INFERENCE ON OPTIMAL TREATMENT ASSIGNMENTS

By

Timothy B. Armstrong and Shu Shen

November 2013
Revised April 2014

COWLES FOUNDATION DISCUSSION PAPER NO. 1927R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Inference on Optimal Treatment Assignments*

Timothy B. Armstrong — Yale University

Shu Shen — University of California, Davis

April 23, 2014

Abstract

We consider inference on optimal treatment assignments. Our methods allow for 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 the Mexican conditional cash transfer program Progresa.

1 Introduction

In recent decades, there has been increasing recognition both in academic and public circles that social experiments or social programs, costly as they are, should be rigorously evaluated to learn lessons from past experience and to better guide future policy decisions. While a

*We thank Colin Cameron, Azeem Shaikh and participants at the 2012 California Econometrics Conference for helpful comments. All remaining errors are our own.

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; Tetenov, 2012; Bhattacharya and Dupas, 2012), the problem of constructing confidence statements for the optimal decision rule has received little attention. The goal of this paper is to formulate and answer the problem of constructing confidence statements that quantify the statistical precision of a treatment assignment rule. This allows researchers to quantify how strong the evidence is in favor of treating certain individuals. This type of analysis is especially useful for policy makers who are interested in designing an efficient large-scale social program following a social experiment or an initial small-scale trial program.

To understand the importance of quantifying the statistical precision of treatment recommendations, consider the case where a policy maker wants to design a social program that gives some selected individuals a treatment intervention (say, school attendance subsidies). The effect of the treatment on the response outcome (say, attendance) is expected to be heterogeneous and varies along certain observed variables (say, distance from the nearest school). A natural goal of the policy maker is to assign treatment only to those individuals whose treatment effect is 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 is 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 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 describing how strong the evidence is in favor of treating the selected individuals. Obviously, a large scale experimental intervention with many observations would provide more compelling evidence for or against treatment than an otherwise identical experiment with fewer observations. Quantifying this requires statements about statistical precision of the treatment decision rule; this is the question that we formulate and answer in this paper.

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 to be treated by testing multiple hypotheses that the conditional average treatment effect is positive 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 social program studies. In this paper we apply the inference method to the Mexican conditional cash transfer program Progresa. Progresa gives cash to households in poor villages if their children attend school regularly. The program is often criticized by the literature for its "leakage of benefit", or transferring cash to households who would send children to school anyway. Therefore, it is very important to statistically quantify the evidence in favor of Progresa treatments. Using data collected from the experimental phase of Progresa, we demonstrate how the proposed inference procedure can be used to distinguish households for which the cash transfer increases children school attendance and statistically quantify the confidence level associated with treating these selected households.

The problem of optimal treatment assignment[1] has been considered by Manski (2004),

[1] 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, Hirano, and Karlan, 2011; Kasy, 2013). In contrast to this literature, which con-

Dehejia (2005), Hirano and Porter (2009), Stoye (2009), Tetenov (2012), Bhattacharya and Dupas (2012), and others. In this literature, individuals are assigned to different treatments by a social planner who maximizes social welfare or minimizes the risk associated with different treatment assignment rules.

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 rule, 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 a measure of statistical precision, 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.

While we are not aware of other papers that consider inference on the treatment assignment rule that would be optimal in the population, 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 (2013). Those papers use finite sample randomization tests to construct subsets of a discrete conditioning

siders the design phase of the experiment, we take data from the initial experiment as given and focus on implications for future policy.

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, Hong, and Tamer (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, Hong, and Tamer (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, Lee, and Rosen, 2011). In particular, we use results from the latter paper, which those authors use in a different class of hypothesis testing problems.[2]

Our proposed inference procedure on optimal treatment assignments is also related to the test for treatment effect heterogeneity considered by Crump, Hotz, Imbens, and Mitnik

[2] Chernozhukov, Lee, and Rosen (2011) are interested in testing a single null hypothesis involving many values of the covariate, while our formulation leads us to the multiple hypothesis testing problem of determining which values of the covariates lead to rejection; the stepdown improvement gains precision in our context, but would be irrelevant in their case.

(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 shows that our proposed inference procedure have good size and power properties in small sample. 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 statistical precision. The size and power properties of the proposed inference procedure is 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 the 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 the Progresa example. Section 6 concludes.

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 shall 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.[3]

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 pretreatment observables that takes on values in a set $\tilde{\mathcal{X}}$. The data are assumed to take the following

[3] This allows costs to be incorporated by being subtracted from the treatment, and budget constraints can be incorporated by estimating a shadow cost (see Bhattacharya and Dupas, 2012). The only major restriction here is that outcomes are considered individually, so peer effects are ruled out.

unconfoundedness assumption.

Assumption 1.

$$E(Y_i(j)|D_i = j, X_i = x) = E(Y_i(j)|X_i = x), \qquad j = 1, 2.$$

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 average potential outcomes for both the treatment and the control group in the sample give the same average outcomes as if both 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 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, Imbens, and Mortimer, 2005). In general, one needs to assume that the conditional average treatment effect is the same for whatever new population is being considered for treatment in order for 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 and therefore leads to treatment. Given a prespecified significance level α, we seek a set $\hat{\mathcal{X}}_+$ that satisfies

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

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 subpopulation 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.[4] 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 infinite number of null hypotheses $H_x : t(x) \leq 0$ for all $x \in \tilde{\mathcal{X}}$. 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, $\hat{\mathcal{X}}_+$ in fact 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 $\hat{\mathcal{X}}_+$.

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 $\hat{t}(x)$. For any set $\mathcal{X}$ on the support of X_i, let the critical value

[4] In any case, the role of $Y_i(0)$ and $Y_i(1)$ can be reversed to apply this framework to obtain a set that contains $\mathcal{X}_+$ with $1-\alpha$ probability.

$\hat{c}_{u,\alpha}(\mathcal{X})$ satisfy

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

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) or Chernozhukov, Lee, and Rosen (2011) as we describe later. For some of the results, we will require that these critical values be nondecreasing in $\mathcal{X}$ in the sense that

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

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

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

Clearly $\hat{\mathcal{X}}_+^1$ satisfies (1), since the event in (2) implies the event in (1). However, we can make improvement on inference using a step-down procedure (see Holm, 1979; Romano and Wolf, 2005). That is, we could find some 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

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

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

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

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 (2013) 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, $\tilde{\mathcal{X}}$ is discrete and takes on a finite number of values. We write

$$\tilde{\mathcal{X}} = \{x_1, \ldots, x_\ell\}. \tag{4}$$

In this setting, we may estimate the treatment effect $\hat{t}(x)$ with the sample analogue. 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

$$\hat{t}(x_j) = \frac{1}{N_{1,x_j}} \sum_{1 \le i \le n, D_i = 1, X_i = x_j} Y_i - \frac{1}{N_{0,x_j}} \sum_{1 \le i \le n, D_i = 0, X_i = x_j} Y_i \tag{5}$$

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, $\{(\hat{t}(x_j) - t(x_j))/\hat{\sigma}(x_j)\}_{j=1}^{\ell}$ converges in distribution to an ℓ dimensional joint normal random variable. 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

$$\hat{c}_{u,\alpha}(\mathcal{X}) = \Phi^{-1}\left((1-\alpha)^{1/|\mathcal{X}|}\right) \tag{6}$$

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

$$\hat{c}_{u,\alpha}(\mathcal{X}) = \Phi^{-1}\left(1 - \alpha/|\mathcal{X}|\right). \tag{7}$$

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 iid 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 the conditional exogeneity assumption 1 holds. Then the critical values defined in (6) and (7) 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, Lee, and Rosen, 2011). 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

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

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, Lee, and Rosen (2011) to our case.

Assumption 2. 1. *The observations $\{(X_i, D_i, Y_i)\}_{i=1}^n$ are iid 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 $\int uK(u)\,du = 0$ and $\int K(u)\,du = 1$.*

7. *The bandwidth for the untreated group, h_0, satisfies the following asymptotic relations as $n \to \infty$: $nh_0^{d+2} \to \infty$ and $nh_0^{d+4} \to 0$ at polynomial rates. In addition, the same conditions hold for the bandwidth h_1 for the treated group.*

To approximate the supremum of this distribution over a nondegenerate set, we follow Neumann and Polzehl (1998) and Chernozhukov, Lee, and Rosen (2011) 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 iid standard normal variables drawn independently of the data. To form critical values $\hat{c}_{u,\alpha}(\mathcal{X})$, we simulate S replications of n iid 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

$$\sup_{x\in\mathcal{X}} \frac{\hat{t}^*(x)}{\hat{\sigma}(x)} = \sup_{x\in\mathcal{X}} \frac{\hat{m}_1^*(x) - \hat{m}_0^*(x)}{\hat{\sigma}(x)}. \tag{8}$$

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.

Theorem 3. *Under Assumptions 1 and 2, the multiplier bootstrap critical value $\hat{c}_{u,\alpha}(\mathcal{X})$ defined above satisfies (2) and (3).*

Proof. The critical value satisfies (2) by the arguments in Example 7 of Chernozhukov, Lee, and Rosen (2011) (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 $(\hat{m}_1(x) - \hat{m}_0(x))/\hat{\sigma}(x))$. Condition (3) is satisfied by construction. □

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 an estimated set that satisfies

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

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.

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 to determine for which values of the conditioning covariate X_i the conditional average treatment effect differs from its average over the entire population. The estimation of $\hat{\mathcal{X}}_{+-}$ is a simple extension to the estimation of

$\hat{\mathcal{X}}_+$ described in the previous section. Let $\hat{c}_{|u|,\alpha}(\mathcal{X})$ be a critical value that satisfies

$$\liminf_n P\left(\sup_{x\in\mathcal{X}}\left|\frac{\hat{t}(x)-\hat{t}-(t(x)-t)}{\hat{\sigma}(x)}\right|\leq \hat{c}_{|u|,\alpha}(\mathcal{X})\right)\geq 1-\alpha,$$

where $\hat{t}=\frac{1}{n}\sum_{i=1}^n\hat{t}(X_i)$ is a $\sqrt{n}$-consistent estimator of t. Let $\hat{\mathcal{X}}_{+-}^1\equiv\{x\in\tilde{\mathcal{X}}||(\hat{t}(x)-\hat{t})/\hat{\sigma}(x)|>\hat{c}_{|u|,\alpha}(\tilde{\mathcal{X}})\}$. For $k>1$, let $\hat{\mathcal{X}}_{+-}^k=\{x||(\hat{t}(x)-\hat{t})/\hat{\sigma}(x)|>\hat{c}_{|u|,\alpha}(\tilde{\mathcal{X}}\backslash\hat{\mathcal{X}}_{+-}^{k-1})\}$. 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, $\hat{c}_{|u|,\alpha}(\mathcal{X})$ could be set as the $1-\alpha$ quantile of the empirical distribution of the bootstrap test statistic $\sup_{x\in\mathcal{X}}|\frac{\hat{t}^*(x)-t^*}{\hat{\sigma}(x)}|$, where $\hat{t}^*(x)$ is the multiplier process defined earlier and $\hat{t}^*=\frac{1}{n}\sum_{i=1}^n\hat{t}^*(X_i)$.

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.

DGP1: $X_i\sim U(0,1)$, $e_i\sim N(0,1/9)$, $v_i\sim N(0,1)$, $D_i=1(0.1X_i+v_i>0.55)$, $Y_i=10(X_i-1/2)^2 1(D_i=1)+e_i$;

DGP2: $X_i\sim U(0,1)$, $e_i\sim N(0,1/9)$, $v_i\sim N(0,1)$, $D_i=1(0.1X_i+v_i>0.55)$, $Y_i=\sin(10X_i+1)1(D_i=1)+e_i$;

DGP3: $X_i\sim U(0,1)$, $e_i\sim N(0,1/9)$, $v_i\sim N(0,1)$, $D_i=1(0.1X_i+v_i>0.55)$, $Y_i=(X_i-1/2)^2+e_i$.

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 $t(x)=10(x-1/2)^2$ always lies above zero except for one tangent point. In the second DGP $t(x)=\sin(10x+1)$ is positive in some parts of the X_i support and negative in the other parts. $t(x)$ is uniformly zero in the third DGP.

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). For each of the repeated simulations of each DGP and sample size, the local linear estimator $\hat{t}(x)$ is evaluated at 500 equally spaced grids on the support of X, or $[0, 1]$. The supremum of the studentized $\hat{t}(x)$ is equal to the maximum of the 500 estimates. The critical value is dependent on the sample distribution and is 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 Figure 1 reports the true CATEs (in black) and the local linear estimates (in blue) of the CATEs based on one randomly simulated sample of size 500. The right panel reports studentized CATE estimates (in dark red), the true optimal treatment set $\mathcal{X}_+$ (in black) and the proposed inference region $\hat{\mathcal{X}}_+$ (in blue) 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 smallest step-down critical value (shown by the lowest blue horizontal line). The step-down critical values are different in each step until convergence because, as is discussed in the theoretical section, the sets of x values used to calculate the supremum of the CATE estimates are different. On the graphs, these different sets are shown by the region covered by the blue horizontal line. The total number of steps taken in critical value calculation is reported in the subtitle of each graph in the right panel.

The confidence region $\hat{\mathcal{X}}_+$ for the optimal treatment set controls familywise error rates properly. As a comparison, the right panel of Figure 1 also reports treatment sets (in red) 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 local linear estimator works reasonably well. As is expected, the proposed confidence regions are always smaller than the pointwise treatment sets. That is because the latter actually does not control the error rate correctly. The figure for DGP3 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 DGP1 and DGP2. As is noted in the figure subtitle, the total number of steps for critical value calculation is 4 for DGP1 and 3 for DGP2. The step-down refinement does not lead to improvement for DGP3 because the initial confidence region is a null set.

Although the simulation that makes Figure 1 is specially selected for illustration purposes, the good performance of the proposed inference procedure holds throughout all simulations. Columns (3)-(6) and (9)-(12) in Table 1 report the size and power of the proposed treatment set $\hat{\mathcal{X}}_+$ 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 the treatment set $\hat{\mathcal{X}}_+^1$ ($\hat{\mathcal{X}}_+$) is not included in the true set $\mathcal{X}_+$. The power is measured by the average proportion of false hypothesis 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 stepdown method. The power measure is denoted as FHR and FHR-SD for the stepdown method. We see from results reported in these columns that the proposed confidence region for the optimal treatment set controls familywise error rates well. In the case of DGP3 where the least favorable condition of the multiple hypothesis testing holds and the conditional average treatment effect equals to zero uniformly, the empirical familywise error rates are very close to the nominal familywise error rate. Comparing results 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 critical value calculation. The increment in power is larger when the sample size is smaller. In our empirical section below, we show an example where applying the step-down refinement method substantially improves the inference of optimal treatment assignment.

For comparison purposes, we also report in Table 1 the size and power properties of treatment sets obtained from pointwise confidence intervals, or all x values that reject the pointwise null hypothesis that $t(x)$ is negative. Comparing results in columns (1)-(2) and (7)-(8) to their uniform counterparts, we see that the pointwise treatment sets, as expected, fail to control the familywise error rate at all. In the case of DGP3, where the true average treatment effect is zero for all x values, more than 39% (57%) of the time the pointwise set estimator discover some "fake" nonempty positive treatment set when the significance level 5% (10%) is used. The probability of reporting a "fake" treatment set does not decrease with the increase of sample size.

5 Optimal Treatment Assignment for Progresa

In this section, we demonstrate how the proposed inference procedure for optimal treatment assignment can be used to study the treatment assignment of real world social programs using a dataset collected from the Mexican welfare program Progresa (now named Oportunidades). Progresa is a conditional cash transfer program that provides cash transfers to households in poor rural and semi-urban localities (villages) conditional on the regular school attendance of their children, family visits to health centers and women's participation in health and nutrition workshops. Progresa collected high quality data in its experiment phase and is widely studied in the literature (c.f. Schultz, 2004 and Attanasio, Meghir, and Santiago, 2011). Like other conditional cash transfer programs, Progresa is often criticized by the literature for its "leakage of benefit", meaning that many cash transfers were given to households who would have met the conditions of the program regardless of the transfer. Therefore, it is very important to utilize the available data to statistically quantify the evidence in favor of Progresa treatments. In this section, we apply the proposed inference procedure for this purpose.

The experimental phase of Progresa was conducted across seven states in 506 poor localities, which were randomly assigned into treated and controlled groups with probabilities 2/3 and 1/3. The impact of the cash transfer can be evaluated by comparing the average outcome between the treatment and control localities. After the experimental phase, Progresa has been continually expanding. In 2012, Progresa had about 5.8 million recipient households in more than 187,000 localities. The dataset we use for analysis focuses on the educational program of Progresa and is originally from Attanasio, Meghir, and Santiago (2011). We look at boys 10-14 years old. Table 2 reports some summary statistics of the dataset. We see that although schooling is compulsory in Mexico through high school (preparatoria), it is common for children not to attend school regularly in rural areas of Mexico during the late '90s. The attendance rate drops sharply when children reach 12 years old and are ready to enter secondary school. It has been documented in the literature that the school attendance rate for children at secondary school age is negatively correlated to households' distance to closest secondary school. As a result, we use information on households' log distance to secondary school in 1998 and children's age to estimate the average treatment effect of Progresa on school attendance. We then quantify the statistical evidence of treating different individual households.

In the upper panel of Figure 2 we estimate the average school attendance in both treatment and control villages in October 1998. We split the sample using the discrete covariate age. Then we estimate average school attendance conditional on log distance by local linear estimation with Epanechnikov kernel and the rule-of-thumb bandwidth described in Section 4. We see from the graphs that the school attendance rates drop with both age and households' distance to closest secondary school. The graph shows that the leakage of benefit is more of an issue for households with younger children (for example, households with 10-year-olds) and households living closer to secondary school (for example, households living within 2-km distance to closest secondary school).

The lower panel of Figure 2 reports the studentized conditional average treatment effect estimate as well as the proposed confidence region for optimal treatment set. First, notice that the shape of the studentized CATE estimates in the bottom panel is different from the vertical difference between the treatment and control estimates in the upper panel. This is because a large CATE estimate for households very far away from nearest secondary school may be mitigated by its large standard error, since the log distance is still right-skewed in the dataset and has a small density for really large values. Second, since households living in the same locality may not be independent, we modify the inference of optimal treatment assignment described in Section 3 to account for data clustering. Use $i = 1, 2, .., N$ to denote individuals and $j = 1, 2, ..., J$ localities in Progresa. To account for potential within-locality dependence, we substitute the multiplier processes used in (8) 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. The critical value is then taken to be the $1 - \alpha$ quantile of the empirical distribution of the supremum estimator described in (8).[5]

Our proposed confidence region for optimal treatment set is outlined in blue in the lower panel of Figure 2. With probability larger than 95%, every child included in the confidence region is expected to increase regular school attendance as a result of the transfer. From the figure, we see that there is insufficient evidence in the data for treating the 10-year-olds. The same holds for 11-14 year olds in households living close to secondary school. Households who live very far away from a secondary school are sometimes also excluded from the confidence region due to the lack of statistical precision at the right tail of the log distance distribution where data is sparse. This problem of lack of inference precision for underrepresented populations could be solved if these groups were given a higher sampling

[5] This modified version with multiplier η_j that is fixed within a cluster can be viewed as corresponding to the wild cluster bootstrap discussed in Cameron, Gelbach, and Miller (2008) in a parametric context, extended to the local linear nonparametric estimator used here, and with a different multiplier weight (the terms "wild bootstrap" and "multiplier bootstrap" appear to be used interchangeably in the literature). We conjecture that, as with other settings with nonparametric smoothing, accounting for dependence is not technically necessary under conventional asymptotics, but will lead to substantial finite sample improvement.

weights in the experimental phase of a social program.

Strictly speaking, both the true (unknown) optimal treatment set and its confidence region plotted in Figure 2 are two-dimensional. The confidence region is the joint of the blue intervals in all five bottom graphs. Therefore the step-down refined critical values are the same in each of the five graphs. If, instead, policy makers would like to control the error rate of treatment assignment separately for children of different age cohorts, the confidence region would be single-dimensional and the step-down critical values would vary by age group. In Figure 3 we report such single-dimensional confidence regions. We notice that the blue intervals in each graphs of Figure 3 are wider than their counterparts in Figure 2. Also, interestingly, the step-down refinement for critical value calculation improves the inference significantly for the group of 12-year-olds.

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 preponderence of empirical evidence in favor of treatment.

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.

Attanasio, O. P., C. Meghir, and A. Santiago (2011): "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," *Review of Economic Studies*, 79, 37–66.

Bhattacharya, D., and P. Dupas (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167(1), 168–196.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 90(3), 414–427.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2011): "Intersection bounds: estimation and inference," *Arxiv preprint arXiv:0907.3503*.

Claeskens, G. (2003): "Bootstrap confidence bands for regression curves and their derivatives," *The Annals of Statistics*, 31(6), 1852–1884.

Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik (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(12), 141–173.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications.* Chapman and Hall/CRC.

Hahn, J., K. Hirano, and D. Karlan (2011): "Adaptive Experimental Design Using the Propensity Score," *Journal of Business & Economic Statistics*, 29, 96–108.

Hayfield, T., and J. S. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 5.

Hirano, K., and J. R. Porter (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77(5), 1683–1701.

Holm, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6(2), 65–70.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005): "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 125, 241–270.

Kasy, M. (2013): "Why Experimenters Should Not Randomize, and What They Should Do Instead," working paper.

Lee, S., and A. M. Shaikh (2013): "Multiple Testing and Heterogeneous Treatment Effects: Re-evaluating the Effect of PROGRESA on School Enrollment," *Journal of Applied Econometrics.*

Manski, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72(4), 1221–1246.

Neumann, M. H., and J. Polzehl (1998): "Simultaneous bootstrap confidence bands in nonparametric regression," *Journal of Nonparametric Statistics*, 9(4), 307–333.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

Romano, J. P., and M. Wolf (2005): "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100(469), 94–108.

Schultz, T. P. (2004): "School subsidies for the poor: evaluating the Mexican Progresa poverty program," *Journal of Development Economics*, 74, 199–250.

Stoye, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151(1), 70–81.

Tetenov, A. (2012): "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 166(1), 157–165.

Table 1: Size and Power Properties of Treatment Sets

	PW, $\alpha = 0.05$		Uniform, $\alpha = 0.05$				PW, $\alpha = 0.1$		Uniform, $\alpha = 0.1$			
	EFER	FHR	EFER	FHR	EFER-SD	FHR-SD	EFER	FHR	EFER	FHR	EFER-SD	FHR-SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
DGP1: $t(x) \geq 0$ for all $x \in \tilde{\mathcal{X}}$												
N=250	0*	0.9763	0*	0.9150	0*	0.9688	0*	0.9863	0*	0.9431	0*	0.9824
N=500	0*	0.9798	0*	0.9289	0*	0.9747	0*	0.9899	0*	0.9520	0*	0.9876
N=1000	0*	0.9854	0*	0.9391	0*	0.9824	0*	0.9921	0*	0.9590	0*	0.9904
DGP2: $t(x) \geq 0$ only for some $x \in \tilde{\mathcal{X}}$												
N=250	0.1120	0.8974	0.0000	0.6630	0.0000	0.7417	0.2280	0.9180	0.0000	0.7159	0.0080	0.7859
N=500	0.1180	0.9330	0.0000	0.7681	0.0000	0.8174	0.2400	0.9423	0.0000	0.8042	0.0080	0.8500
N=1000	0.1140	0.9547	0.0000	0.8418	0.0000	0.8740	0.1900	0.9637	0.0000	0.8662	0.0040	0.8966
DGP3: $t(x) = 0$ for all $x \in \tilde{\mathcal{X}}$												
N=250	0.3960	/$^{\#}$	0.0740	/$^{\#}$	0.0740	/$^{\#}$	0.5720	/$^{\#}$	0.1400	/$^{\#}$	0.1400	/$^{\#}$
N=500	0.4400	/$^{\#}$	0.0620	/$^{\#}$	0.0620	/$^{\#}$	0.6100	/$^{\#}$	0.1200	/$^{\#}$	0.1200	/$^{\#}$
N=1000	0.4640	/$^{\#}$	0.0520	/$^{\#}$	0.0520	/$^{\#}$	0.6800	/$^{\#}$	0.1000	/$^{\#}$	0.1000	/$^{\#}$

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 rejected is not defined in DGP 3 since the set where the null hypothesis is false has by construction measure zero.

Table 2: Summary Statistics

Age	Sample Size		Attendance Rate		Years of Education	
	Control	Treatment	Control	Treatment	Control	Treatment
10	644	1121	0.95 (0.21)	0.96 (0.20)	3.12 (1.04)	3.26 (0.97)
11	652	1096	0.94 (0.24)	0.95 (0.21)	3.98 (1.20)	4.04 (1.23)
12	700	1192	0.83 (0.37)	0.89 (0.31)	4.83 (1.41)	4.84 (1.43)
13	658	1024	0.78 (0.42)	0.83 (0.38)	5.44 (1.56)	5.58 (1.57)
14	673	1065	0.61 (0.49)	0.73 (0.44)	6.07 (1.71)	6.22 (1.73)

Note: Standard deviations are reported in the parentheses.

Figure 1: CATE Estimates, Critical Values, and Treatment Sets

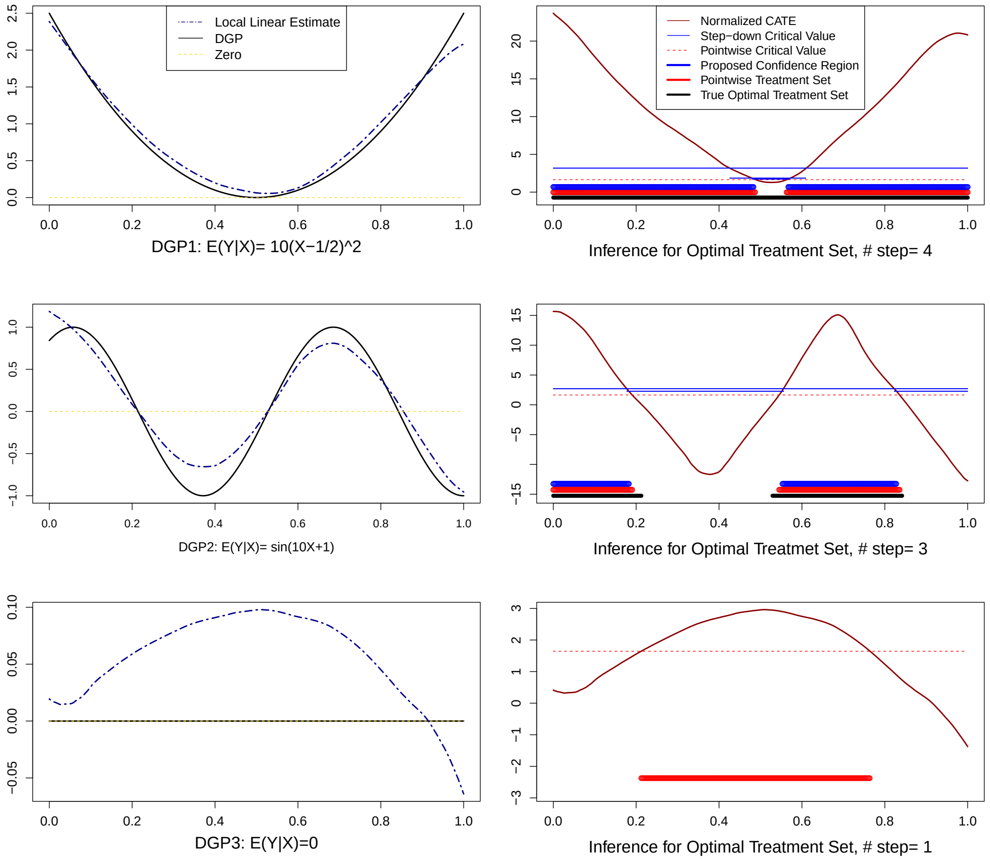
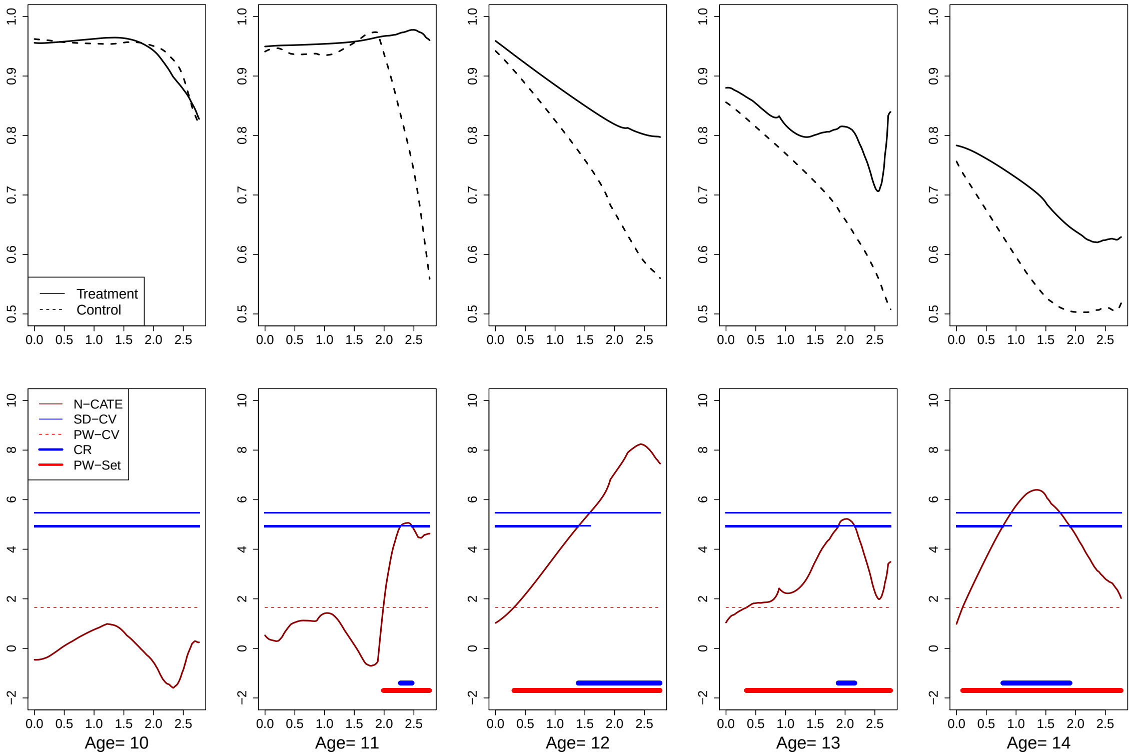
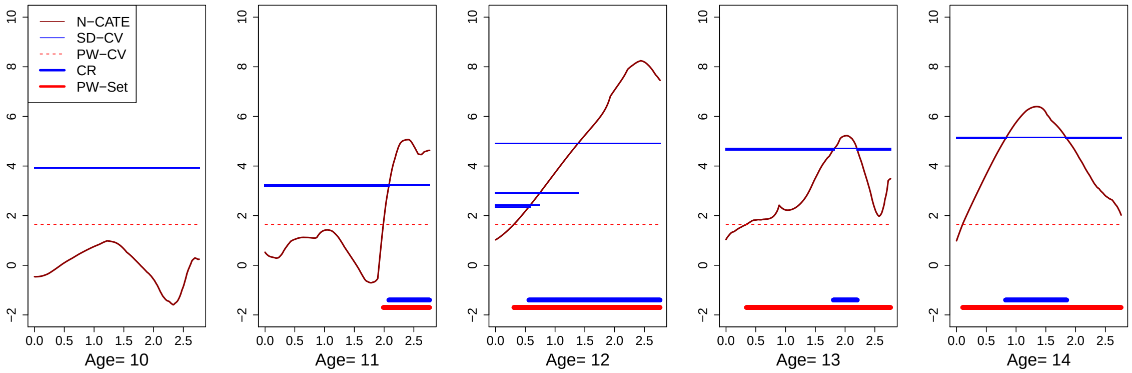


Figure 2: 2-Dimensional Confidence Region of the Optimal Treatment Set



Note: N-CATE denotes the conditional average treatment effect estimates normalized by standard errors; SD-CV are critical values following the step-down procedure; PW-CV is the pointwise critical value 1.645 for 5% one-sided tests; CR is the proposed confidence region for optimal treatment set; PW-Set is the treatment set calculated from the pointwise testing problem.

Figure 3: 1-Dimensional Confidence Region of Optimal Treatment Sets



Note: N-CATE denotes the conditional average treatment effect estimates normalized by standard errors; SD-CV are critical values following the step-down procedure; PW-CV is the pointwise critical value 1.645 for 5% one-sided tests; CR is the proposed confidence region for optimal treatment set; PW-Set is the treatment set calculated from the pointwise testing problem.