}(x)$ is evaluated between 0.2 and 0.8. The critical values for the proposed step-down procedure are data dependent and calculated using the multiplier bootstrap method with $S = 500$ for each simulated dataset.

Before reporting the Monte Carlo results for all 500 simulations, we first illustrate the implementation of our proposed inference procedure using graphs. The left panel of Fig. 1 reports the true CATEs and the local linear estimates of the CATEs based on one randomly simulated sample of size 500. The right panel reports studentized CATE estimates, the true optimal treatment set $\mathcal{X}_+$ and the proposed inference region $\hat{\mathcal{X}}_+$ for the optimal treatment set. The optimal treatment set contains all x values with positive CATE. The confidence region $\hat{\mathcal{X}}_+$ includes all x values with studentized CATE estimates lying above the final step-down critical value.

The confidence region $\hat{\mathcal{X}}_+$ for the optimal treatment set controls familywise error rates properly. As a comparison, the right panel of Fig. 1 also reports treatment sets based on pointwise confidence bands. These sets are constructed as the region where the studentized CATE estimates lie above 1.645, the 95% quantile of standard normal distribution.

We see from the graphs that the proposed confidence regions are no wider than the pointwise treatment sets. That is expected because the latter does not control the error rate correctly. The figure for DGP 3 gives an example where the pointwise treatment set gives very misleading treatment assignment information regarding a policy treatment that has no effect at all. The step-down method improves the power of the inference procedure for both DGP 1 and DGP 2. As is noted in the figure subtitle, the total number of steps for critical value calculation is 4 for DGP 1 and 3 for DGP 2. The step-down refinement does not lead to improvement for DGP 3 because the initial confidence region is a null set.

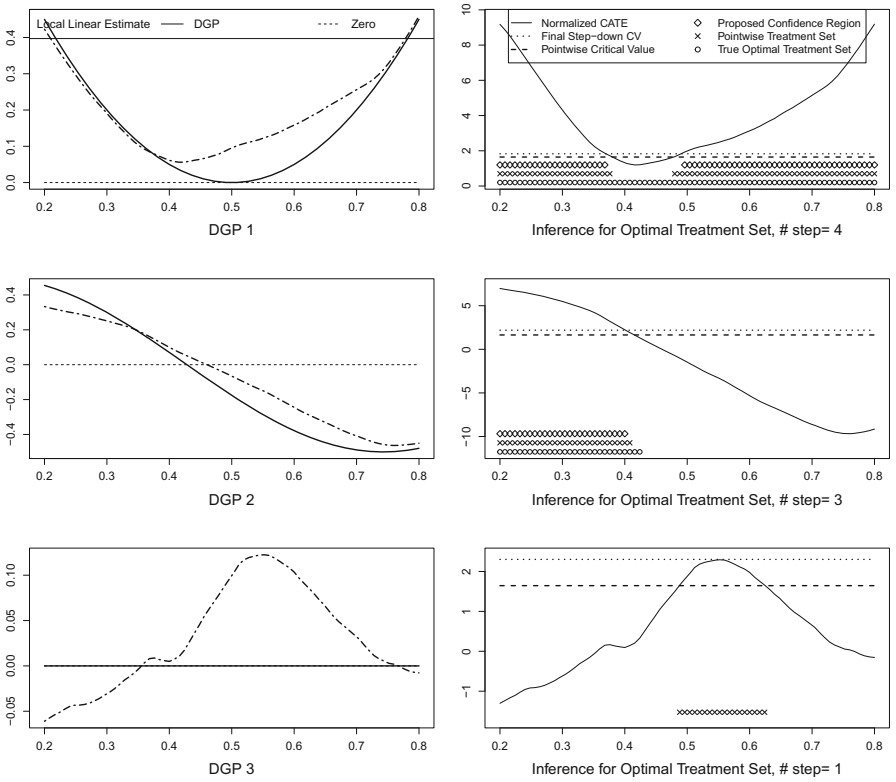

Fig. 1 CATE estimates, critical values, and treatment sets

Although the simulation that makes Fig. 1 is specially selected for illustration purposes, the good performance of the proposed inference procedure holds when we look at results from all 500 simulations. Columns (3)–(6) and (9)–(12) in Table 1 report the size and power of the proposed confidence region $\hat{\mathcal{X}}_+$ obtained with and without applying the step-down refinement of critical values. The associated nominal familywise error rate is 0.05 for columns (3)–(6) and 0.1 for columns (9)–(12). The size measure used is the empirical familywise error rates (EFER), the proportion of simulation repetitions for which $\hat{\mathcal{X}}_+^1$ ($\hat{\mathcal{X}}_+$) is not included in the true set $\mathcal{X}_+$. The power is measured by the average proportion of false hypothesis (correctly) rejected (FHR), or the average among 500 repetitions of the ratio between the length of $\hat{\mathcal{X}}_+^1 \cap \mathcal{X}_+$ ($\hat{\mathcal{X}}_+ \cap \mathcal{X}_+$) and the length of the true optimal treatment set $\mathcal{X}_+$. The size measure is denoted in the table as EFER and EFER-SD for the step-down method. The power measure is denoted as FHR and FHR-SD for the step-down method. We see from results reported in these columns that the proposed confidence region for the optimal treatment set controls familywise error rates very well. In the case of DGP 3 where the least favorable condition of the multiple hypothesis testing holds and the conditional average treatment effect equals to zero uniformly, the familywise error rates are close to the targeted significance level especially when the sample size is

Table 1 Size and power properties of the proposed inference method

	PW, $\alpha = 0.05$		Uniform, $\alpha = 0.05$				PW, $\alpha = 0.1$		Uniform, $\alpha = 0.1$			
	EFER (1)	FHR (2)	EFER (3)	FHR (4)	EFER-SD (5)	FHR-SD (6)	EFER (7)	FHR (8)	EFER (9)	FHR (10)	EFER-SD (11)	FHR-SD (12)
DGP1: $t(x) \geq 0$ for all $x \in \tilde{\mathcal{X}}$												
N = 500	0*	0.6814	0*	0.5870	0*	0.6327	0*	0.7392	0*	0.6295	0*	0.6917
N = 1000	0*	0.7141	0*	0.6316	0*	0.6739	0*	0.7639	0*	0.6664	0*	0.7221
N = 2000	0*	0.7584	0*	0.6807	0*	0.7228	0*	0.8009	0*	0.7112	0*	0.7647
DGP2: $t(x) \geq 0$ only for some $x \in \tilde{\mathcal{X}}$												
N = 500	0.0540	0.8717	0.0080	0.8129	0.0120	0.8234	0.0980	0.9005	0.0200	0.8410	0.0320	0.8527
N = 1000	0.0420	0.9033	0.0060	0.8579	0.0060	0.8665	0.0840	0.9255	0.0140	0.8774	0.0220	0.8880
N = 2000	0.0300	0.9375	0.0040	0.9001	0.0060	0.9064	0.0660	0.9516	0.0060	0.9137	0.0160	0.9206
DGP3: $t(x) = 0$ for all $x \in \tilde{\mathcal{X}}$												
N = 500	0.2640	/#	0.0840	/#	0.0840	/#	0.4120	/#	0.1400	/#	0.1400	/#
N = 1000	0.2860	/#	0.0740	/#	0.0740	/#	0.4220	/#	0.1260	/#	0.1260	/#
N = 2000	0.2720	/#	0.0560	/#	0.0560	/#	0.4460	/#	0.1160	/#	0.1160	/#

Note: *, EFER is equal to 0 by construction for DGP 1 since the set where the null hypothesis is false is the support of X. #, the proportion of false hypotheses (correctly) rejected is not defined in DGP 3 since the set where the null hypothesis is false has by construction measure zero

larger. Comparing results of DGPs 1 and 2 in columns (5)–(6), (11)–(12) to those in columns (3)–(4), (9)–(10), we also see that the power of our procedure increases when the step-down refinement is used for the calculation of the critical values.

For comparison purposes, we also report in Table 1 the size and power properties of confidence regions obtained from pointwise confidence intervals, or all x values that reject the pointwise null hypothesis that $t(x)$ is negative. Comparing the results in columns (1)–(2) and (7)–(8) to their uniform counterparts, we see that the pointwise sets, as expected, fail to control the familywise error rate. In the case of DGP 3, where the true average treatment effect is zero for all x values, the chance that the pointwise set estimator discover some falsely identified nonempty positive treatment set more than quadruples the significance level, regardless of the sample size.

5 Empirical example: the STAR project

Project STAR was a randomized experiment designed to study the effect of class size on students' academic performance. The experiment took place in Tennessee in the mid-1980s. Teachers as well as over eleven thousand students in 79 public schools were randomly assigned to either a small class (13–17 students), regular-size class (22–25 students), or regular-size class with a full time teacher aide from grade K to 3. Previous papers in the literature find that attending small classes improves student outcomes both in the short run in terms of Stanford Achievement Test scores (Krueger, 1999) and in the long run in terms of likelihoods of taking college-entrance exam (Krueger and Whitmore, 2001), attending college and in terms of earnings at age 27 (Chetty et al., 2010). Previous papers also find that students benefit from being assigned to more experienced teacher in kindergarten, but little has been said about whether and how the effect of reducing class size varies with teacher experience. The nonparametric analysis in this section sheds new light on this question. We find small class matters most for students taught by inexperienced teachers, especially those in poor schools. We use this heterogeneity to study the optimal assignment of the small class treatment.

The upper panel of Fig. 2 plots the conditional mean estimates of grade K test score percentiles (defined in the footnote of Fig. 2) conditional on teacher experience and class type.[3] Regardless of whether schools are located in disadvantaged neighborhoods (defined by whether more than half of the students receive free lunch), the positive effect of attending a small class is larger if the student is taught by a teacher with some but not a lot of experience in teaching. The nonparametric estimates also suggest that reducing class size may hurt student performance in classes taught by very experienced teachers. One might argue that very experienced teachers have set ways of teaching and hence less incentive to adapt to a small class size. But one needs to keep in mind that these negative effects are imprecisely

[3] We do not look at test scores at higher grades because after grade K there is self-selection into small classes. According to Krueger (1999), "Approximately 10% of students switched between small and regular classes between grades, primarily because of behavioral problems or parental complaints."

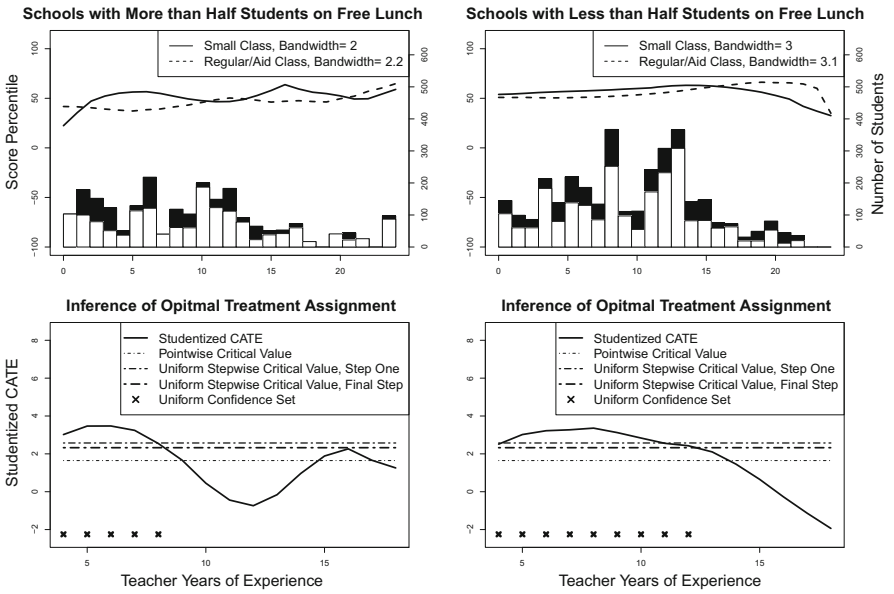

Fig. 2 Optimal Treatment Assignment Based Teacher Experience. Note: In the top panel, score percentiles are defined following Krueger (1999), where student scores from all type of classes are translated to score percentile ranks based on the total score distribution in regular and regular/aide classes. The shaded bars in the top panel represent the number of students assigned to small classes given teacher experience and white bars represent the number of students assigned to regular/aide classes. In the bottom panel, Studentized CATEs are conditional average treatment effects divided by their pointwise standard error. Pointwise Critical Value is equal to 1.645 for one-sided testing with 5% significance level. Uniform Stepwise Critical Values and Confidence Sets are obtained following our proposed optimal treatment assignment procedure. Nonparametric estimation uses the Epanechnikov kernel and the rule-of-thumb bandwidth discussed in Sect. 4. Multiplier bootstraps for inference are carried out 1000 times. Codes for replication are available on the authors' websites

estimated due to the small sample size at the right tail of the teacher experience distribution. Therefore, it is important to apply the proposed inference method to determine whether the data are precise enough to give evidence for this negative effect.

Since we examine treatment effect heterogeneity along both dimensions of teacher experience and school characteristic, the conditioning set x for the treatment effect effect $t(x)$ defined in inequality (2) is two dimensional. Specifically, let $t(x_1, 1)$ (and $t(x_1, 0)$) be the treatment effect of the small classroom intervention for students in schools with (and without) more than half of students on free lunch and in classrooms with a teacher having x_1 years of experience. Treating the year of teacher experience as a continuous covariate, the treatment effect is estimated through subsample local linear regressions combining the discussions in Sects. 3.1 and 3.2. We take the supremum over the multiplier processes for the studentized estimates of both $t(x_1, 1)$ and $t(x_0, 0)$ when forming our critical values. Following our proposed inference method, the resulted optimal treatment set is also two dimensional and derived based on uniform inference over both dimensions of teacher experience and school characteristic.

In addition, since students in the same school may face common shocks that affect test scores, we modify the inference procedure described in Sect. 3 to allow for data clustering. Let $i = 1, 2, ..., N$ denotes individuals and $j = 1, 2, \ldots, J$ denotes schools. To account for potential within-school error term dependence, we substitute the multiplier processes used in (6) by $\hat{m}_0^{**}(x)$ and $\hat{m}_1^{**}(x)$ with

$$\hat{m}_l^{**}(x) \equiv \frac{\sum_{1 \le i \le n, D_i = l} \eta_j \hat{\varepsilon}_{ij} K((X_{ij} - x)/h_l)}{\sum_{1 \le i \le n, D_i = l} K((X_{ij} - x)/h_l)}, \quad l = 0, 1,$$

where $\eta_1, \ldots, \eta_J$ are i.i.d. standard normal random variables drawn independently of the data following the wild cluster bootstrap suggestion in Cameron et al. (2008). We also substitute the standard error $\hat{\sigma}(x)$ used to construct the test statistic in equation (2) with a null-imposed wild cluster bootstrap standard error suggested in Cameron et al. (2008) using the Rademacher weights (+1 with probability 0.5 and 1 with probability 0.5). The critical value is then taken to be the $1 - \alpha$ quantile of the empirical distribution of the supremum estimator described in (6). We conjecture that, as with other settings with nonparametric smoothing, accounting for dependence is not technically necessary under conventional asymptotics but will lead to nontrivial finite sample improvement.

The bottom panel of Fig. 2 studies the statistical inference of optimal treatment assignment assuming zero cost relative to the small class treatment. Given the rule-of-thumb bandwidth (reported in graphs in the top panel) and the support of teacher experience, we conduct the inference exercise for teachers with 4–18 years of experience to avoid the boundary issue described in Sect. 3. With a 95% confidence level, the confidence set contains teachers with 4–8 years of experience in schools with more than half students on free lunch, as well as teachers with 4–12 years of experience in schools with less than half students on free lunch. The results suggest that for both types of schools, assigning small classes to teachers who are relatively new but not completely new to teaching improves students' test score on average. One should notice that although the confidence sets for optimal treatment assignment (assuming zero cost) are similar for both types of schools, the average score improvement is much larger in the first type of disadvantaged schools. If one takes into consideration that the cost of reducing class size is roughly equivalent to the benefit of a 2.22 percentile score increase (roughly calculated break-even point for the intervention as explained in the footnote of Fig. 3), the confidence set for optimal treatment assignment will only include classrooms taught by teachers with 4–7 years of experience in disadvantaged schools, as is shown in graph (a) of Fig. 3.

What about the very experienced teachers? Does the inference method say anything against assigning experience teachers to small classes? If the null hypothesis is whether the effect of the small class intervention is zero, the rejection region does not include very experienced teachers across both types of schools. If the null hypothesis is whether the effect of small class intervention is 2.22 percentile, the step-wise method picks out both inexperienced teachers in disadvantage schools and teachers with 18 years of experience in schools with less than half students on free lunch, as is demonstrated in Fig. 3b. Apart from this one group of very experienced teachers, the graph suggest that the effect of small class intervention is not

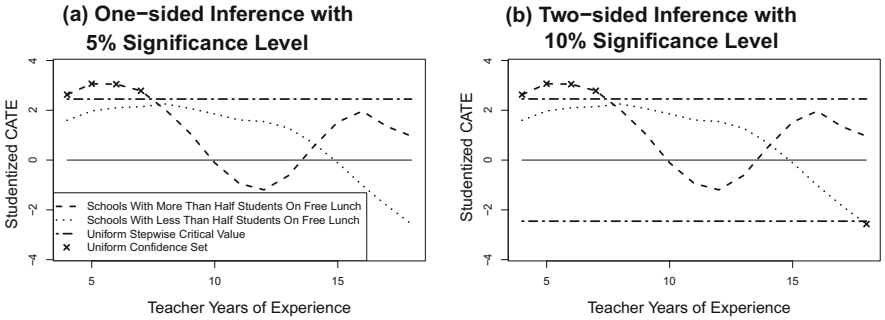

Fig. 3 Optimal Treatment Assignment With Nonzero Treatment Cost. Note: The graph is based on the cost-benefit analysis conducted in Chetty et al. (2010), Online Appendix C. Specifically, the cost of reducing class size is roughly $(22.56/15.1 - -1) \times \$8848 = \$4371$ per student per year in 2009 dollars. (The annual cost of school for a child is \$8,848 per year. Small classes had 15.1 students on average, while large classes had 22.56 students on average.) On the other hand, the benefit of 1 percentile increase in test score is roughly \$1968 (Chetty et al., 2010 states a \$9,460 benefit for a 4.8 percentile increase in test score, derived assuming constant wage return to score increase) per student when life-time earning increase driven by early childhood test score increase is discounted at present values and measured in 2009 dollars. Therefore, the break-even point of class size reduction for the STAR project is an average test score increase of 2.22 percentile. Studentized CATEs are the conditional average treatment effects divided by their pointwise standard errors. Uniform Stepwise Critical Values and Confidence Sets are obtained following our proposed optimal treatment assignment procedure. Nonparametric estimation uses the Epanechnikov kernel and the rule-of-thumb bandwidth discussed in Sect. 4. Multiplier bootstraps for inference are carried out 1,000 times. Codes for replication are available on the authors' websites

distinguishable from the alternative cost of the intervention for very experienced teachers.

Next we provide a nonparametric analysis of treatment effect heterogeneity using the method discussed in Sect. 3.3. Here, we form our estimates at the level of the individual student, and we condition on teacher experience as well as student gender and free lunch status. As with our classroom level specification, we form critical values that are uniform over both the continuous variable (teacher experience) and the discrete variable (given here by student gender interacted with free lunch status).

Previous papers in the literature find that the effect of attending small class is larger for boys and for students from disadvantaged backgrounds. The nonparametric estimates plotted in Fig. 4a reinforce these findings. Specifically, the multiple testing for the positive treatment effect reported in Fig. 4b shows that the score improvement reported in Fig. 2 is driven by boys and by girls who receive free lunch. This finding supports the theoretical results in Lazear (2001) who predicts that the effect of reducing class size is larger for students with worse initial performance. Furthermore, in contrast to Whitmore (2005) who finds no significant gender and ratio differences in the effect of attending small classes, our nonparametric analysis rejects the null hypothesis of treatment effect homogeneity with a 5% significance level. The corresponding test statistic is 3.35, and the simulated critical value is 3.13. Figure 4c shows that the rejection of treatment effect homogeneity is driven by boys who do not receive free lunch assigned to teachers with 4 and 5 years of experience.

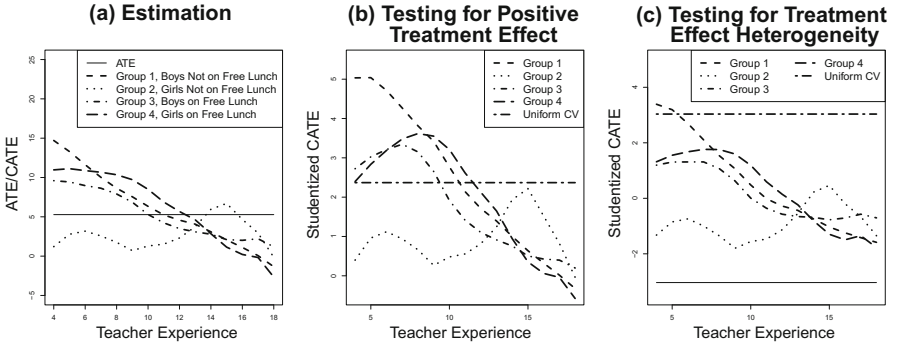

Fig. 4 Treatment Effect Heterogeneity Across Student Groups. Note: The ATE reported in figure (**a**) is the unconditional average treatment effect of the small class intervention. CATEs reported in graph (**a**) are conditional average treatment effects given teacher experience and group definition. Studentized CATEs reported in graphs (**b**) and (**c**) are CATEs divided by their standard errors. Uniform CVs are obtained following our proposed stepwise procedure. Nonparametric estimation uses the Epanechnikov kernel and the rule-of-thumb bandwidth discussed in Sect. 4. Codes for replication are available on the authors' websites

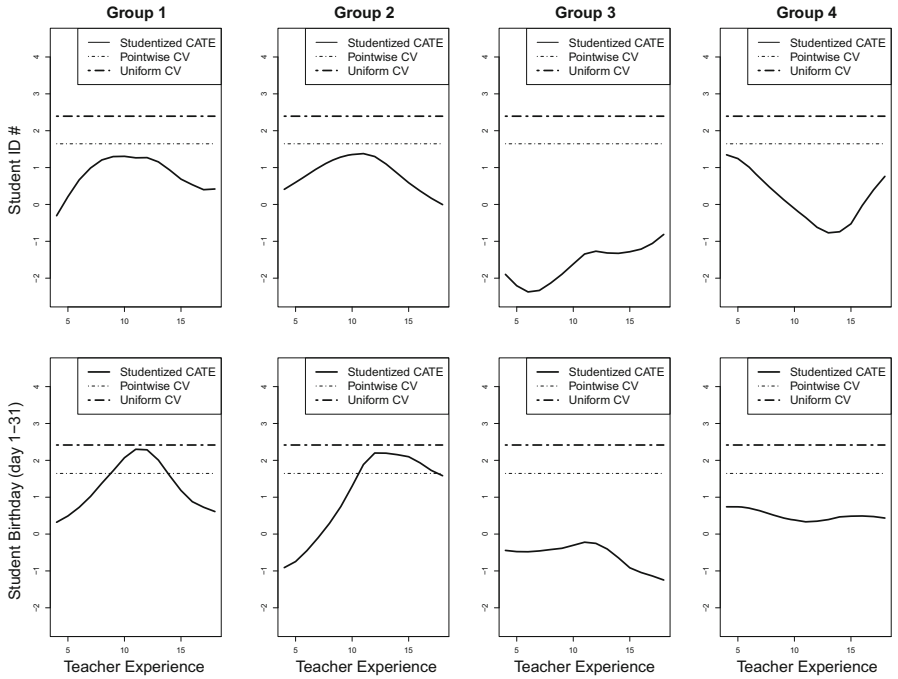

Fig. 5 Falsification Tests. Note: Groups are defined as in Fig. 4. Studentized CATEs are the average conditional treatment effect divided by their standard errors. Pointwise CV is 1.645. Uniform CV is the critical value obtained following our proposed stepwise procedure. Nonparametric estimation uses the Epanechnikov kernel and the rule-of-thumb bandwidth discussed in Sect. 4. Codes for replication are available on the authors' websites

Figure 5 provides a check on our method using student id number and student birthday (day 1 to day 31) as falsification outcomes. We expect that the treatment will have no effect on these outcomes, so that our 95% confidence region will be empty with probability.95 if it delivers the promised coverage. With 95% confidence level, our proposed confidence region for optimal treatment assignment is empty, indicating that the treatment is not at all helpful in improving the falsification outcomes. However, if one constructs confidence region based on the pointwise one-sided critical value 1.645, one would falsely select out some teacher experience and student characteristic combinations and conclude that, for such combinations, the small class treatment has positive effects on students' birthday (day 1 to day 31). Figure 5 reinforces the motivation of our proposed optimal treatment assignment procedure.

6 Conclusion

This paper formulates the problem of forming a confidence region for treatment rules that would be optimal given full knowledge of the distribution of outcomes in the population. We have proposed a solution to this problem by pointing out a relationship between our notion of a confidence region for this problem and a multiple hypothesis testing problem. The resulting confidence regions provide a useful complement to the statistical treatment rules proposed in the literature based on other formulations of treatment as a statistical decision rule. Just as one typically reports confidence intervals in addition to point estimates in other settings, we recommend that the confidence regions proposed here be reported along with the statistical treatment rule resulting from a more liberal formulation of the treatment problem. In this way, readers can assess for which subgroups there is a preponderance of empirical evidence in favor of treatment.

Appendix

A Motivation for the coverage criterion

In this appendix, we provide a more detailed discussion of settings where our confidence set $\hat{\mathcal{X}}_+$ may be of interest.

A.1 Intrinsic interest in the conditional average treatment effect

The CATE and the population optimal treatment rule it leads to are often of intrinsic interest in their relation to economic theory and its application to the design of policy interventions. In such settings, one is interested in how the evidence from a particular study adds to the body of scientific knowledge and overall evidence in favor of a particular theory, rather than (or in addition to) the more immediate question of how the intervention at hand should be implemented. Our confidence set $\hat{\mathcal{X}}_+$ guarantees frequentist coverage properties, which can be used by researchers in forming or testing scientific theories regarding the treatment being studied.

In our application to the class size intervention in Project STAR, we estimate the average treatment effects of assigning kindergarten students to small class on their test scores conditional on teacher experience and whether a school has more than half of its students receiving free or reduced-price lunch (later on we follow the literature and just use the term "free lunch"). The estimated CATE is larger for less experienced teachers, and is smaller and even negative for more experienced teachers. One may speculate about the reasons for this (perhaps experience allows teachers to overcome the negative effects of large class sizes, or perhaps more experienced teachers have difficulty adapting their teaching to take advantage of smaller class sizes), but, before doing so, it is of interest to determine whether the data are precise enough to give evidence in favor of this effect at all. Our confidence region determines at a 5% level that the effect is indeed positive for less experienced teachers. However, using a version of our procedure with the definitions of $Y_i(1)$ and $Y_i(0)$ switched, one can see that the negative effect for more experienced teachers is not statistically significant.

We note that, for certain hypotheses regarding the population optimal treatment rule, one can use a standard hypothesis test that does not correct for multiplicity. In our application, one can test, for example, the null hypothesis that students in the first type of school taught by teachers with 10 years of experience do not benefit from smaller classes using a standard z-test. This works as long as (1) the researcher chooses the hypothesis without using the data, and (2) the null hypothesis takes a simple form involving a predetermined value of the covariate. While there are certainly applications where these criteria are met, (1) becomes an issue whenever correlations in the data suggest new hypotheses to test. Indeed, given that our data set contains information on school characteristics and student demographics as well as teacher experience, it seems difficult to argue convincingly that teacher experience should have been the focus of our study a priori, particularly if the interaction effect described above was not expected. Using our approach, we uncover heterogeneity over teacher experience as well as other dimensions, while controlling for the possibility that the decision to focus on heterogeneity along a particular dimension was influenced by correlations in the data.

Regarding issue (2), one could try to determine whether the optimal treatment assignment favors less experienced teachers by testing the null hypothesis, say, that teachers with less than six years of experience have a larger CATE than those with six or more years of experience, but this clearly involves some arbitrary choices (such as the number six in this case). With our methods, if one finds that $\hat{\mathcal{X}}_+$ includes

all teachers with experience within a certain range, while the set formed analogously by reversing the roles of $Y_i(1)$ and $Y_i(0)$ shows that teachers with experience above a certain threshold should not be treated, one can conclude that the population optimal treatment assignment favors less experienced teachers.

In other settings as well, studies of optimal treatment assignment often have, as an additional goal, the formulation and testing of theories explaining observed heterogeneity in treatment effects. In their study of insecticide-treated nets, Bhattacharya and Dupas (2012) discuss reasons why uptake may vary along covariates such as wealth and children's age (see Section 7.1). The multiple testing approach of the present paper could be used to assess whether heterogeneity in CATEs is consistent with these theories, while taking into account the possibility that the theories themselves were formulated based on an initial inspection of the data.

A.2 "Do No Harm" and other decision criteria

Policy decisions regarding treatment assignment often involve considerations other than expected welfare maximization of the form that would lead to a decision theoretic formulation with the negative of the sum of individual utility functions defining the loss function. When evaluating new medical treatments, the United States Food and Drug Administration requires clinical trials that include statistical hypothesis tests with a null hypothesis of ineffective or harmful treatment. This can be interpreted as following the "do no harm" directive of the Hippocratic Oath. For economic policy, a similar interpretation may be given to arguments that government intervention should be justified by a certain degree of certainty that the policy will not be ineffective or harmful. The notion of coverage satisfied by our confidence set $\hat{\mathcal{X}}_+$ can be regarded as an extension of this criterion to the setting where treatment rules based on stratification by a covariate are a possibility.

Given that a policy-maker with an objective function based on the above interpretation of the "do no harm" objective may indeed wish to implement our confidence set $\hat{\mathcal{X}}_+$ as a treatment rule, it is of interest to ask how this rule performs under the expected welfare risk considered by Manski (2004). That is, how much will a policy maker who cares about expected welfare lose by implementing $\hat{\mathcal{X}}_+$ (perhaps out of a desire to avoid debate with a rival policy maker who prefers "do no harm")? In Appendix C, we derive some of the expected welfare properties of $\hat{\mathcal{X}}_+$ when used as a statistical treatment rule.

A.3 Political economy

In addition to asking about average welfare, one may be interested in how welfare gains and losses are distributed over the population, and how this relates to observed variables. This can have implications for political economy questions regarding what types of individuals one might expect to support a given policy. Depending on the assumptions made about the information sets and objectives of policy makers and those affected by the policy, our confidence set $\hat{\mathcal{X}}_+$ can be used to answer such questions, as we now illustrate.

Suppose that a policy is being considered that would have CATE $t(x)$. In forming an opinion about this policy, individual i knows his or her covariate X_i and the distributions $F_1(s|X_i) = P(Y_i(1) \leq s|X_i)$ and $F_0(s|X_i) = P(Y_i(0) \leq s|X_i)$ of outcomes for treatment and non-treatment conditional on this value of the covariate, but has no further information about his or her place within these distributions. If Y_i is given in units of individual i's Bernoulli utility, individual i will be in favor of the policy if $t(X_i) > 0$. Thus, under these assumptions, our criterion gives a set $\hat{\mathcal{X}}_+$ such that all individuals with $X_i \in \hat{\mathcal{X}}_+$ will be in favor of the policy. Our approach can then be used to see whether heterogeneity of the treatment effect along the support of the covariate X_i can explain political opinions and voting behavior.

Note the importance of information assumptions. The information assumptions above will be reasonable when an individual's standing in the outcome distribution is sufficiently uncertain. For example, it may be reasonable to assume that individuals have little knowledge of whether they will benefit from a job training program more or less than their peers. In contrast, if each individual i has full knowledge of $Y_i(0)$ and $Y_i(1)$ before the treatment, then support for the policy among individuals with covariate X_i will be determined by $P(Y_i(1) > Y_i(0)|X_i)$, which is not identified without further assumptions (see Fan and Park, 2010).

As another example, suppose that, in addition to experimental data satisfying our conditions, we observe another setting where a policy maker assigns treatment to some group $\mathcal{X}^*$. We wish to test the null hypothesis that the policy-maker is fully informed about the CATE and is maximizing expected welfare against the alternative that the policy maker has a different objective or information set. In our notation, this null hypothesis can be written as $H_0 : \mathcal{X}_+ = \mathcal{X}^*$, and rejecting when $\hat{\mathcal{X}}_+ \not\subseteq \mathcal{X}^*$ provides a level α test, which, in the case of rejection, can further be used to find groups that would be treated differently by a fully informed, welfare maximizing policy maker. Considering our application to the STAR experiment, we find that the population optimal treatment assignment gives small classes to less experienced teachers. If we found that small classes were given to a different set of teachers in a similar setting, this could be taken as evidence about the motives or information sets of the decision-makers involved.

B Two-sided confidence sets for $\mathcal{X}_+$

In this appendix, we develop two-sided confidence sets for the population optimal treatment set $\mathcal{X}_+$ based on a single step version of our procedure with a two-sided critical value. Formally, our goal is to form sets $\hat{\mathcal{X}}_+^{\text{inner}}$ and $\hat{\mathcal{X}}_+^{\text{outer}}$ such that

$$\liminf_n P\big(\hat{\mathcal{X}}_+^{\text{inner}} \subseteq \mathcal{X}_+ \subseteq \hat{\mathcal{X}}_+^{\text{outer}}\big) \geq 1 - \alpha \tag{7}$$

We also note that, for the notion of two-sided coverage given by (7), inverting a two-sided step-down test will not guarantee coverage when the number of steps is greater than one, unless additional conditions hold. The issues have to do with so-called directional errors, as we discuss further below.

For a set $\tilde{\mathcal{X}}$ of values of x under consideration, let the critical value $\hat{c}_{|\cdot|,\alpha}(\tilde{\mathcal{X}})$ satisfy

$$\liminf_n P\left(\sup_{x\in\tilde{\mathcal{X}}}\frac{|\hat{t}(x)-t(x)|}{\hat{\sigma}(x)}\le\hat{c}_{|\cdot|,\alpha}(\tilde{\mathcal{X}})\right)\ge 1-\alpha. \tag{8}$$

Let the sets $\hat{\mathcal{X}}_+^{\text{inner}}$ and $\hat{\mathcal{X}}_+^{\text{outer}}$ be defined as

$$\hat{\mathcal{X}}_+^{\text{inner}}=\{x\in\tilde{\mathcal{X}}|\hat{t}(x)>\hat{c}_{|\cdot|,\alpha}(\tilde{\mathcal{X}})\cdot\hat{\sigma}(x)\},\quad \hat{\mathcal{X}}_+^{\text{outer}}=\{x\in\tilde{\mathcal{X}}|\hat{t}(x)\ge -\hat{c}_{|\cdot|,\alpha}(\tilde{\mathcal{X}})\cdot\hat{\sigma}(x)\}. \tag{9}$$

Theorem 4 *If $\mathcal{X}_+\subseteq\tilde{\mathcal{X}}$, then the sets $\hat{\mathcal{X}}_+^{\text{inner}}$ and $\hat{\mathcal{X}}_+^{\text{outer}}$ defined in (9) satisfy the coverage condition (7).*

Proof On the event in (8), we have, for all $x\in\tilde{\mathcal{X}}$,

$$\hat{t}(x)-\hat{c}_{|\cdot|,\alpha}(\mathcal{X})\cdot\hat{\sigma}(x)\le t(x)\le\hat{t}(x)+\hat{c}_{|\cdot|,\alpha}(\mathcal{X})\cdot\hat{\sigma}(x).$$

For $x\in\mathcal{X}_+\cap\tilde{\mathcal{X}}$, $t(x)\ge 0$, so the second inequality implies that $0\le\hat{t}(x)+\hat{c}_{|\cdot|,\alpha}(\mathcal{X})\cdot\hat{\sigma}(x)$, which implies that $x\in\hat{\mathcal{X}}_+^{\text{outer}}$. For $x\in\hat{\mathcal{X}}_+^{\text{inner}}$, $\hat{t}(x)-\hat{c}_{|\cdot|,\alpha}(\mathcal{X})\cdot\hat{\sigma}(x)>0$, so the first inequality implies that $x\in\mathcal{X}_+$. □

From a multiple hypothesis testing standpoint, the sets $\hat{\mathcal{X}}_+^{\text{inner}}$ and $\tilde{X}\backslash\hat{\mathcal{X}}_+^{\text{outer}}$ can be considered sets where the null $H_0: t(x)=0$ has been rejected in favor of $t(x)>0$ or $t(x)<0$ respectively, while controlling the FWER. To ensure that (7) is satisfied, it is necessary to control not only the FWER for these null hypotheses, but also the probability that the decision $t(x)>0$ is made when in fact $t(x)<0$, or vice versa. Proving the control of these errors, called "directional" or "type III" errors in the multiple testing literature, can be an issue for stepwise procedures (see Shaffer, 1980; Finner, 1999).

While the single step procedure given above controls these error rates (thereby giving two-sided coverage as defined above), the multistep extension of this procedure requires further conditions. Indeed, the counterexample in Section 3 of Shaffer (1980) shows that the multistep extension of the above procedure will not satisfy (7) if $\tilde{X}$ has two elements and $\hat{t}(x)$ follows the Cauchy distribution independently over x. For the case where $\hat{t}(x)$ is asymptotically normal and independent over x, as in Sect. 3.1, Theorems 1 and 2 of Shaffer (1980) show that the directional error rate for the step-down procedure is controlled. However, results that would apply to the normal approximation used in Sect. 3.2 (which involves a sequence of Gaussian processes with complicated dependence structures that do not even settle down in the sense of weak convergence) are, to our knowledge, not available. The control of directional errors in such settings is an interesting topic for future research.

C Performance of $\hat{\mathcal{X}}_+$ under average welfare loss

This appendix considers the performance of $\hat{\mathcal{X}}_+$ under average welfare loss, when implemented as a statistical treatment rule. Consider the setup with discrete covariates with support $\{x_1, \ldots, x_\ell\}$. For a treatment rule assigning treatment to $\hat{\mathcal{X}}$, the regret risk under average welfare loss, considered by Manski (2004), is given by the difference between expected welfare under the optimal treatment and under $\hat{\mathcal{X}}$, and can be written in our setup as

$$R(\hat{\mathcal{X}}, P) \equiv \sum_{j=1}^{\ell} |t_P(x_j)| f_X(x_j) \left[P(x \in \mathcal{X}_{+,P} \backslash \hat{\mathcal{X}}) + P(x \in \hat{\mathcal{X}} \backslash \mathcal{X}_{+,P}) \right],$$

where we now index $t_P(\cdot)$ and $\mathcal{X}_{+,P}$ by P to denote the dependence on the underlying distribution explicitly (note that f_X depends on P as well, but we will consider classes of distributions where f_X is fixed).

Let P_n be a sequence of distributions such that $f_X(x) = P(X_i = x)$ does not change with n, and such that $\sqrt{n} t_{P_n}(x) \to t_\infty(x)$ for some t_∞.

Suppose that

$$\frac{\hat{t}(x_j) - t_{P_n}(x_j)}{\hat{\sigma}(x_j)}$$

converges in distribution, jointly over j, to a vector of independent standard normal variables, and that $\sqrt{n}\hat{\sigma}(x_j) \xrightarrow{p} s(x_j)$ for each j. Let $c_{\alpha,k} = c_\alpha(\mathcal{X})$ for $|\mathcal{X}| = k$.

Theorem 5 *Under the above assumptions with* $\alpha \leq 1/2$,

$$\limsup_n \sqrt{n} R(\hat{\mathcal{X}}_+, P_n) \leq \sup_{t>0} t \cdot \Phi\left(c_{\alpha,\ell} - \frac{t}{\max_x s(x)} \right).$$

Proof For any x with $t_\infty(x) > 0$, we will have $x \in \mathcal{X}_{+,P_n}$ under P_n for large enough n and

$$\begin{aligned}
\sqrt{n} t_{P_n}(x) P_n\left(x \notin \hat{\mathcal{X}}_+\right) &\leq \sqrt{n} t_{P_n}(x) P_n\left(\frac{\hat{t}(x)}{\hat{\sigma}(x)} \leq c_{\alpha,\ell} \right) \\
&= \sqrt{n} t_{P_n}(x) P_n\left(\frac{\hat{t}(x) - t_{P_n}(x)}{\hat{\sigma}(x)} \leq c_{\alpha,\ell} - \frac{t_{P_n}(x)}{\hat{\sigma}(x)} \right) \overset{n\to\infty}{\to} t_\infty(x) \Phi\left(c_{\alpha,\ell} - \frac{t_\infty(x)}{s(x)} \right).
\end{aligned}$$

For x with $t_\infty(x) < 0$, we have $x \notin \mathcal{X}_+$ under P_n for large enough n and

$$\begin{aligned}
\sqrt{n} |t_{P_n}(x)| P_n\left(x \in \hat{\mathcal{X}}_+\right) &\leq \sqrt{n} |t_{P_n}(x)| P_n(\hat{t}(x) \geq 0) \\
&= \sqrt{n} |t_{P_n}(x)| P_n\left(\frac{\hat{t}(x) - t_{P_n}(x)}{\hat{\sigma}(x)} \geq -\frac{t_{P_n}(x)}{\hat{\sigma}(x)} \right) \overset{n\to\infty}{\to} |t_\infty(x)| \Phi\left(-\frac{|t_\infty(x)|}{s(x)} \right),
\end{aligned}$$

where the first inequality uses the fact that $c_{n,k} \geq 0$ for all k. For $t_\infty(x) = 0$, we have $\sqrt{n} t_\infty(x) \to 0$. Thus,

$$\limsup_n \sqrt{n} R(\hat{\mathcal{X}}, P) \le \sum_{t_\infty(x) > 0} t_\infty(x) \Phi\left(c_{\alpha,\ell} - \frac{t_\infty(x)}{s(x)} \right) f_X(x) + \sum_{t_\infty(x) < 0} |t_\infty(x)| \Phi\left(- \frac{|t_\infty(x)|}{s(x)} \right) f_X(x) \le \sup_{t > 0} t \cdot \Phi\left(c_{\alpha,\ell} - \frac{t}{\max_x s(x)} \right).$$

□

We now give a lower bound specialized to the case where $t_\infty(x)$ and $s(x)$ are constant.

Theorem 6 *Suppose that the above assumptions hold with* $t_\infty(x) = C > 0$ *and* $s(x) = 1$ *for all* x. *Then, for any* $m \in \{1, \ldots, \ell\}$

$$\liminf_n \sqrt{n} R(\hat{\mathcal{X}}_+, P_n) \ge C \cdot \Phi\big(c_{\alpha,m} - C\big) \sum_{k=m-1}^{\ell-1} \binom{\ell-1}{k} \Phi\big(c_{\alpha,m} - C\big)^k \big[1 - \Phi\big(c_{\alpha,m} - C\big)\big]^{\ell-1-k}.$$

Proof For each x we have, for any $m \in \{1, \ldots, \ell\}$,

$$\begin{aligned} P_n\big(x \notin \hat{\mathcal{X}}_+\big) &\ge P_n\left(\frac{\hat{t}(x)}{\hat{\sigma}(x)} \le c_{\alpha,m} \text{ and } \left| \left\{ x' \in \{1, \ldots, \ell\} \backslash x \,\middle|\, \frac{\hat{t}(x')}{\hat{\sigma}(x')} \le c_{\alpha,m} \right\} \right| \ge m - 1 \right) \\ &\overset{n \to \infty}{\to} \Phi\left(c_{\alpha,m} - \frac{t_\infty(x)}{s(x)} \right) \cdot \sum_{\bar{\mathcal{X}} s.t. x \notin \bar{\mathcal{X}}, |\bar{\mathcal{X}}| \ge m-1} \prod_{x' \in \bar{\mathcal{X}}} \Phi\left(c_{\alpha,m} - \frac{t_\infty(x')}{s(x')} \right) \\ &\quad \cdot \prod_{x' \notin \bar{\mathcal{X}}} \left[1 - \Phi\left(c_{\alpha,m} - \frac{t_\infty(x')}{s(x')} \right) \right] \\ &= \Phi\big(c_{\alpha,m} - C\big) \sum_{k=m-1}^{\ell-1} \binom{\ell-1}{k} \Phi\big(c_{\alpha,m} - C\big)^k \big[1 - \Phi\big(c_{\alpha,m} - C\big)\big]^{\ell-1-k}. \end{aligned}$$

Thus,

$$\begin{aligned}\liminf_n \sqrt{n}R(\hat{\mathcal{X}}_+, P_n) &= \liminf_n \sum_{j=1}^{\ell} \sqrt{n} t_{P_n}(x_j) f_X(x_j) P\big(x_j \notin \hat{\mathcal{X}}_+\big) \\ &\geq \sum_{j=1}^{\ell} C f_X(x_j) \cdot \Phi\big(c_{\alpha,m} - C\big) \\ &\quad \sum_{k=m-1}^{\ell-1} \binom{\ell-1}{k} \Phi\big(c_{\alpha,m} - C\big)^k \big[1 - \Phi\big(c_{\alpha,m} - C\big)\big]^{\ell-1-k} \\ &= C \cdot \Phi\big(c_{\alpha,m} - C\big) \\ &\quad \sum_{k=m-1}^{\ell-1} \binom{\ell-1}{k} \Phi\big(c_{\alpha,m} - C\big)^k \big[1 - \Phi\big(c_{\alpha,m} - C\big)\big]^{\ell-1-k}\end{aligned}$$

□

The comparison between $\hat{\mathcal{X}}_+$ as a treatment rule and other statistical treatment rules will depend on how one compares risk functions. Consider the conditional empirical success (CES) rule, defined by the set $\hat{\mathcal{X}}_{\text{CES}} = \{x | \hat{t}(x) > 0\}$ (see Manski, 2004). For $\alpha \leq 1/2$, $\hat{\mathcal{X}}_+$ is contained in $\hat{\mathcal{X}}_{\text{CES}}$ with probability one, so the risk function will always be smaller for $\hat{\mathcal{X}}_+$ when $t(x) \leq 0$ for all x. Thus, if one chooses a criterion that puts a large enough amount of weight on cases with negative treatment effects, such as Bayes risk where the prior puts enough of the mass on negative treatment effects, $\hat{\mathcal{X}}_+$ will be preferred to $\hat{\mathcal{X}}_{\text{CES}}$. However, the situation will be reversed in cases where less weight is given to negative effects.

Consider the minimax regret criterion, which takes the supremum of $R(\cdot, P)$ over an given class of distributions $P \in \mathcal{P}$. Theorem 6 gives an asymptotic lower bound on minimax regret for any class of distributions that contains a sequence P_n satisfying the conditions of that theorem. Theorem 5 essentially gives an upper bound on minimax regret, although additional technical conditions would be needed to ensure that the lim sup is uniform over sequences P_n under consideration.

We now consider these bounds and how they relate to the number of values ℓ taken by the covariate. Consider the case where $s(x)$ is constant for all x, and let us make the normalization $s(x) = 1$. For the CES rule and other rules that assign treatment to x based only on observations with $X_i = x$, the minimax regret will not increase as ℓ increases (the normalization $s(x) = 1$ means that, as ℓ increases, the variance of Y_i decreases so that the difficulty of the estimation problem for each x stays the same despite each x having fewer observations). However, the minimax regret for $\hat{\mathcal{X}}_+$ will increase without bound, as we now show.

First, consider the lower bound. Let $C = C_n = c_{\alpha,\lfloor \ell/2 \rfloor}$, where $\lfloor s \rfloor$ denotes the greatest integer less than or equal to s. Then, applying the bound with $m = \lfloor \ell/2 \rfloor$, we obtain a lower bound of

$$c_{\alpha,\lfloor \ell/2 \rfloor}(1/2) \sum_{k=\lfloor \ell/2 \rfloor - 1}^{\ell-1} \binom{\ell-1}{k} (1/2)^k (1/2)^{\ell-1-k} = c_{\alpha,\lfloor \ell/2 \rfloor}(1/2) \sum_{k=\lfloor \ell/2 \rfloor - 1}^{\ell-1} \binom{\ell-1}{k} (1/2)^{\ell-1}.$$

The last term is the probability of a binom$(1/2, \ell-1)$ being at least $\lfloor \ell/2 \rfloor - 1$, which converges to 1/2 as $\ell \to \infty$. Since $c_{\alpha,\lfloor \ell/2 \rfloor}/\sqrt{2\log \ell}$ converges to one as $\ell \to \infty$, it follows that the asymptotic minimax regret is bounded from below by $\sqrt{2\log \ell}/4$ times a sequence that converges to one as ℓ increases.

For the upper bound, note that, using the fact that $\Phi(-s) \leq \phi(s)/s$ for $s > 0$, where $\phi(s)$ is the standard normal pdf, we have, for $t > c_{\alpha,\ell}$, letting $s = t - c_{\alpha,\ell}$

$$t\Phi(c_{\alpha,\ell} - t) = (s + c_{\alpha,\ell})\Phi(-s) \leq c_{\alpha,\ell} + \phi(s) \leq c_{\alpha,\ell} + \phi(0).$$

For $t \leq c_{\alpha,\ell}$, $t\Phi(c_{\alpha,\ell} - t)$ is clearly bounded by $c_{\alpha,\ell}$, so the right hand side of the above display gives a bound for $\sup_{t \geq 0} t\Phi(c_{\alpha,\ell} - t)$. Since $(c_{\alpha,\ell} + \phi(0))/\sqrt{2\log \ell}$ converges to one as ℓ increases, this gives an approximate upper bound of $\sqrt{2\log \ell}$ for large ℓ.

Thus, the minimax regret (under average welfare loss) for implementing $\hat{\mathcal{X}}_+$ as a statistical treatment rule increases with ℓ at a $\sqrt{\log \ell}$ rate. This reflects the fact that, due to its incorporation of multiple hypothesis tests, this rule becomes increasingly conservative with ℓ. Since the average welfare loss function along with minimax regret leads to a symmetric treatment of "overestimation" and "underestimation" of $\mathcal{X}_+$, this increasing conservativeness leads to worse behavior for large ℓ. On the other hand, $\sqrt{\log \ell}$ increases slowly with ℓ, so the increase in minimax regret as ℓ increases is not too large.

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

Data availability Replication codes for the simulation and empirical sections are provided on the authors' websites.

References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian. *Perry Preschool, and Early Training Projects, Journal of the American Statistical Association, 103*, 1481–1495.

Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica, 89*, 133–161. https://doi.org/10.3982/ECTA15732

Bhattacharya, D., & Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics, 167*, 168–196.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics, 90*, 414–427.

Chamberlain, G. (2011). Bayesian Aspects of Treatment Choice. *Oxford Handbook of Bayesian Econometrics,* 11–39.

Chernozhukov, V., Hong, H., & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica, 75*, 1243–1284.

Chernozhukov, V., Lee, S., & Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica, 81*, 667–737.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., Yagan, D. (2010). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR, Tech. rep., National Bureau of Economic Research.

Claeskens, G. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics, 31*, 1852–1884.

Crump, Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics, 90*(3), 389–405.
Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics, 125*(1), 141–173.
Fan, J., & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC.
Fan, Y., & Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory, 26*, 931–951.
Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *The Annals of Statistics, 27*, 274–289.
Hahn, J., Hirano, K., & Karlan, D. (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics, 29*, 96–108.
Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software, 27*(5), 1–32.
Hirano, K., & Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica, 77*, 1683–1701.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.
Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics, 125*, 241–270.
Kitagawa, T., & Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica, 86*, 591–616.
Kong, E., Linton, O., & Xia, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory, 26*, 1529–1564.
Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*(2), 497–532.
Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *The Economic Journal, 111*(468), 1–28.
Kwon, K. (2022). Essays in Inference for Nonparametric Regression Models. Yale Graduate School of Arts and Sciences Dissertations.
Lazear, E. P. (2001). Educational production. *Quarterly Journal of Economics, 116*(3), 777–803.
Lee, S., & Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of PROGRESA on school enrollment. *Journal of Applied Econometrics, 29*(4), 612–626.
Luedtke, A. R., & Laan, M. J. V. U. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics, 44*, 713–742.
Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica, 72*, 1221–1246.
Manski, C. F., & Tetenov, A. (2016). Sufficient trial size to inform clinical practice. *Proceedings of the National Academy of Sciences, 113*, 10518–10523.
Manski, C. F., & Tetenov, A. (2019). Trial size for near-optimal choice between surveillance and aggressive treatment: Reconsidering MSLT-II. *The American Statistician, 73*, 305–311.
Mbakop, E., & Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica, 89*, 825–848.
Neumann, M. H., & Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics, 9*, 307–333.
Romano, J. P., & Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica, 78*, 169–211.
Romano, J. P., & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association, 100*, 94–108.
Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *The Annals of Statistics, 8*, 1342–1347.
Stoye. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics, 151*, 70–81.
Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics, 166*, 157–165.
Whitmore, D. (2005). Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *American Economic Review, 95*(2), 199–203.