Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Estimating Conditional Average Treatment Effects

Jason Abrevaya, Yu-Chin Hsu & Robert P. Lieli

To cite this article: Jason Abrevaya, Yu-Chin Hsu & Robert P. Lieli (2015) Estimating Conditional Average Treatment Effects, Journal of Business & Economic Statistics, 33:4, 485-505, DOI: 10.1080/07350015.2014.975555

To link to this article: https://doi.org/10.1080/07350015.2014.975555

Published online: 27 Oct 2015.

Estimating Conditional Average Treatment Effects

Jason ABREVAYA
Department of Economics, University of Texas, Austin, TX 78712 (*abrevaya@eco.utexas.edu*)

Yu-Chin HSU
Institute of Economics, Academia Sinica, Taipei 115, Taiwan (*ychsu@econ.sinica.edu.tw*)

Robert P. LIELI
Department of Economics, Central European University, H-1051 Budapest, Hungary and Magyar Nemzeti Bank, 1850 Budapest, Hungary (*lielir@ceu.hu*)

We consider a functional parameter called the conditional average treatment effect (CATE), designed to capture the heterogeneity of a treatment effect across subpopulations when the unconfoundedness assumption applies. In contrast to quantile regressions, the subpopulations of interest are defined in terms of the possible values of a set of continuous covariates rather than the quantiles of the potential outcome distributions. We show that the CATE parameter is nonparametrically identified under unconfoundedness and propose inverse probability weighted estimators for it. Under regularity conditions, some of which are standard and some are new in the literature, we show (pointwise) consistency and asymptotic normality of a fully nonparametric and a semiparametric estimator. We apply our methods to estimate the average effect of a first-time mother's smoking during pregnancy on the baby's birth weight as a function of the mother's age. A robust qualitative finding is that the expected effect becomes stronger (more negative) for older mothers.

KEY WORDS: Birth weight; Inverse probability weighted estimation; Nonparametric method; Treatment effect heterogeneity.

1. INTRODUCTION

When individual treatment effects in the population are heterogeneous, but treatment assignment is unconfounded given a vector X of observable covariates, it is a well-known result that the average treatment effect (ATE) in the population is nonparametrically identified (see, e.g., Rosenbaum and Rubin 1983, 1985). Given the heterogeneity of individual effects, it may also be of interest to estimate ATE in various subpopulations defined by the possible values of some component(s) of X. We will refer to the value of the ATE parameter within such a subpopulation as a conditional average treatment effect (CATE). For example, if one of the covariates is gender, one might be interested in estimating ATE separately for males and females. As treatment assignment in the two subpopulations is unconfounded given the rest of the components of X, one can simply split the sample by gender and apply standard nonparametric estimators of ATE to the two subsamples. A second example, considered by a number of authors, is to define CATE as a function of the full set of conditioning variables X. In this case CATE(x) gives the conditional mean of the treatment effect for any point x in the support of X.

Though not referred to by this name, the CATE function introduced in the second example already appears in Hahn (1998) and Heckman, Ichimura, and Todd (1997, 1998) as a "first stage" estimand in the (imputation-based) nonparametric estimation of ATE. Heckman and Vytlacil (2005) discussed the identification and estimation of CATE(x), which they called ATE(x), in terms of the marginal treatment effect in a general structural model. Khan and Tamer (2010) mentioned CATE(x) explicitly, but their focus was on ATE. Lee and Whang (2009) and Hsu (2012) considered estimating and testing hypotheses about CATE(x) when X is absolutely continuous, and provided detailed asymptotic theory. MaCurdy, Chen, and Hong (2011) also discussed the identification and estimation of CATE(x).

In this article we extend the concept of CATE to the technically more challenging situation in which the conditioning covariates X_1 are continuous and form a strict subset of X. As the unconfoundedness assumption will not generally hold conditional on X_1 alone, it is not possible to simply apply, say, the Lee and Whang (2009) CATE estimator with X_1 playing the role of X. Rather, one needs to estimate CATE as a function of X, and then average out the unwanted components by integrating with respect to the conditional distribution of $X_{(1)}$ given X_1, where $X_{(1)}$ denotes those components of X that are not in X_1. This distribution is, however, generally unknown and has to be estimated.

When X_1 is a discrete variable (such as in the first example), averaging with respect to the empirical distribution of $X_{(1)}|\, X_1$ is accomplished "automatically" by virtue of the ATE estimator being implemented subsample-by-subsample. The result is an estimate of CATE(x_1) for each point x_1 in the support of X_1. This suggests that when X_1 is continuous one could at least

Journal of Business & Economic Statistics
October 2015, Vol. 33, No. 4
DOI: 10.1080/07350015.2014.975555
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jbes*.

approximate CATE by discretizing X_1 and estimating ATE on the resulting subsamples provided that they are large enough. However, the CATE estimate obtained this way will depend on the discretization used and will be rather crude and discontinuous, just as a histogram is generally a crude and discontinuous estimate of the underlying density function.

The technical contribution of this article consists of proposing "smooth" nonparametric and semiparametric estimators of CATE when X_1 is continuous and a strict subset of X, and developing the first-order asymptotic theory of these estimators. The estimators are constructed as follows. First, the propensity score, the probability of treatment conditional on X, is estimated by either a kernel-based regression (the fully nonparametric case) or by a parametric model (the semiparametric case). In the second step the observed outcomes are weighted based on treatment status and the inverse of the estimated propensity score, and local averages are computed around points in the support X_1, using another set of kernel weights. (Intuitively, the second stage can be interpreted as integrating with respect to a smoothed estimate of the conditional distribution of the inverse propensity weighted outcomes given X_1.) Under regularity conditions the estimator is shown to be consistent and asymptotically normal; the results allow for pointwise inference about CATE as a function of X_1. Of the conditions used to prove these results, the most noteworthy ones are those that are used in the fully nonparametric case to restrict the relative convergence rates of the two smoothing parameters employed in steps one and two, and prescribe the order of the kernels.

The CATE estimators described above can be regarded as generalizations of the inverse probability weighted ATE estimator proposed by Hirano, Imbens, and Ridder (2003). An alternative (first-order equivalent) estimator of CATE could be based on nonparametric imputation (e.g., Hahn 1998). In this article we restrict attention to the first approach.

We present simulation results as well as an empirical exercise to illustrate the finite sample properties and the practical implementation of our estimators. Specifically, we estimate the expected effect of a first-time mother's smoking during pregnancy on the birth weight of her child conditional on the mother's age, using vital statistics data from North Carolina. In this exercise we focus on the semiparametric estimator, as in most applied settings, such as the one at hand, the applicability of a fully nonparametric procedure is likely to be hampered by curse of dimensionality problems.

The intended contribution of the application to the pertaining empirical literature is to explore the heterogeneity of the smoking effect along a given dimension in an unrestricted and intuitive fashion. Previous estimates reported in the literature are typically constrained to be a single number by the functional form of the underlying regression model. If the effect of smoking is actually heterogeneous, such an estimate is of course not informative about how much the effect varies across relevant subpopulations and may not even be consistent for the overall population mean.

Nevertheless, there have been some attempts in the literature to capture the heterogeneity of the treatment effect in question. Most notably, Abrevaya and Dahl (2008) estimated the effect of smoking separately for various quantiles of the birth weight distribution. Though insightful, a drawback of the quantile regression approach is that it allows for heterogeneity in the treatment effect across subpopulations that are not identifiable based on the mother's characteristics alone. Hence, the estimated effects are hard to translate into targeted "policy" recommendations. For instance, Abrevaya and Dahl (2008) reported that the negative effect of smoking on birth weight is more pronounced at the median of the birth weight distribution than at the 90th percentile. However, it is not clear, before actual birth, or at least without additional modeling, which quantile should be "assigned" to a mother with a given set of observable characteristics. In addition, the treatment effect for any given quantile could also be a function of these characteristics (this is assumed away by the linear specification they use). In contrast, the CATE parameter is defined as a function of variables that are observable a priori, and the estimator proposed in this article places only mild restrictions on the shape of this function.

Qualitatively, the main story that emerges from our empirical exercise is that the predicted average effect of smoking becomes stronger (more negative) at higher age. This finding is reasonably robust with respect to race, smoothing parameters, and the specification of the propensity score. Nevertheless, there is a fair amount of specification and estimation uncertainty about the numerical extent of this variation.

The rest of the article is organized as follows. Section 2 introduces the CATE parameter and discusses its identification and estimation. The first-order asymptotic properties of the proposed estimators are developed. Section 3 presents the simulation results, and Section 4 is devoted to the empirical exercise. Section 5 outlines possible extensions of the basic framework, including multivalued treatments and instrumental variables. Section 6 concludes.

2. THEORY

2.1 The Formal Framework and the Proposed Estimators

Let D be a dummy variable indicating treatment status in a population of interest with $D = 1$ if an individual (unit) receives treatment and $D = 0$ otherwise. Define $Y(1)$ as the potential outcome for an individual if treatment is imposed exogenously; $Y(0)$ is the corresponding potential outcome without treatment. Let X be a k-dimensional vector of covariates with $k \geq 2$. The econometrician observes D, X, and $Y \equiv D \cdot Y(1) + (1 - D) \cdot Y(0)$. In particular, we make the following assumption.

Assumption 1 (Sampling). The data, denoted $\{(D_i, X_i, Y_i)\}_{i=1}^n$, is a random sample of size n from the joint distribution of the vector (D, X, Y).

Throughout the article we maintain the assumption that the observed vector X can fully control for any endogeneity in treatment choice. Stated formally:

Assumption 2 (Unconfoundedness). $(Y(0), Y(1)) \perp D \big| X$.

Assumption 2 is also known as (strongly) "ignorable treatment assignment" (Rosenbaum and Rubin 1983), and it is a rather strong but standard identifying assumption in the treatment effect literature. In particular, it rules out the existence of unobserved factors that affect treatment choice and are also correlated with the potential outcomes.

Let $X_1 \in \mathbb{R}^\ell$ be a subvector of $X \in \mathbb{R}^k$, $1 \le \ell < k$, X absolutely continuous. The *conditional average treatment effect* (CATE) given $X_1 = x_1$ is defined as

$$\tau(x_1) \equiv E[Y(1) - Y(0)|X_1 = x_1].$$

Under Assumption 2, $\tau(x_1)$ can be identified from the joint distribution of (X, D, Y) as

$$\tau(x_1) = E[E[Y|D = 1, X] - E[Y|D = 0, X]|\, X_1 = x_1] \quad (1)$$

or

$$\tau(x_1) = E\left[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\middle| X_1 = x_1\right], \quad (2)$$

where $p(x) = P[D = 1|X = x]$ denotes the propensity score function. These identification results follow from a simple string of equalities justified by the law of iterated expectations and unconfoundedness. While Equation (1) identifies CATE somewhat more intuitively, we will base our estimators on Equation (2).

In particular, we propose the following procedure for estimating $\tau(x_1)$. The first step consists of estimating the propensity score. We consider two options. Option (i) is a nonparametric estimator given by a kernel-based (Nadaraya-Watson) regression, that is,

$$\hat{p}(x) = \frac{\frac{1}{nh^k}\sum_{i=1}^n D_i K\left(\frac{X_i - x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, \quad (3)$$

where $K(\cdot)$ is a kernel function and h is a smoothing parameter (bandwidth). Option (ii) is a parametric estimate of $p(x)$, for example, a logit or probit model estimated by maximum likelihood.

Given an estimator $\hat{p}(x)$ for the propensity score, in the second stage we estimate $\tau(x_1)$ by inverse probability weighting and kernel-based local averaging, that is, we propose

$$\hat{\tau}(x_1) = \frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n \left(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\right) K_1\left(\frac{X_{1i} - x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i} - x_1}{h_1}\right)},$$

where $K_1(u)$ is a kernel function and h_1 is a bandwidth (different from K and h). As in the second stage $\hat{p}(x)$ is evaluated at the sample observations X_i, in the nonparametric case we employ the "leave-one-out" version of (3).

As far as econometric theory is concerned, it is the development of the asymptotics of the fully nonparametric case that is the central contribution of this article. Though the asymptotics are less novel and challenging, the semiparametric estimator is easier to implement and a more robust practical alternative (at the cost of potential misspecification bias and loss of efficiency).

Even though we work out the asymptotic theory of $\hat{\tau}(x_1)$ for any $\ell < k$, in our assessment the most relevant case in practice is $\ell = 1$ (and maybe $\ell = 2$). When X_1 is a scalar, $\hat{\tau}(x_1)$ can easily be displayed as a two-dimensional graph while for higher dimensional X_1 the presentation and interpretation of the CATE estimator can become rather cumbersome.

2.2 CATE in a Linear Regression Framework

The standard linear regression model for program evaluation combines a weaker version of the unconfoundedness assumption with the assumption that the conditional expectation of the potential outcomes is a linear function. As noted by Imbens and Wooldridge (2009), the treatment effect literature has gradually moved away from this baseline model over the last 10–15 years. The main reason is that the estimated average treatment effect can be severely biased if the linear functional form is not correct. Nevertheless, as the general CATE parameter introduced in this article has not yet been in use in the treatment effect literature, it is useful to develop further intuition by relating it to a standard linear regression framework.

Given the vector X of covariates, we can, without loss of generality, write

$$E[Y(d) \mid X] = \mu_d + r_d(X), \quad d = 0, 1,$$

where $\mu_d = E[Y(d)]$ and $r_d(\cdot)$ is some function with $E[r_d(X)] = 0$. Under the unconfoundedness assumption, the mean of the observed outcome conditional on D and X can be represented as

$$E(Y \mid X, D) = \mu_0 + (\mu_1 - \mu_0)D + r_0(X) + [r_1(X) - r_0(X)]D. \quad (4)$$

If one assumes $r_0(X) = [X - E(X)]'\beta$ and $r_1(X) = [X - E(X)]'(\beta + \delta)$, then CATE$(X_1)$ is given by

$$\text{CATE}(X_1) = \mu_1 - \mu_0 + E[(X - E(X))' \mid X_1]\delta.$$

Further assuming that the above conditional expectation w.r.t. X_1 is a linear function of X_1 gives rise to a three-step parametric estimator of CATE. The first step consists of regressing Y on a constant, D, X and $D \cdot (X - \bar{X})$; specifically, we write

$$Y_i = \hat{\kappa} + \hat{\alpha} D_i + X_i'\hat{\beta} + D_i(X_i - \bar{X})'\hat{\delta} + \hat{\epsilon}_i, \quad i = 1, \ldots, n. \quad (5)$$

The second step consists of regressing each component of $X - \bar{X}$ on a constant and X_1:

$$X_i^{(j)} - \bar{X}^{(j)} = \tilde{X}_{1i}'\hat{\gamma}^{(j)} + \hat{u}_i^{(j)}, \quad i = 1, \ldots, n,\ j = 1, \ldots, k, \quad (6)$$

where $\tilde{X}_1 \equiv (1, X_1')'$, and $X^{(j)}$ and $\bar{X}^{(j)}$ denote the jth component of X and $\bar{X}$, respectively. Finally, for $X_1 = x_1$, one takes

$$\hat{\alpha} + (\tilde{x}_1'\hat{\gamma})\hat{\delta} \quad (7)$$

as an estimate of CATE(x_1), where $\hat{\gamma} \equiv (\hat{\gamma}^{(1)}, \ldots, \hat{\gamma}^{(k)})$ is an $(\ell + 1) \times k$ matrix.

The special case in which $\delta = 0$, that is, $r_0(X) = r_1(X)$, corresponds to assuming that individual treatment effects do not systematically depend on X. Accordingly, CATE reduces to a constant function whose value is equal to ATE$= \mu_1 - \mu_0$ everywhere, and ATE itself can be estimated as the coefficient on D from regression (5) without the interaction terms.

Though it requires entirely standard methods, calculating the standard error of (7) is somewhat cumbersome. Specifically, one can write the $k + 1$ regressions in (5) and (6) as a SUR system (see, e.g., Wooldridge 2010, Ch. 7) and estimate the joint variance-covariance matrix of all regression coefficients. Then one can invoke the multivariate delta method to obtain the standard error of (7) for any given x_1. The construction is described in detail in Appendix A. Alternatively, one could resample from the empirical distribution of the residuals and compute bootstrapped standard errors.

2.3 Asymptotic Properties of $\hat{\tau}(x_1)$: The Fully Nonparametric Case

In the fully nonparametric case we estimate the propensity score by a kernel-based nonparametric regression:

Assumption 3 (Estimated propensity score). $\hat{p}(X_i)$ is given by the leave-i-out version of the estimator in (3).

We derive the asymptotic properties of the resulting CATE estimator under the following regularity conditions.

Assumption 4 (Distribution of X). The support $\mathcal{X}$ of the k-dimensional covariate X is a Cartesian product of compact intervals, and the density of X, $f(x)$, is bounded away from 0 on $\mathcal{X}$.

Let s and s_1 denote positive even integers such that $s \geq k$ and $s_1 \geq k$.

Assumption 5 (Conditional moments and smoothness). (i) $\sup_{x\in\mathcal{X}} E[Y(j)^2 | X = x] < \infty$ for $j = 0, 1$; (ii) the functions $m_j(x) = E[Y(j) \mid X = x]$, $j = 0, 1$ and $f(x)$ are s-times continuously differentiable on $\mathcal{X}$.

Assumption 6 (Population propensity score). (i) $p(x)$ is bounded away from 0 and 1 on $\mathcal{X}$; (ii) $p(x)$ is s-times continuously differentiable on $\mathcal{X}$.

A function $\kappa : \mathbb{R}^k \to \mathbb{R}$ is a kernel of order s if it integrates to one over $\mathbb{R}^k$, and $\int u_1^{p_1} \cdot \ldots \cdot u_k^{p_k} \kappa(u) du = 0$ for all nonnegative integers $p_1, \ldots, p_k$ such that $1 \leq \sum_i p_i < s$.

Assumption 7 (Kernels).

(i) $K(u)$ is a kernel of order s, is symmetric around zero, is equal to zero outside $\prod_{i=1}^k [-1, 1]$, and is continuously differentiable.
(ii) $K_1(u)$ is a kernel of order s_1, is symmetric around zero, and is s times continuously differentiable.

Assumption 8 (Bandwidths). The bandwidths h and h_1 satisfy the following conditions as $n \to \infty$:

(i) $h \to 0$ and $\log(n)/(nh^{k+s}) \to 0$.
(ii) $nh_1^{2s_1+\ell} \to 0$ and $nh_1^{\ell} \to \infty$.
(iii) $h^{2s}h_1^{-2s-\ell} \to 0$ and $nh_1^{\ell}h^{2s} \to 0$.

Define the function $\psi(x, y, d)$ as

$$\psi(x, y, d) \equiv \frac{d(y - m_1(x))}{p(x)} - \frac{(1 - d)(y - m_0(x))}{1 - p(x)} + m_1(x) - m_0(x).$$

The following theorem states our main theoretical result. The proof is given in Appendices B and C.

Theorem 1. Suppose that Assumptions 1 through 8 are satisfied. Then, for each point x_1 in the support of X_1,

$$\text{(a)} \quad \sqrt{nh_1^{\ell}}(\hat{\tau}(x_1) - \tau(x_1)) = \frac{1}{\sqrt{nh_1^{\ell}}} \frac{1}{f_1(x_1)} \sum_{i=1}^{n} [\psi(X_i, Y_i, D_i) - \tau(x_1)] K_1\left(\frac{X_{1i} - x_1}{h_1}\right) + o_p(1)$$

$$\text{(b)} \quad \sqrt{nh_1^{\ell}}(\hat{\tau}(x_1) - \tau(x_1)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\|K_1\|_2^2 \sigma_{\psi}^2(x_1)}{f_1(x_1)}\right),$$

where $f_1(x_1)$ is the pdf of X_1, $\|K_1\|_2 \equiv (\int K_1(u)^2 du)^{1/2}$, and $\sigma_{\psi}^2(x_1) \equiv E[(\psi(X, Y, D) - \tau(x_1))^2 | X_1 = x_1]$.

Comments

1. The technical restrictions imposed on the distribution of X and on various conditional moment functions in Assumptions 4 and 5 are analogous to those in Hirano, Imbens, and Ridder (2003) and are common in the literature on nonparametric estimation. As pointed out by Khan and Tamer (2010), the assumption that the propensity score is bounded away from zero and one plays an important role in determining the convergence rate of inverse probability weighted estimators.
2. Assumptions 7(i) and 8(i) ensure that $\hat{p}(x) - p(x) = o_p(h^{s/2})$, uniformly in x; see Lemma 6.1 part (b) in Appendix B. This is the convergence rate needed to establish the influence function representation in Theorem 1(a). The influence function itself is analogous to the influence function that efficient nonparametric estimators of ATE possess; see, for example, Hahn (1998) and Hirano, Imbens, and Ridder (2003).
3. The influence function $[\psi(X_i, Y_i, D_i) - \tau(x_1)] K_1\left(\frac{X_{1i} - x_1}{h_1}\right)$ does not have mean zero; rather, it decomposes into a "bias term" that depends on h_1 and a mean zero term to which Lyapunov's CLT directly applies. Assumptions 7(ii) and 8(ii) ensure the asymptotic negligibility of the bias term (and the applicability of the CLT). See Appendix C for further details.
4. Assumption 8(iii) underlies the novel aspects of our asymptotic theory. It controls the relative and "joint" convergence rates of h (the bandwidth used to estimate the propensity score) and h_1 (the bandwidth used in the integration step). These rates, along with the kernel orders, are chosen subject to numerous tradeoffs that need to be considered to ensure the asymptotic negligibility of all remainder terms in our expansion of $\hat{\tau}(x_1)$.
5. More specifically, our asymptotic analysis builds on the expansion of the Hirano, Imbens, and Ridder (2003) ATE estimator by Ichimura and Linton (2005). However, the integration step causes the factor $K_1(\cdot/h_1)$ to appear in each term, which has a number of consequences. First, the leading terms in the expansion converge at the rate of $\sqrt{nh_1^{\ell}}$ rather than $\sqrt{n}$. Second, as the convergence rates of the original remainder terms depend on h, the presence of K_1 and the

scaling by $\sqrt{nh_1^\ell}$ introduce interactions between h and h_1. These interactions require that h_1 converges to zero slower than h. In particular, if one were to set h_1 equal to a constant, then all remainder terms could be made to vanish by requiring $h \to 0$ at an appropriate rate as in Ichimura and Linton (2005) or Donald, Hsu, and Lieli (2014b). However, the bias in the leading term, described in comment 3 above, would of course not disappear. Hence, one also needs $h_1 \to 0$, but slowly enough to satisfy part two of Assumption 8(ii) and part one of 8(iii). One can then employ a kernel K_1 of sufficiently high order to satisfy the first part of Assumption 8(ii), and a kernel K of sufficiently high order to satisfy part two of Assumption 8(iii), which is needed to ensure that the (conditional) bias of $\hat{p}(x)$ remains asymptotically negligible when scaled by $\sqrt{nh_1^\ell}$. Note, however, that increasing s, the order of K, is not costless—it slows the convergence of h to zero via Assumption 8(i), which then slows the convergence of h_1 to zero via the first part of 8(iii), which again requires an increase in s_1 and possibly s, etc.

6. We have yet to show that there actually exist bandwidth sequences and kernel orders satisfying all the requirements posed by Assumption 8 (otherwise the asymptotic theory would be vacuous). We set

$$h = a \cdot n^{\frac{-1}{k+s+\delta}}, \quad a > 0,\ \delta > 0,$$
$$h_1 = a_1 \cdot n^{\frac{-1}{\ell+2s_1-\delta_1}}, \quad a_1 > 0,\ \delta_1 > 0,$$

where δ and δ_1 can be made as small as necessary or desired. It is clear that Assumptions 8(i) and (ii) hold with these choices. To satisfy Assumption 8(iii), we further set the kernel orders as $s = k$ for k even, $s = k+1$ for k odd, and $s_1 = s+2$.

To verify $h^{2s}h_1^{-2s-\ell} \to 0$, note that δ and δ_1 can be arbitrarily small, so it is sufficient to check

$$\frac{-2s}{k+s} + \frac{2s+\ell}{2s+4+\ell} < 0.$$

This is obviously true because $-2s/(k+s) < -1$ and $(2s+\ell)/(2s+4+\ell) < 1$ under our selections.

To verify $nh_1^\ell h^{2s} \to 0$, note that by Assumption 8(ii), $nh_1^\ell h^{2s} = nh_1^{2s_1+\ell} \cdot h^{2s}h_1^{-2s_1} \to 0$ when $h^s h_1^{-s_1} \to 0$, so it is sufficient to check the latter. Again, since δ and δ_1 can be arbitrarily small, we only need

$$\frac{-s}{k+s} + \frac{s+2}{2s+4+\ell} < 0,$$

which is obvious because $-s/(k+s) < -1/2$ and $(s+2)/(2s+4+\ell) < 1/2$ under our selections.

7. To use Theorem 1 for statistical inference, one needs to consistently estimate $\sigma_\psi^2(x_1)$ and $f_1(x_1)$. The latter is easily accomplished by, say, $\hat{f}_1(x_1) = \frac{1}{nh_1^\ell}\sum_{i=1}^n K_1[(X_{1i}-x_1)/h_1]$. It is more involved to estimate $\sigma_\psi^2(x_1)$ because ψ includes unknown functions: $m_1(x)$, $m_0(x)$, and $p(x)$. Let $\hat{m}_1(x)$ be a uniformly consistent estimator for $m_1(x)$ over $\mathcal{X}$ in that $\sup_{x\in\mathcal{X}}|\hat{m}_1(x) - m_1(x)| = o_p(1)$. Similarly, let $\hat{m}_0(x)$ and $\hat{p}(x)$ be uniformly consistent estimators for $m_0(x)$ and $p(x)$ over $\mathcal{X}$. In particular, the $\hat{p}(x)$ we use is uniformly consistent for $p(x)$. Also, such $\hat{m}_1(x)$ and $\hat{m}_0(x)$ can be obtained by performing kernel regressions of Y on X in the treated and nontreated subpopulations, respectively. Then, we estimate $\sigma_\psi^2(x_1)$ by

$$\hat{\sigma}_\psi^2(x_1) = \left[\frac{1}{nh_1^\ell}\sum_{i=1}^n (\hat{\psi}(X_i, Y_i, D_i) - \hat{\tau}(x_1))^2 K_1\left(\frac{X_{1i}-x_1}{h_1}\right)\right] \Big/ \hat{f}_1(x_1), \tag{8}$$
$$\hat{\psi}(x, y, d) = \frac{d(y-\hat{m}_1(x))}{\hat{p}(x)} - \frac{(1-d)(y-\hat{m}_0(x))}{1-\hat{p}(x)} + \hat{m}_1(x) - \hat{m}_0(x).$$

The consistency of $\hat{\sigma}_\psi^2(x_1)$ can be shown as follows. First, let $\tilde{\sigma}_\psi^2(x_1)$ be the (infeasible) estimator for $\sigma_\psi^2(x_1)$ where we replace $\hat{\psi}(X_i, Y_i, D_i) - \hat{\tau}(x_1)$ with $\psi(X_i, Y_i, D_i) - \tau(x_1)$ in (8). It is easy to see that $\tilde{\sigma}_\psi^2(x_1)$ is a consistent estimator for $\sigma_\psi^2(x_1)$. Next, note that $\hat{\psi}(x, y, d)$ is a uniformly consistent estimator for $\psi(x, y, d)$ and $\hat{\tau}(x_1)$ is a consistent estimator for $\tau(x_1)$. Therefore, using $\hat{\psi}(X_i, Y_i, D_i) - \hat{\tau}(x_1)$ in (8) is as good as $\psi(X_i, Y_i, D_i) - \tau(x_1)$ in that the estimation error will disappear in the limit.

8. Assumption 4 does not allow X to have discrete components, which is of course restrictive in applications. One way to incorporate discrete covariates into the analysis is as follows. For concreteness, suppose that in addition to some continuous variables, X contains gender. Let M denote the indicator of the male subpopulation and define $p(x, m) = P(D = 1 \mid X = x, M = m)$ for $m = 0, 1$. We can estimate these functions by kernel based regressions of D on X in each subsample:

$$\hat{p}(x, 1) = \frac{\frac{1}{nh^k}\sum_{\{i:M_i=1\}} D_i K\left(\frac{X_i-x}{h}\right)}{\frac{1}{nh^k}\sum_{\{i:M_i=1\}} K\left(\frac{X_i-x}{h}\right)} = \frac{\frac{1}{nh^k}\sum_i^n M_i D_i K\left(\frac{X_i-x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^n M_i K\left(\frac{X_i-x}{h}\right)},$$

and $\hat{p}(x, 0)$ is obtained analogously. We claim that $\tau(x)$ can be estimated by

$$\hat{\tau}(x) = \frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n \left(\frac{D_i Y_i}{\hat{p}(X_i, M_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i, M_i)}\right) K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}. \tag{9}$$

In particular, as we show in Appendix D, the following influence function representation applies to $\hat{\tau}(x_1)$:

$$\sqrt{nh_1^\ell}(\hat{\tau}(x_1) - \tau(x_1)) = \frac{1}{f_1(x_1)\sqrt{nh_1^\ell}}\sum_{i=1}^n [\psi(M_i, X_i, Y_i, D_i) - \tau(x_1)] \times K_1\left(\frac{X_{1i}-x_1}{h_1}\right) + o_p(1), \tag{10}$$

where

$$\psi(M_i, X_i, Y_i, D_i) \equiv \frac{D_i(Y_i - m(X_i, M_i))}{p(X_i, M_i)} - \frac{(1 - D_i)(Y_i - m_0(X_i, M_i))}{1 - p(X_i, M_i)} + m_1(X_i, M_i) - m_0(X_i, M_i),$$

and $m_j(x, m) \equiv E[Y(j)|X = x, M = m]$ for $j = 0, 1$ and $m = 0, 1$. Therefore, Theorem 1 continues to hold for (9) after replacing $\psi(X_i, Y_i, D_i)$ with $\psi(M_i, X_i, Y_i, D_i)$.

9. Kernels satisfying Assumption 7 can be constructed by taking products of higher order univariate kernels. A general method for obtaining higher order kernels from "regular" ones is described, for example, by Imbens and Ridder (2009). The "support" condition imposed on K is for expositional convenience only; we can extend the proof of Theorem 1 to kernels with exponential tails.

2.4 Asymptotic Properties of $\hat{\tau}(x_1)$: The Semiparametric Case

The asymptotic theory of estimating CATE simplifies considerably if a parametric model is postulated for the propensity score. In particular, we replace Assumption 3 with the following.

Assumption 9 (Parametric propensity score estimator). The estimator $\hat{\theta}_n$ of the propensity score model $p(x; \theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, $d < \infty$, satisfies $\sup_{x \in \mathcal{X}} |p(x; \hat{\theta}_n) - p(x; \theta_0)| = O_p(n^{-1/2})$ where $\theta_0 \in \Theta$ such that $p(x) = p(x; \theta_0)$ for all $x \in \mathcal{X}$.

Assumption 9 will typically hold for standard parametric estimation methods under reasonably mild regularity conditions. For example, a logit model or a probit model based on a linear index and estimated by maximum likelihood will satisfy (9) if $\mathcal{X}$ is bounded. Obviously, Assumption 9 eliminates the need for those conditions stated in Section 2.3 whose role is to govern the behavior of the Nadaraya–Watson regression estimator and the interaction between the two bandwidths. Of course, the bandwidth h_1 used in the second, local averaging, stage still needs to be controlled to ensure consistency and asymptotic normality. Assumption 8(ii) with any $s_1 \geq 2$ is sufficient in this regard.

To state the result formally, define

$$\psi_\theta(x, y, d) \equiv \frac{dy}{p(x)} - \frac{(1 - d)y}{1 - p(x)}.$$

The following theorem corresponds closely to Theorem 1.

Theorem 2. Suppose that Assumptions 1, 2, 8(ii), and 9 are satisfied for some $s_1 \geq 2$. Then, under some additional regularity conditions, the following statements hold for each point x_1 in the support of X_1:

(a) $$\sqrt{nh_1^\ell}(\hat{\tau}(x_1) - \tau(x_1)) = \frac{1}{\sqrt{nh_1^\ell}} \frac{1}{f_1(x_1)} \sum_{i=1}^{n} [\psi_\theta(X_i, Y_i, D_i) - \tau(x_1)] K_1\left(\frac{X_{1i} - x_1}{h_1}\right) + o_p(1)$$

(b) $$\sqrt{nh_1^\ell}(\hat{\tau}(x_1) - \tau(x_1)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\|K_1\|_2^2 \sigma_{\psi_\theta}^2(x_1)}{f_1(x_1)}\right),$$

where $f_1(x_1)$ is the pdf of X_1, $\|K_1\|_2 \equiv (\int K_1(u)^2 du)^{1/2}$, and

$$\sigma_{\psi_\theta}^2(x_1) \equiv E[(\psi_\theta(X, Y, D) - \tau(x_1))^2 | X_1 = x_1].$$

The proof of Theorem 2 is given in Appendix E.

Comments

1. The form of the influence function highlights an important difference between estimating ATE and CATE. If ATE is estimated by inverse probability weighting, then even a $\sqrt{n}$-consistent parametric estimate of the propensity score will make a nontrivial contribution to the influence function (since the ATE estimator itself converges at the same rate). In contrast, the CATE estimator converges at a rate slower than $\sqrt{n}$, so employing a (correctly specified) parametric estimator is asymptotically equivalent to the propensity score being known.
2. The semiparametric approach offers several practical advantages over the fully nonparametric estimator: (i) It can help circumvent the curse of dimensionality problem when X is large. (ii) Discrete and continuous covariates can be treated the same way, that is, one can simply include, say, a gender dummy in a logit or probit regression rather than follow the partitioning approach described in comment 8 after Theorem 1. This is very useful if one of the categories has a small number of observations. (iii) Only the bandwidth used in the integration step needs to be chosen.
3. Of course, the advantages listed above do not come without costs. While the semparametric approach is still reasonably flexible, misspecification of the score function will generally bias the resulting CATE estimates. Furthermore, the semiparametric CATE estimator is less efficient than the nonparametric one. In particular, we can show that

$$\sigma_\psi^2(x_1) = E\left[(m_1(X) - m_0(X) - \tau(x_1))^2 + \frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1 - p(X)} \middle| X_1 = x_1\right],$$

$$\sigma_{\psi_\theta}^2(x_1) = \sigma_\psi^2(x_1) + E\left[p(X)(1 - p(X))\left(\frac{m_1(X)}{p(X)} + \frac{m_0(X)}{1 - p(X)}\right)^2 \middle| X_1 = x_1\right],$$

where $\sigma_d^2(x) = \text{var}(Y(d)|X = x)$ for $d = 0$ and 1. Hence, clearly $\sigma_{\psi_\theta}^2(x_1) \geq \sigma_\psi^2(x_1)$. Therefore, for a given choice of h_1 and K_1 satisfying the conditions in Theorems 1 and 2, the semiparametric CATE estimator is less efficient. This result is not surprising given that the influence function of the semiparametric estimator is the same as if $p(x)$ were known (see comment 1 above). It is well known from the work of Hahn (1998) and Hirano, Imbens, and Ridder (2003) that such estimators of ATE do not attain the semiparametric efficiency bound constructed with or without the knowledge of $p(x)$.
4. The "additional regularity conditions" mentioned in the theorem are needed to ensure that (i) Lyapunov's CLT can be applied to the leading term (A.17) in the expansion of the estimator in Appendix E, and (ii) that the remainder term is

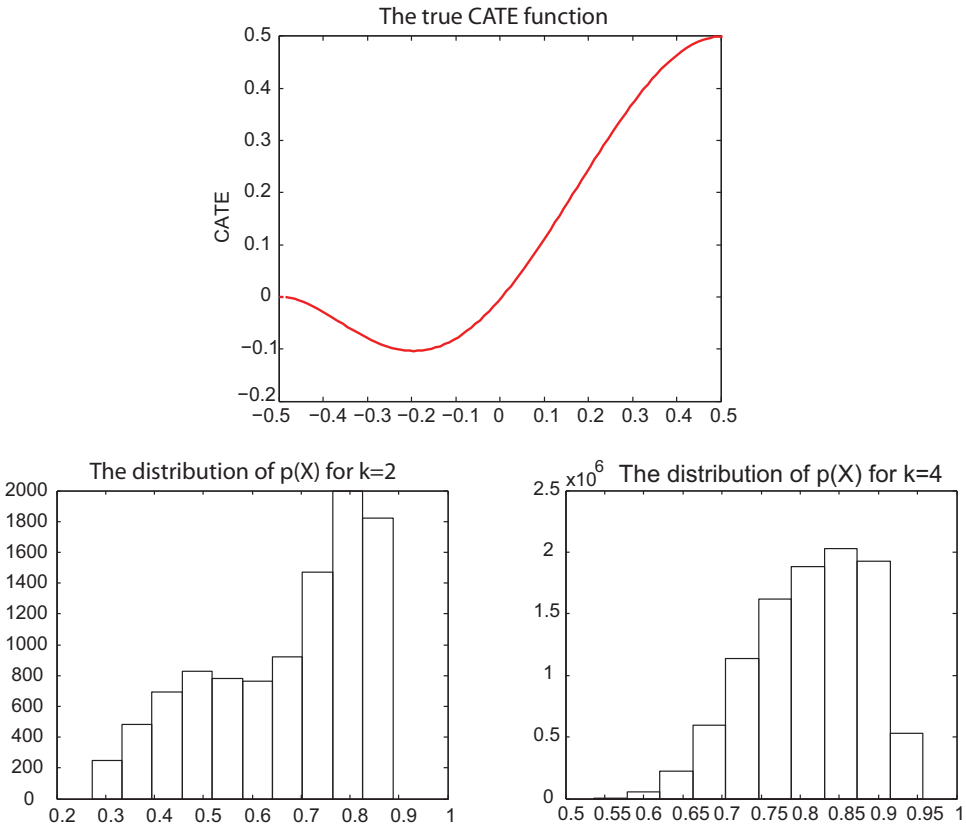

Figure 1. The CATE(x_1) function and the distribution of $p(X)$.

well behaved in that the second factor in (A.18) is $O_p(1)$. A set of primitive sufficient conditions could be obtained by suitable changes to Assumptions 4, 5, 6, and 7(ii).

5. K_1 no longer needs to be a higher order kernel in the semiparametric case, but such kernels could still be used in practice.

3. MONTE CARLO SIMULATIONS

In this section we present a Monte Carlo exercise aimed at evaluating the finite sample accuracy of the asymptotic approximations given in Theorems 1 and 2. Part of the design is admittedly artificial in that we set $k = \dim(X) \in \{2, 4\}$, while in practice justifying the unconfoundedness assumption typically requires conditioning on a much larger set of covariates. Nevertheless, the setup is rich enough to allow us to make a number of salient and practically relevant points without taking on an undue computational burden. We pay particular attention to exploring the sensitivity of the finite sample distribution of $\hat{\tau}(x_1)$ to the smoothing parameter(s) as first-order asymptotic theory has virtually no implications about how to choose these parameters in practice. Curse of dimensionality issues in the fully nonparametric case are already apparent with four covariates.

3.1 The Data-Generating Process and the Exercise

We consider two data-generating processes (DGPs); one with $k = 2$ and another with $k = 4$. In the bivariate case the covariates $X = (X_1, X_2)$ are given by

$$X_1 = \epsilon_1, \;\; X_2 = (1 + 2X_1)^2(-1 + X_1)^2 + \epsilon_2,$$

where $\epsilon_i \sim$ iid unif$[-0.5, 0.5]$, $i = 1, 2$. The potential outcomes are defined as

$$Y(1) = X_1X_2 + \nu \qquad \text{and} \qquad Y(0) = 0,$$

where $\nu \sim N(0, 0.25^2)$ and is independent of (ϵ_1, ϵ_2). These definitions imply

$$\text{CATE}(x_1) = x_1(1 + 2x_1)^2(-1 + x_1)^2,$$

plotted in the top panel of Figure 1. The propensity score is set as $p(X) = \Lambda(X_1 + X_2)$, where $\Lambda(\cdot)$ is the c.d.f. of the logistic distribution. The distribution of the random variable $p(X_1, X_2)$ is shown in the lower left panel of Figure 1.

To investigate the impact of the curse of dimensionality on our estimator, we also consider the following modification to the DGP:

$$X_1 = \epsilon_1, \;\; X_2 = 1 + 2X_1 + \epsilon_2, \;\; X_3 = 1 + 2X_1 + \epsilon_3,$$
$$X_4 = (-1 + X_1)^2 + \epsilon_4$$
$$Y(1) = X_1X_2X_3X_4 + \nu, \;\; Y(0) = 0$$
$$p(X) = \Lambda(.5(X_1 + X_2 + X_3 + X_4)),$$

where again $\epsilon_i \sim$ iid unif$[-0.5, 0.5]$, $i = 1, 2, 3, 4$ and $\nu \sim N(0, 0.25^2)$, independent of the ϵ's. The distribution of the the random variable $p(X_1, \ldots, X_4)$ is shown in lower right panel of Figure 1. These modifications leave the CATE(x_1) function unchanged, but make it harder to estimate.

We estimate CATE(x_1), $x_1 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$, for samples of size $n = 500$ and $n = 5000$. The number of Monte Carlo repetitions is 1000 in the fully nonparametric case and 5000 in the semiparametric case.

Let $T(x_1) = \sqrt{nh_1}[\widehat{\text{CATE}}(x_1) - \text{CATE}(x_1)]$ and $S(x_1) = (T(x_1) - ET(x_1))/\widehat{\text{s.e.}}(T(x_1))$. For each x_1, we report (the Monte Carlo estimate of) the mean of $T(x_1)/\sqrt{nh_1}$ (i.e., the raw bias of CATE(x_1)), the standard deviation of $T(x_1)$, the estimated standard deviation of $T(x_1)$, the MSE of $T(x_1)$, the probability that $S(x_1)$ exceeds 1.645 and 1.96, respectively, and the probability that $S(x_1)$ is below -1.645 and -1.96, respectively.

We implement the fully nonparametric and semiparametric version of the estimator for various choices of the bandwidths h and h_1. Motivated by Comment 6 after Theorem 1, for $k = 2$ we set $s = 2, s_1 = 4, h = an^{-1/4}$ for $a \in \{0.5, 0.167, 1.5\}$ and $h_1 = a_1 n^{-1/9}$ for $a_1 \in \{0.2, 0.067, 0.6\}$. For $k = 4$, the corresponding choices are $s = 4$, $s_1 = 6$, $h = an^{-1/8}$ for $a \in \{0.5, 0.167, 1.5\}$ and $h_1 = a_1 n^{-1/13}$ for $a_1 \in \{0.2, 0.067, 0.6\}$. We will refer to the choices $a = 0.5$ and $a_1 = 0.2$ as "baseline;" these values were calibrated to ensure that for $n = 5000$ the actual distribution of $\hat{\tau}(x_1)$ is reasonably close to its theoretical limit for most of the five points considered, as measured by the statistics listed above. We then vary h and/or h_1 (by tripling and/or thirding a and a_1) to illustrate how oversmoothing or undersmoothing relative to this crude baseline affects the quality of the inferences drawn from the asymptotic results given in Theorems 1 and 2. Note that in specifying h and h_1 as above, we ignore the (small) positive constants δ and δ_1 and simply set them to zero.

Regarding further computational details, we use a normal kernel (and higher order kernels derived from it) throughout. In computing the variance of $\hat{\tau}(x_1)$, we use the bandwidth h and a regular kernel to estimate the functions m_0 and m_1 (in the fully nonparametric case); the bandwidth h_1 and the kernel K_1 to estimate $f_1(x_1)$, and the bandwidth h_1 and a regular kernel to estimate $\sigma^2_{\psi}(x_1)$ and $\sigma^2_{\psi_\theta}(x_1)$. The estimated propensity score is trimmed to lie in the interval [0.005, 0.995]. This constraint affects the fully nonparametric estimator for $n = 500$, but it is basically inconsequential for the semiparametric estimator or in larger samples.

There are other relevant issues we do not investigate here in detail due to constraints on space and scope. One is the sensitivity of our asymptotic approximations to the propensity score approaching zero or one with high probability. Based on results, both theoretical (Khan and Tamer 2010) and experimental (Huber et al. 2012; Donald, Hsu, and Lieli 2014a), we expect the asymptotic distributions given in Theorems 1 and 2 to be worse small sample approximations in this case. Another issue not addressed by the simulations is the sensitivity of the semiparametric estimator to the misspecification of the propensity score model. Nevertheless, the empirical exercise presented in Section 4 will provide an illustration of the possible impact and handling of these problems in practice.

3.2 Simulation Results

We highlight several aspects of the simulation results reported in Tables 1 through 3.

First, as seen in Panel 1 of Tables 1 and 2, the distribution of the fully nonparametric estimator for $n = 5000$ is reasonably close to its theoretical limit under the baseline bandwidth choices, and the studentized estimator leads to reliable inferences without major size distortions. In fact, for $k = 2$ the asymptotic theory seems to work well even for $n = 500$, but for $k = 4$ there is a significant improvement in the approximations as one goes from $n = 500$ to $n = 5000$. In particular, there is a marked reduction in the bias of the estimator at most points x_1, as well as the bias of the estimated standard error. More generally, comparing Tables 1 and 2 across panels provides evidence on the curse of dimensionality—for a given sample size, CATE(x_1) can be estimated much less precisely in terms of MSE when the dimensionality of X is larger, and the asymptotic approximation captures finite sample properties to a much lesser degree. (As explained in the previous section, the bandwidths across the two tables are adjusted for the change in the dimensionality of X by adjusting the exponent of n but not the scaling factor.) Viewed somewhat differently, for $k = 2$, results tend to change much less across the two sample sizes, while for $k = 4$, asymptotic theory generally starts "kicking in" much slower.

Second, as can be expected, the bias of $\hat{\tau}(x_1)$ is generally larger if x_1 is close to the boundary of the support of X_1. (Moreover, boundary bias is often accompanied by evidence of skewness.) Bias can be a result of oversmoothing with a large h_1 (compare, e.g., Panels 1 and 3 in Tables 1 and 2) or undersmoothing with a small h (compare Panels 1 and 4 in Table 1 and especially Table 2). Choosing h too small for the sample size has the additional effect of causing severe bias in the estimated standard errors, which, in turn, can throw the p-values of $S(x_1)$ completely off. Oversmoothing with h_1 also biases the standard errors, but this problem seems much milder in comparison. Nevertheless, biased standard errors do not always lead to bad inference. For example, as seen in Panel 1 of Table 2, for $n = 5000$ the estimated standard errors are still biased downward at points close to the boundary ($x_1 = -0.4$ and $x_1 = 0.4$), but the p-values of the studentized estimator are very close to their nominal levels. This suggests that the observed s.e. of $T(\pm 0.4)$ is blown up by just a few outlier estimates.

Third, while Assumption 8(iii) implies that the ratio h/h_1 should go to zero as $n \to \infty$, the results described in the previous paragraph caution against necessarily imposing $h << h_1$ in practice. Panel 6 in Tables 1 and 2 contains results for h small and h_1 large. As expected, there is severe bias in both the estimator and the standard errors, and inference is misleading (especially for $k = 4$). In contrast, if one takes the opposite extreme, that is, oversmooths with h and undersmooths with h_1, as shown in Panel 7 of Tables 1 and 2, the asymptotic theory delivers very nice approximations even for $n = 500$ and $k = 4$ and despite the fact that h/h_1 is very large. This does not mean that the asymptotics are invalid; it just means, as usual, that first-order asymptotic theory has little to say about how to pick the smoothing parameters in finite samples.

The observation that oversmoothing the propensity score estimate can be beneficial in finite samples leads naturally to considering the semiparametric version of the estimator. The results are displayed in Table 3.

Our fourth observation then is that the asymptotic normal approximation for the semiparametric estimator performs well

Table 1. The distribution of $\sqrt{nh_1}[\hat{\tau}(x_1) - \tau(x_1)]$: the fully nonparametric case with $k = 2$

	$n = 500$								$n = 5000$							
x_1	$\frac{\text{Mean}}{\sqrt{nh_1}}$	s.e.	E($\widehat{\text{s.e.}}$)	MSE	p-val. 1.65	p-val. -1.65	p-val 1.96	p-val -1.96	$\frac{\text{Mean}}{\sqrt{nh_1}}$	s.e.	E($\widehat{\text{s.e.}}$)	MSE	p-val. 1.65	p-val. -1.65	p-val 1.96	p-val -1.96
Panel 1	$h = 0.5n^{-1/4} = 0.106$, $h_1 = 0.2n^{-1/9} = 0.100$								$h = 0.5n^{-1/4} = 0.059$, $h_1 = 0.2n^{-1/9} = 0.078$							
−0.4	−0.002	0.337	0.276	0.114	0.074	0.080	0.044	0.047	−0.000	0.277	0.275	0.077	0.051	0.053	0.026	0.028
0.2	0.002	0.245	0.237	0.061	0.059	0.056	0.030	0.026	0.001	0.235	0.235	0.056	0.052	0.050	0.025	0.025
0	0.001	0.213	0.218	0.046	0.050	0.044	0.025	0.022	0.000	0.205	0.211	0.042	0.047	0.044	0.023	0.022
0.2	0.003	0.209	0.215	0.044	0.043	0.044	0.024	0.019	0.000	0.202	0.209	0.041	0.042	0.043	0.021	0.024
0.4	0.001	0.222	0.218	0.049	0.052	0.055	0.027	0.029	0.000	0.210	0.209	0.044	0.048	0.052	0.025	0.027
Panel 2	$h = 0.5n^{-1/4} = 0.106$, $h_1 = 0.067n^{-1/9} = 0.034$								$h = 0.5n^{-1/4} = 0.059$, $h_1 = 0.067n^{-1/9} = 0.026$							
−0.4	0.002	0.324	0.275	0.105	0.074	0.078	0.043	0.045	0.000	0.287	0.280	0.082	0.052	0.059	0.027	0.029
−0.2	−0.001	0.246	0.233	0.060	0.063	0.060	0.035	0.030	−0.000	0.240	0.233	0.057	0.058	0.055	0.030	0.028
0	−0.000	0.213	0.206	0.045	0.056	0.054	0.030	0.025	−0.000	0.204	0.204	0.042	0.051	0.047	0.025	0.026
0.2	0.001	0.213	0.199	0.045	0.063	0.059	0.035	0.031	0.000	0.208	0.198	0.043	0.058	0.062	0.031	0.033
0.4	0.007	0.238	0.208	0.057	0.071	0.081	0.041	0.042	0.001	0.235	0.207	0.056	0.074	0.072	0.043	0.040
Panel 3	$h = 0.5n^{-1/4} = 0.106$, $h_1 = 0.6n^{-1/9} = 0.301$								$h = 0.5n^{-1/4} = 0.059$, $h_1 = 0.6n^{-1/9} = 0.233$							
−0.4	−0.042	0.375	0.308	0.405	0.076	0.075	0.046	0.045	−0.038	0.325	0.299	1.746	0.067	0.066	0.037	0.036
−0.2	0.062	0.257	0.259	0.648	0.048	0.047	0.027	0.022	0.033	0.235	0.248	1.344	0.042	0.038	0.021	0.019
0	0.064	0.218	0.248	0.660	0.034	0.031	0.015	0.012	0.038	0.209	0.242	1.743	0.025	0.033	0.009	0.013
0.2	−0.030	0.232	0.258	0.185	0.033	0.033	0.016	0.016	−0.007	0.217	0.245	0.099	0.027	0.030	0.012	0.012
0.4	−0.100	0.272	0.305	1.582	0.030	0.034	0.012	0.017	−0.059	0.258	0.280	4.096	0.036	0.037	0.014	0.017
Panel 4	$h = 0.167n^{-1/4} = 0.035$, $h_1 = 0.2n^{-1/9} = 0.100$								$h = 0.167n^{-1/4} = 0.020$, $h_1 = 0.2n^{-1/9} = 0.078$							
−0.4	0.074	2.236	0.315	5.271	0.239	0.504	0.225	0.466	0.002	0.334	0.275	0.113	0.080	0.081	0.050	0.051
−0.2	−0.018	0.790	0.237	0.640	0.199	0.179	0.160	0.142	−0.002	0.251	0.235	0.065	0.063	0.059	0.034	0.034
0	−0.000	0.330	0.216	0.109	0.090	0.090	0.060	0.056	0.000	0.210	0.211	0.044	0.052	0.047	0.025	0.025
0.2	0.013	0.247	0.215	0.070	0.059	0.061	0.035	0.033	0.003	0.203	0.209	0.044	0.046	0.041	0.021	0.024
0.4	0.031	0.621	0.224	0.435	0.084	0.142	0.064	0.090	0.003	0.211	0.208	0.049	0.047	0.053	0.026	0.026
Panel 5	$h = 1.5n^{-1/4} = 0.317$, $h_1 = 0.2n^{-1/9} = 0.100$								$h = 1.5n^{-1/4} = 0.178$, $h_1 = 0.2n^{-1/9} = 0.078$							
−0.4	−0.006	0.273	0.259	0.076	0.059	0.063	0.031	0.035	−0.001	0.268	0.266	0.072	0.051	0.056	0.026	0.029
−0.2	0.006	0.235	0.234	0.057	0.054	0.047	0.027	0.024	0.002	0.237	0.236	0.057	0.054	0.047	0.025	0.026
0	0.001	0.211	0.221	0.044	0.046	0.038	0.023	0.018	0.000	0.204	0.213	0.042	0.044	0.043	0.022	0.020
0.2	0.008	0.218	0.221	0.051	0.047	0.047	0.026	0.021	0.002	0.210	0.210	0.045	0.046	0.052	0.023	0.025
0.4	0.014	0.248	0.230	0.071	0.060	0.064	0.031	0.038	0.007	0.228	0.213	0.069	0.058	0.063	0.034	0.033
Panel 6	$h = 0.167n^{-1/4} = 0.035$, $h_1 = 0.6n^{-1/9} = 0.301$								$h = 0.167n^{-1/4} = 0.020$, $h_1 = 0.6n^{-1/9} = 0.233$							
−0.4	−0.003	2.354	0.323	5.543	0.259	0.480	0.241	0.442	−0.037	0.421	0.299	1.771	0.103	0.104	0.070	0.065
−0.2	0.073	1.096	0.256	2.011	0.239	0.322	0.220	0.276	0.032	0.259	0.247	1.263	0.059	0.055	0.033	0.027
0	0.066	0.450	0.244	0.856	0.124	0.116	0.093	0.085	0.038	0.215	0.241	1.725	0.029	0.036	0.011	0.015
0.2	−0.020	0.396	0.257	0.216	0.087	0.087	0.059	0.055	−0.005	0.220	0.245	0.074	0.030	0.031	0.013	0.014
0.4	−0.076	0.819	0.311	1.542	0.069	0.118	0.054	0.068	−0.056	0.265	0.280	3.676	0.037	0.041	0.017	0.017
Panel 7	$h = 1.5n^{-1/4} = 0.317$, $h_1 = .067n^{-1/9} = 0.034$								$h = 1.5n^{-1/4} = 0.178$, $h_1 = .067n^{-1/9} = 0.026$							
−0.4	−0.001	0.271	0.255	0.073	0.057	0.064	0.029	0.034	−0.000	0.276	0.270	0.076	0.052	0.057	0.027	0.030
−0.2	0.003	0.236	0.233	0.056	0.058	0.048	0.030	0.024	0.001	0.241	0.235	0.058	0.058	0.054	0.028	0.028
0	−0.000	0.210	0.210	0.044	0.052	0.043	0.029	0.021	−0.000	0.203	0.206	0.041	0.049	0.046	0.023	0.023
0.2	0.006	0.220	0.204	0.049	0.065	0.061	0.035	0.034	0.001	0.214	0.199	0.046	0.063	0.070	0.034	0.035
0.4	0.021	0.252	0.222	0.071	0.066	0.084	0.037	0.045	0.008	0.251	0.212	0.071	0.081	0.082	0.049	0.049

with a fairly wide range of bandwidths, but small choices of h_1 seem to work particular well. Results for the baseline h_1 are shown in Table 3, Panel 1 ($k = 2$) and Panel 4 ($k = 4$). In both cases, bias is minimal (even for $n = 500$) and the p-values are acceptable (though there are some instances where they are somewhat below their nominal levels). When one triples h_1 relative to baseline (Panels 3 and 6), bias starts creeping into both the estimator and the standard errors, and the p-values become somewhat attenuated. In contrast, as shown by Panels 2 and 5, cutting h_1 by a factor of three yields a textbook-perfect approximation for $n = 5000$. In fact, additional experimentation shows that making h_1 an order of magnitude smaller still works very well.

Fifth, as stated in Comment 3 after Theorem 2, theory predicts that for a given choice of h_1 and K_1, the asymptotic variance of the fully nonparametric CATE estimator is smaller than that of the semiparametric one. Comparing the results in Tables 1 versus 3 and Tables 2 versus 3 shows that in finite samples the efficiency ranking can go either way, depending also on the choice of h. Take, for example, $k = 2$ and $n = 5000$ with the baseline choice of h_1 (i.e., $h_1 = 0.078$). Though the differences are rather small, in this case the standard error of the nonparametric estimator is indeed smaller than that of the semiparametric estimator almost uniformly in h and x_1 (compare Panel 1 of Table 3 with Panels 1, 4 and 5 of Table 1). Exceptions arise only when the nonparametric estimator undersmooths the propensity score (Panel 4 of

Table 2. The distribution of $\sqrt{nh_1}[\hat{\tau}(x_1) - \tau(x_1)]$: the fully nonparametric case with $k = 4$

	$n = 500$								$n = 5000$							
x_1	$\frac{\text{Mean}}{\sqrt{nh_1}}$	s.e.	E($\widehat{\text{s.e.}}$)	MSE	p-val. 1.65	p-val. -1.65	p-val 1.96	p-val -1.96	$\frac{\text{Mean}}{\sqrt{nh_1}}$	s.e.	E($\widehat{\text{s.e.}}$)	MSE	p-val. 1.65	p-val. -1.65	p-val 1.96	p-val -1.96
Panel 1	$h = 0.5n^{-1/8} = 0.230$, $h_1 = 0.2n^{-1/13} = 0.124$								$h = 0.5n^{-1/8} = 0.172$, $h_1 = 0.2n^{-1/13} = 0.104$							
−0.4	−0.006	3.050	1.141	9.305	0.133	0.136	0.089	0.094	−0.000	0.437	0.322	0.191	0.051	0.047	0.026	0.020
−0.2	−0.035	1.862	0.837	3.543	0.159	0.059	0.099	0.045	−0.001	0.310	0.263	0.097	0.054	0.052	0.024	0.026
0	−0.002	1.430	0.653	2.044	0.059	0.047	0.035	0.028	−0.000	0.261	0.252	0.068	0.045	0.036	0.020	0.018
0.2	0.031	1.704	0.646	2.962	0.042	0.108	0.035	0.063	0.000	0.284	0.288	0.081	0.034	0.035	0.016	0.017
0.4	0.095	4.155	0.873	17.82	0.062	0.458	0.057	0.358	0.003	0.527	0.395	0.283	0.059	0.062	0.030	0.034
Panel 2	$h = 0.5n^{-1/8} = 0.230$, $h_1 = 0.067n^{-1/13} = 0.042$								$h = 0.5n^{-1/8} = 0.172$, $h_1 = 0.067n^{-1/13} = 0.035$							
−0.4	−0.010	2.671	0.999	7.139	0.076	0.057	0.043	0.032	0.000	0.342	0.284	0.117	0.052	0.050	0.025	0.026
−0.2	−0.039	1.798	0.603	3.266	0.147	0.040	0.093	0.029	−0.001	0.311	0.249	0.097	0.065	0.054	0.030	0.032
0	0.004	1.468	0.493	2.157	0.050	0.064	0.029	0.031	−0.000	0.283	0.232	0.080	0.048	0.051	0.025	0.023
0.2	0.032	1.648	0.437	2.738	0.032	0.118	0.023	0.067	0.001	0.263	0.250	0.069	0.048	0.052	0.027	0.027
0.4	0.075	3.601	0.802	13.08	0.028	0.193	0.021	0.128	0.003	0.443	0.393	0.197	0.052	0.056	0.027	0.030
Panel 3	$h = 0.5n^{-1/8} = 0.230$, $h_1 = 0.6n^{-1/13} = 0.372$								$h = 0.5n^{-1/8} = 0.172$, $h_1 = 0.6n^{-1/13} = 0.312$							
−0.4	−0.077	3.432	1.289	12.87	0.211	0.158	0.157	0.123	−0.051	0.644	0.376	4.481	0.078	0.075	0.041	0.043
−0.2	0.040	1.997	1.026	4.290	0.153	0.107	0.093	0.072	0.031	0.331	0.323	1.621	0.027	0.032	0.012	0.012
0	0.067	1.448	0.934	2.926	0.063	0.064	0.042	0.037	0.045	0.258	0.320	3.237	0.017	0.016	0.005	0.006
0.2	0.007	2.253	0.955	5.086	0.086	0.188	0.075	0.118	−0.007	0.341	0.353	0.184	0.026	0.032	0.012	0.014
0.4	−0.018	4.578	1.145	21.02	0.081	0.378	0.074	0.295	−0.056	0.770	0.458	5.450	0.059	0.069	0.031	0.033
Panel 4	$h = 0.167n^{-1/8} = 0.077$, $h_1 = 0.2n^{-1/13} = 0.124$								$h = 0.167n^{-1/8} = 0.058$, $h_1 = 0.2n^{-1/13} = 0.104$							
−0.4	−0.612	14.03	2.266	219.9	0.427	0.397	0.410	0.380	−0.403	11.84	4.111	224.6	0.294	0.288	0.257	0.248
−0.2	−1.805	13.57	2.657	386.2	0.409	0.347	0.391	0.321	−1.395	11.89	4.423	1152	0.278	0.263	0.246	0.224
0	0.001	11.52	2.224	132.7	0.387	0.382	0.365	0.361	−0.005	10.02	4.174	100.5	0.248	0.245	0.210	0.211
0.2	3.089	15.31	3.231	826.0	0.321	0.420	0.290	0.404	2.347	13.69	4.365	3048	0.291	0.317	0.252	0.283
0.4	4.212	23.66	3.962	1659	0.330	0.462	0.308	0.447	3.263	20.50	4.363	5950	0.336	0.383	0.310	0.356
Panel 5	$h = 1.5n^{-1/8} = 0.690$, $h_1 = 0.2n^{-1/13} = 0.124$								$h = 1.5n^{-1/8} = 0.517$, $h_1 = 0.2n^{-1/13} = 0.104$							
−0.4	−0.004	0.245	0.247	0.061	0.048	0.047	0.024	0.026	−0.001	0.236	0.243	0.056	0.042	0.044	0.024	0.023
−0.2	−0.001	0.239	0.240	0.057	0.055	0.042	0.025	0.019	−0.002	0.238	0.238	0.058	0.052	0.049	0.025	0.026
0	−0.001	0.228	0.249	0.052	0.040	0.035	0.018	0.017	−0.000	0.223	0.240	0.050	0.041	0.037	0.017	0.018
0.2	0.004	0.263	0.296	0.070	0.031	0.033	0.014	0.013	0.001	0.255	0.282	0.066	0.032	0.035	0.015	0.016
0.4	0.009	0.414	0.390	0.177	0.055	0.065	0.029	0.034	0.008	0.401	0.383	0.195	0.060	0.061	0.031	0.035
Panel 6	$h = 0.167n^{-1/8} = 0.077$, $h_1 = 0.6n^{-1/13} = 0.372$								$h = 0.167n^{-1/8} = 0.058$, $h_1 = 0.6n^{-1/13} = 0.312$							
−0.4	−1.411	16.92	2.853	656.7	0.412	0.383	0.395	0.363	−1.040	14.60	4.977	1897	0.289	0.283	0.259	0.245
−0.2	−0.745	13.05	2.166	273.3	0.412	0.387	0.395	0.368	−0.831	11.38	4.166	1205	0.274	0.266	0.237	0.233
0	0.609	11.97	1.977	212.3	0.388	0.409	0.368	0.389	0.335	10.46	3.915	284.1	0.264	0.266	0.224	0.231
0.2	2.252	16.55	2.743	1217	0.369	0.429	0.347	0.417	1.975	14.29	4.239	6280	0.301	0.317	0.273	0.289
0.4	3.666	26.04	4.425	3177	0.346	0.438	0.324	0.420	3.096	23.65	5.441	15494	0.334	0.376	0.307	0.346
Panel 7	$h = 1.5n^{-1/8} = 0.690$, $h_1 = 0.067n^{-1/13} = 0.041$								$h = 1.5n^{-1/8} = 0.517$, $h_1 = 0.067n^{-1/13} = 0.035$							
−0.4	−0.001	0.247	0.245	0.061	0.048	0.049	0.025	0.023	−0.001	0.246	0.245	0.060	0.052	0.054	0.025	0.027
−0.2	−0.003	0.244	0.238	0.060	0.060	0.048	0.031	0.023	−0.001	0.240	0.236	0.058	0.056	0.049	0.028	0.027
0	0.000	0.226	0.226	0.051	0.049	0.045	0.026	0.021	−0.000	0.222	0.223	0.049	0.049	0.049	0.026	0.023
0.2	0.004	0.255	0.253	0.065	0.047	0.048	0.025	0.022	0.002	0.251	0.248	0.064	0.050	0.052	0.027	0.027
0.4	0.012	0.424	0.406	0.183	0.052	0.063	0.026	0.032	0.008	0.409	0.397	0.179	0.054	0.056	0.027	0.028

Table 1, $x_1 = -0.4, -0.2$), but even in these cases the expected value of the estimated standard error is smaller for the nonparametric estimator. For $k = 4$, the results are more mixed; we see that the nonparametric estimator is more efficient only when the propensity score estimator is sufficiently smoothed, that is, when h is large.

In conclusion, while our first-order asymptotic theory is unable to guide bandwidth selection in applications, these simulations do offer a few useful pointers for practitioners: (i) for the fully nonparametric estimator, one should avoid undersmoothing the propensity score; (ii) for the semiparametric estimator undersmoothing with h_1 does not seem to be very costly while oversmoothing potentially is; (iii) estimation of CATE close to the boundary of the support of X_1 is by all means problematic.

4. EMPIRICAL ILLUSTRATION

As documented by a number of authors, low birth weight is associated with increased health care costs during infancy as well as adverse health, educational and labor market outcomes later in life; see, for example, Abrevaya (2006) or Almond, Chay, and Lee (2005) for a set of references. Smoking is generally regarded as one of the major modifiable risk factors for low birth weight, and there are many studies that attempt to estimate

Table 3. The distribution of $\sqrt{nh_1}[\hat{\tau}(x_1) - \tau(x_1)]$: the semiparametric case

	$n = 500$								$n = 5000$							
x_1	$\frac{\text{Mean}}{\sqrt{nh_1}}$	s.e.	E(ŝ.e.)	MSE	p-val. 1.65	p-val. −1.65	p-val 1.96	p-val −1.96	$\frac{\text{Mean}}{\sqrt{nh_1}}$	s.e.	E(ŝ.e.)	MSE	p-val. 1.65	p-val. −1.65	p-val 1.96	p-val −1.96
Panel 1	$h_1 = 0.2n^{-1/9} = 0.100$					$k = 2$			$h_1 = 0.2n^{-1/9} = 0.078$					$k = 2$		
−0.4	−0.004	0.302	0.296	0.092	0.053	0.059	0.023	0.029	−0.000	0.294	0.291	0.087	0.051	0.054	0.025	0.027
−0.2	0.003	0.241	0.244	0.059	0.049	0.046	0.025	0.019	0.001	0.237	0.242	0.057	0.047	0.045	0.023	0.024
0	0.001	0.209	0.220	0.044	0.042	0.038	0.025	0.018	0.000	0.205	0.213	0.042	0.044	0.044	0.020	0.023
0.2	0.001	0.216	0.232	0.047	0.031	0.044	0.014	0.022	0.000	0.215	0.226	0.046	0.044	0.044	0.019	0.021
0.4	−0.006	0.233	0.257	0.056	0.029	0.039	0.013	0.021	−0.001	0.229	0.250	0.053	0.036	0.037	0.015	0.017
Panel 2	$h_1 = 0.067n^{-1/9} = 0.034$					$k = 2$			$h_1 = 0.067n^{-1/9} = 0.026$					$k = 2$		
−0.4	0.002	0.300	0.295	0.090	0.046	0.053	0.025	0.026	0.000	0.298	0.296	0.089	0.049	0.050	0.025	0.026
−0.2	0.000	0.248	0.240	0.062	0.061	0.049	0.032	0.020	0.001	0.239	0.241	0.057	0.046	0.049	0.023	0.022
0	0.000	0.204	0.204	0.042	0.049	0.050	0.023	0.024	0.000	0.201	0.204	0.040	0.046	0.048	0.022	0.023
0.2	0.001	0.218	0.216	0.047	0.048	0.058	0.024	0.029	−0.000	0.212	0.214	0.045	0.047	0.047	0.024	0.023
0.4	−0.000	0.241	0.252	0.058	0.045	0.045	0.022	0.023	−0.000	0.250	0.252	0.063	0.047	0.049	0.023	0.027
Panel 3	$h_1 = 0.6n^{-1/9} = 0.301$					$k = 2$			$h_1 = 0.6n^{-1/9} = 0.233$					$k = 2$		
−0.4	−0.043	0.339	0.324	0.390	0.058	0.059	0.029	0.029	−0.037	0.341	0.314	1.739	0.067	0.068	0.036	0.036
−0.2	0.062	0.246	0.270	0.640	0.035	0.032	0.015	0.015	0.034	0.239	0.257	1.369	0.041	0.036	0.018	0.017
0	0.063	0.216	0.258	0.648	0.023	0.027	0.010	0.010	0.038	0.213	0.251	1.749	0.027	0.025	0.011	0.009
0.2	−0.032	0.239	0.273	0.209	0.027	0.031	0.011	0.014	−0.007	0.224	0.261	0.109	0.027	0.028	0.011	0.010
0.4	−0.105	0.282	0.329	1.722	0.023	0.032	0.010	0.014	−0.060	0.269	0.309	4.252	0.026	0.030	0.012	0.014
Panel 4	$h_1 = 0.2n^{-1/13} = 0.124$					$k = 4$			$h_1 = 0.2n^{-1/13} = 0.104$					$k = 4$		
−0.4	−0.002	0.251	0.251	0.063	0.051	0.052	0.024	0.026	−0.000	0.244	0.246	0.059	0.048	0.047	0.024	0.023
−0.2	0.001	0.237	0.236	0.056	0.054	0.047	0.030	0.026	−0.000	0.241	0.240	0.058	0.053	0.054	0.024	0.026
0	0.000	0.225	0.217	0.051	0.055	0.059	0.032	0.034	0.000	0.222	0.240	0.049	0.039	0.037	0.017	0.016
0.2	0.001	0.264	0.251	0.070	0.054	0.067	0.026	0.038	−0.001	0.262	0.292	0.069	0.032	0.038	0.014	0.016
0.4	−0.003	0.411	0.418	0.169	0.046	0.048	0.022	0.026	−0.001	0.396	0.395	0.158	0.051	0.054	0.023	0.027
Panel 5	$h_1 = 0.067n^{-1/13} = 0.041$					$k = 4$			$h_1 = 0.067n^{-1/13} = 0.035$					$k = 4$		
−0.4	0.001	0.250	0.244	0.063	0.061	0.056	0.032	0.031	−0.001	0.249	0.248	0.062	0.048	0.049	0.027	0.026
−0.2	0.000	0.241	0.234	0.058	0.057	0.051	0.035	0.029	0.000	0.240	0.239	0.058	0.050	0.047	0.025	0.023
0	0.001	0.223	0.215	0.050	0.061	0.064	0.031	0.033	0.000	0.218	0.222	0.047	0.048	0.049	0.024	0.024
0.2	0.001	0.256	0.251	0.065	0.048	0.065	0.022	0.036	−0.000	0.252	0.259	0.063	0.043	0.050	0.020	0.025
0.4	−0.000	0.412	0.399	0.170	0.050	0.066	0.028	0.036	−0.001	0.410	0.411	0.168	0.048	0.054	0.021	0.025
Panel 6	$h_1 = 0.6n^{-1/13} = 0.372$					$k = 4$			$h_1 = 0.6n^{-1/13} = 0.312$					$k = 4$		
−0.4	−0.056	0.303	0.256	0.682	0.080	0.086	0.047	0.055	−0.050	0.302	0.306	4.013	0.048	0.048	0.025	0.024
−0.2	0.057	0.236	0.229	0.665	0.059	0.053	0.030	0.027	0.032	0.238	0.281	1.632	0.027	0.026	0.010	0.011
0	0.064	0.233	0.268	0.824	0.030	0.033	0.012	0.012	0.045	0.232	0.295	3.198	0.018	0.021	0.006	0.006
0.2	−0.029	0.307	0.332	0.253	0.031	0.042	0.013	0.021	−0.008	0.291	0.337	0.191	0.027	0.031	0.008	0.012
0.4	−0.093	0.454	0.437	1.801	0.052	0.062	0.026	0.031	−0.059	0.453	0.445	5.663	0.049	0.058	0.024	0.028

its average causal effect. As we point out in the introduction, our goal is not to provide another estimate of the average effect per se, but rather to illustrate how to explore the heterogeneity of this effect across subpopulations defined by the values of some continuous covariates. In particular, we will designate mother's age as X_1, that is, we are interested in estimating how the expected smoking effect changes with age, while averaging out all other confounders. We focus on the semiparametric estimator primarily because of the large number of covariates needed to make the unconfoundedness assumption plausible.

4.1 The Dataset and the Identification Strategy

We use a dataset of vital statistics recorded between 1988 and 2002 by the North Carolina State Center Health Services, accessible through the Odum Institute at the University of North Carolina. We focus on first-time mothers, a restriction which we will motivate shortly. As routine in the literature, we treat blacks and whites as separate populations throughout. The number of observations is 157,989 for the former group and 433,558 for the latter.

In accordance with our theory, the key identifying assumption is that the potential birth weight outcomes are independent of the smoking decision conditional on a sufficiently rich vector of observables X. Almond, Chay, and Lee (2005), da Veiga and Wilder (2008), and Walker, Tekin, and Wallace (2009) also used variants of the unconfoundedness assumption to identify the average effect of smoking on birth weight. Specifically, we include into X the mother's age, education, month of first prenatal visit (=10 if prenatal care is foregone), number of prenatal visits, and indicators for the baby's gender, the mother's marital status, whether or not the father's age is missing, gestational diabetes, hypertension, amniocentesis, ultra sound exams, previous (terminated) pregnancies, and alcohol use.

By contrast, Abrevaya (2006) and Abrevaya and Dahl (2008) controlled for unobserved heterogeneity by using a panel of

mothers with multiple births. Nevertheless, their approach still imposes restrictions on the channels through which unobservables are allowed to operate. In particular, there cannot be feedback from unobserved factors affecting the birth weight of the first child to the decision whether or not to smoke during the second pregnancy. This is likely violated in practice. The presence of such feedback also affects the plausibility of the unconfoundedness assumption; our focus on first births is an attempt to deal with this problem (we cannot identify mothers with multiple births in the data).

4.2 Implementation Issues

We estimate the CATE (mother's age) function over a grid between the first and fourth quintile of the age distribution. In particular, for blacks the corresponding age-range is 17.4–26.0 years and for whites it is 19.6–29.9 years. In both cases CATE is estimated at 50 equally spaced grid points in these intervals. (While it is tempting to estimate CATE further out in the tails of the age distribution, the simulation results suggest that estimates can be very unreliable closer to the boundary.)

We use a logit model based on a linear index to estimate the propensity score. More precisely, the index is constrained to be linear in the parameters, but we allow the components of X to enter through more flexible functional forms. To obtain sensible results, it is particularly important to account for the apparent nonlinearity of the propensity score with respect to age. Specifically, in addition to X itself, the benchmark model also incorporates a constant, the square of the mother's age, and cross products between mother's age and all other variables in X.

The choice of the smoothing parameter h_1 is another critical issue in practice. As is typical for estimators with a nonparametric component, the choice of the smoothing parameter drives a finite sample bias-variance tradeoff. Smaller values of h_1 produce more variable, often nonmonotonic, CATE estimates with a wider, sometimes even extreme, range. On the other hand, very large values of h_1 force the estimated CATE function to be essentially constant; in this case, our estimator can be thought of as an implementation of the Hirano, Imbens, and Ridder (2003) ATE estimator.

Using a grid search, we calibrate h_1 in a way that the resulting CATE estimates are "reasonable" in that they correspond well to previous estimates of ATE. Specifically, by the law of iterated expectations, the average of CATE estimates computed at each sample observation is an implicit estimate of ATE. Typical ATE estimates obtained under some form of the unconfoundedness assumption (OLS, propensity score matching) range from about -120 to -250 grams; see, for example, Abrevaya (2006), da Veiga and Wilder (2008), and Walker, Tekin, and Wallace (2009). Note, however, that no study known to us restricts attention to first time mothers.

A further consideration is to prevent the estimated CATE function from taking on values that appear extreme in light of prior knowledge. The occurrence of such values, particularly for small h_1, is due in part to propensity score estimates that are close to zero. As advocated by Crump et al. (2009), we therefore drop the observations with propensity score estimates outside the interval $[\alpha, 1-\alpha]$ for a suitably small value of α. In our benchmark estimations we pick α using the data-driven method proposed by Crump et al. (2009), which gives $\alpha = 0.03$ for blacks (about 20% of the observations dropped), and $\alpha = 0.08$ for whites (about 33% of the observations dropped).

Compared with the choice of the smoothing parameter and trimming, the specification of the kernel function K_1 is a second-order issue. We use a regular Gaussian kernel (a kernel with unbounded support) rather than the Epanechnikov (a kernel with bounded support) mainly because the former gives smoother, nicer looking estimates that are somewhat less sensitive to the bandwidth.

4.3 Estimation Results and Some Robustness Checks

Our benchmark estimates are depicted in Figure 2 (blacks) and Figure 3 (whites). We report four different CATE estimates corresponding to bandwidths $h_1 = 0.25\hat{\sigma}, 0.5\hat{\sigma}, 1\hat{\sigma}, 5\hat{\sigma}$, where $\hat{\sigma}$ is the sample standard deviation of X_1 (mother's age). In both figures the top panel shows the four estimates together; the bottom panels show them separately along with ± 2 standard errors. As h_1 increases, the estimated CATE functions become flatter, less variable, and their range shrinks. The CATE estimate corresponding to the smallest value of h_1 actually shows a positive and somewhat significant smoking effect around age 19–20 for both races; increasing the bandwidth causes this anomaly to disappear. While more smoothing might guard against CATE estimates with unreasonable values, it also goes against the general philosophy of the exercise—to let the data speak about the heterogeneity of the smoking effect with as mild restrictions as possible. Indeed, for $h_1 = 5\hat{\sigma}$, the estimates are only informative about the overall average effect, estimated to be about -155 grams for blacks and -190 for whites.

Taken globally, all the nonconstant CATE estimators tell a similar story—the predicted average effect of smoking, by and large, becomes stronger (more negative) at higher ages. For blacks, this monotone relationship is possibly broken by a hump around age 19–20, and for whites by a "check mark" shape in the late 20s. However, these features become less pronounced with more smoothing. The overall negative slope is consistent with the results obtained by Walker, Tekin, and Wallace (2009), who reported OLS and matching estimates that are more negative for adult mothers than for teen mothers. The advantage of our approach over simple and somewhat arbitrary sample splitting is that CATE is potentially capable of delivering a more detailed picture of age-related heterogeneity. A natural but speculative explanation for the results is that age is positively correlated with how long an individual has smoked, and long-term smoking has a cumulative negative effect on birth weight. Another possibility is that the correlation between smoking and other unobserved harmful behaviors that are not picked up by our controls becomes stronger with age.

For a given age, the numerical values of the CATE point estimates can vary substantially with h_1. For example, for blacks at age 26 the estimated effect ranges from about -150 g to -480 g, a factor of 3. Similarly, if we consider, say, whites at age 20, we again see differences in point estimates on the order of 300 g, despite the fact that the estimated confidence intervals do not suggest such large sampling variation. Again,

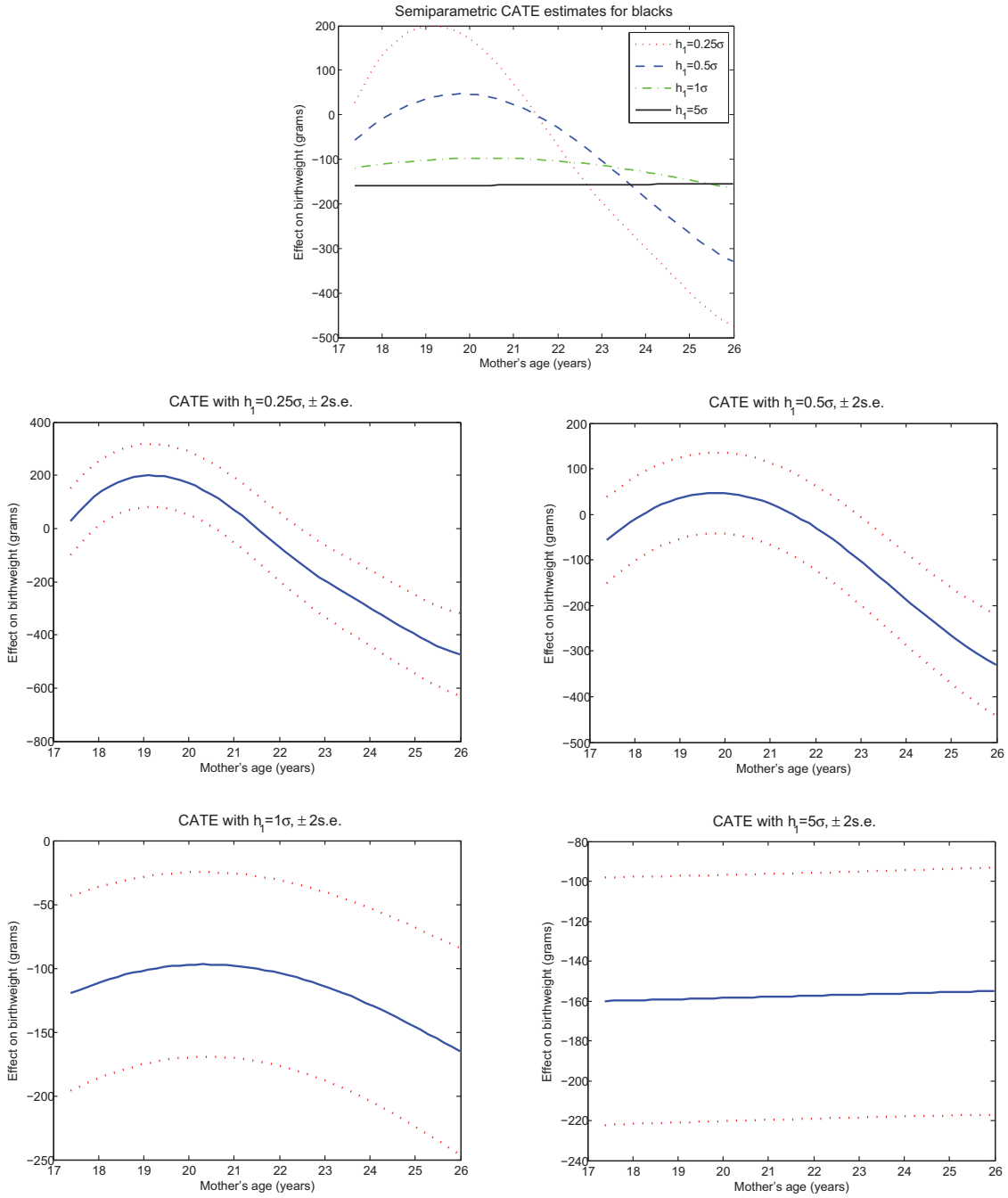

Figure 2. Baseline estimates of CATE (mother's age) for blacks.

this underscores the need for sensitivity analysis and data-driven procedures to guide bandwidth choice in finite samples.

We perform a number of robustness checks in addition to varying h_1. First, we estimate CATE using the linear regression method proposed in Section 2.2; the results are displayed in Figure 4. Consistent with the overall shape of the semiparametric estimates, the regression-based CATE functions also have a negative slope. In fact, while a bit steeper, the semiparametric estimates for $h_1 = 1 \times \hat{\sigma}$ are reasonably close to the OLS estimates for both groups.

Regarding the sufficiency of our controls, the level differences in the estimates for blacks versus whites could be interpreted as evidence that confounding factors are not completely accounted for by the conditioning variables (we are grateful to a referee for pointing this out). Nevertheless, it is not straightforward to pinpoint the omitted factors. A potential concern could be the decline in smoking incidence during the sample period (or that it might have taken place differently among the two groups). However, adding "year fixed effects" (i.e., birth year) to X does not change the baseline results in a noteworthy way. As the North Carolina vital statistics include the mother's zip code, we can link some zip-code level characteristics, such as annual mean income, to individuals. Adding such controls does not appreciably change the results either.

Varying the point at which the propensity score estimates are trimmed has a more substantial effect. For example, setting $\alpha = 2\%$ implies dropping only 4% of observations for whites and 6.5% for blacks. The general shape of the estimated CATE

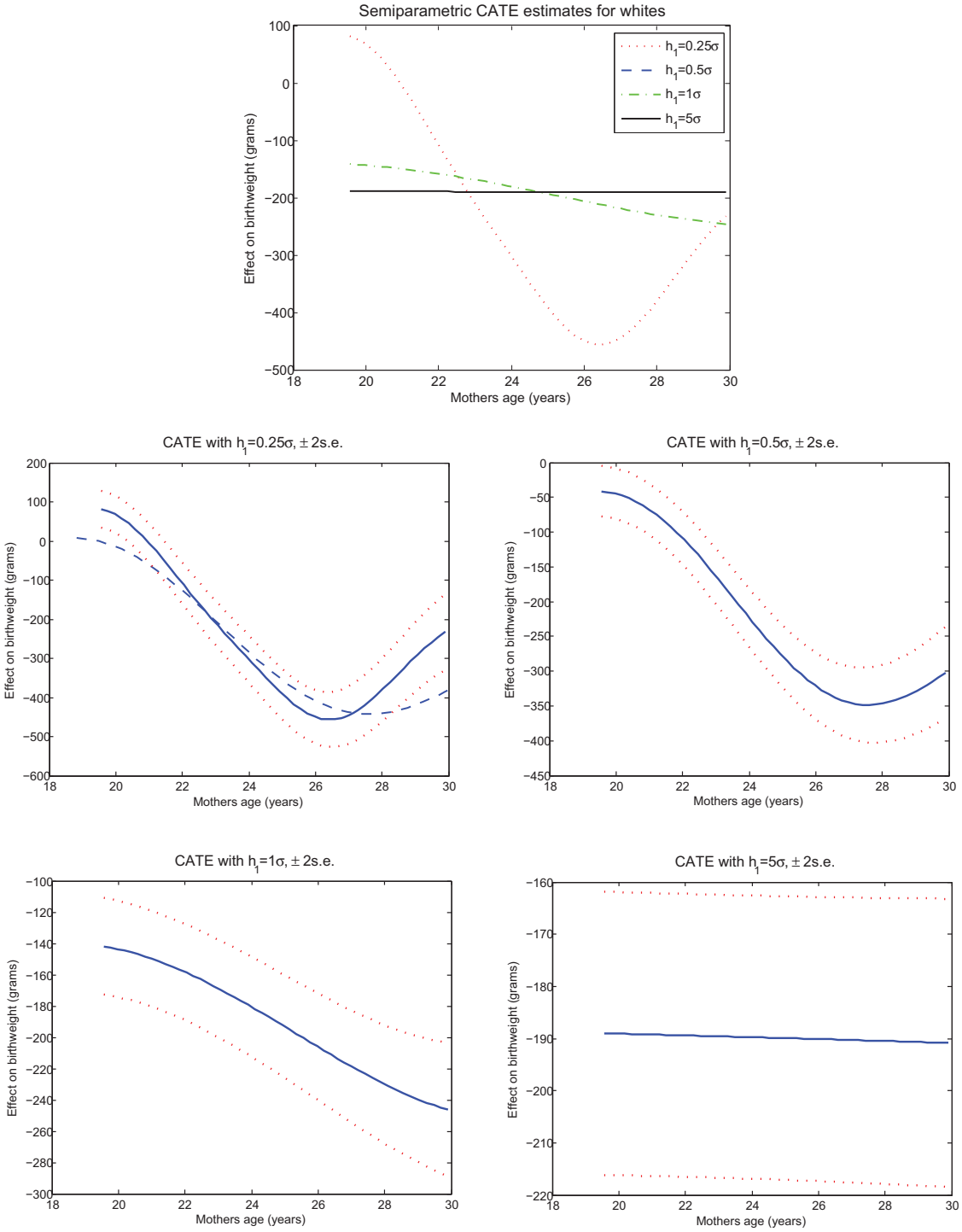

Figure 3. Baseline estimates of CATE (mother's age) for whites.

functions remains similar for all bandwidths, but their means shift downward by about 40 g for blacks and 180 g for whites. For small bandwidths, and particularly for whites, the range of the estimated effect also expands, pushing the boundaries of credibility.

5. THEORETICAL EXTENSIONS

In this section we discuss several extensions of our theory. First, we define the conditional average treatment effect for the treated (CATT) and state the analog of Theorem 1 for this parameter. Second, we allow for selection on unobservables and consider suitable estimands in an instrumental variable (IV) framework. In particular, if the unconfoundedness assumption is violated, but there exists a valid binary instrument, we define the conditional local average treatment effect (CLATE) and the conditional local average treatment effect for the treated (CLATT), and extend our theory to these parameters. Finally, in cases where the treatment is multivalued and unconfoundedness holds, we extend our theory to the conditional marginal average treatment effect (CMATE).

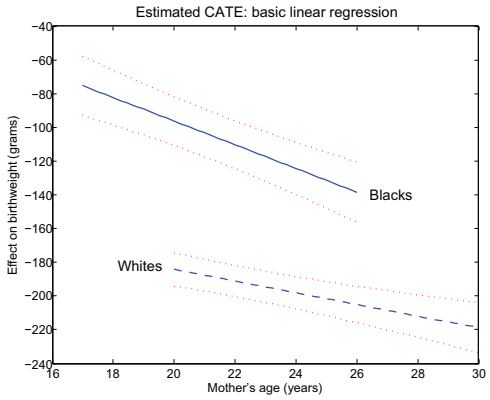

Figure 4. OLS estimates of CATE (mother's age).

5.1 Conditional Average Treatment Effect of the Treated

The average treatment effect for the treated (ATT) is often more relevant for policy making than the average effect for the entire population. Of course, individual treatment effects might be heterogeneous within the treated subpopulation as well. This motivates defining the conditional average treatment effect (CATT) as

$$\text{CATT}(x_1) \equiv \tau_t(x_1) \equiv E[Y(1) - Y(0)|D = 1, X_1 = x_1],$$

where $X_1 \in \mathbb{R}^\ell$ is a (continuous) subvector of $X \in \mathbb{R}^k$. It can be shown that $\tau(x_1)$ can be identified as

$$\tau_t(x_1) = E\left[DY - \frac{p(X)(1-D)Y}{1-p(X)}\middle| X_1 = x_1\right] \Big/ E[p(X)|X_1 = x_1], \tag{11}$$

which suggests the estimator

$$\hat{\tau}_t(x_1) = \frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n \left(D_i Y_i - \frac{\hat{p}(X_i)(1-D_i)Y_i}{1-\hat{p}(X_i)}\right) K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n \hat{p}(X_i) K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}, \tag{12}$$

where $\hat{p}(X_i)$ is again given by the leave-i-out version of (3).

The following theorem is analogous to Theorem 1 and summarizes the first-order asymptotic of $\hat{\tau}_t(x_1)$.

Theorem 3. Suppose that Assumptions 1 through 8 are satisfied. Then, for each point x_1 in the support of X_1,

$$\sqrt{nh_1^\ell}(\hat{\tau}_t(x_1) - \tau_t(x_1)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\|K_1\|_2^2 \sigma_{\psi_t}^2(x_1)}{f_1(x_1)}\right),$$

where $\sigma_{\psi_t}^2(x_1) \equiv E[\psi_t^2(X, Y, D)|X_1 = x_1]$ with

$$\psi_t(x, y, d) \equiv \frac{1}{p_{x_1}}\left(d(y - m_1(x)) - \frac{p(x)(1-d)(y - m_0(x))}{1-p(x)} + d(m_1(x) - m_0(x) - \tau_t(x_1))\right),$$

$$p_{x_1} \equiv E[D = 1|X_1 = x_1].$$

Comments

1. The influence function $\psi_t(x, y, d)$ is analogous to the influence function that efficient nonparametric estimators of ATT possess; see, for example, Hahn (1998) and Hirano, Imbens, and Ridder (2003).
2. Some comments about the proof of Theorem 3 are provided in Appendix F. The details of the derivations are similar to the proof of Theorem 1 and are omitted.
3. The semiparametric results (i.e., Theorem 2) can also be extended in a straightforward manner. In this case the influence function is analogous to the influence function that Hahn (1998) and Hirano, Imbens, and Ridder (2003) derived for their ATT estimators in the case when $p(x)$ is known.

5.2 Endogenous Treatment Assignment

We will now relax the unconfoundedness assumption and allow for selection to treatment based on unobserved confounders. We will, therefore, need an instrument to control for selection bias. In a pure nonparametric framework (C)ATE and (C)ATT are no longer identified by an instrument alone; we will rather extend our theory to the local average treatment effect (LATE) and the local average treatment effect for the treated (LATT).

The following IV framework, augmented by covariates, is now standard in the treatment effect literature; see, for example, Abadie (2003), Frolich (2007), Hong and Nekipelov (2010), and Donald, Hsu, and Lieli (2014a). In addition to Y, D, and X, we observe the value of a binary instrument $Z \in \{0, 1\}$ for each individual in the sample. For $Z = z$, the random variable $D(z) \in \{0, 1\}$ specifies individuals' potential treatment status with $D(z) = 1$ corresponding to treatment and $D(z) = 0$ to no treatment. The actually observed treatment status is then given by $D \equiv D(Z) = D(1)Z + D(0)(1 - Z)$. The following assumptions, taken from Donald, Hsu, and Lieli (2014a) with some modifications, describe the relationships between the variables defined above.

Assumption 10.

(i) (Instrument Validity): $\big(Y(0), Y(1), D(1), D(0)\big) \perp Z\big|X$.
(ii) (First stage): $P[D(1) = 1 \mid X] > P[D(0) = 1 \mid X]$ and $0 < P(Z = 1 \mid X) < 1$.
(iii) (Monotonicity): $P[D(1) \geq D(0)] = 1$.

Assumption 10(i) is the analog of the unconfoundedness assumption in the IV framework—it requires that conditional on X, Z is independent of the potential outcomes and the potential treatment status. Part (ii) postulates that the instrument is (positively) related to the probability of being treated and implies that the distributions $X|Z = 0$ and $X|Z = 1$ have common support. Finally, the monotonicity of $D(z)$ in z, required in part (iii), implies that there are no defiers $[D(0) = 1, D(1) = 0]$ in the population.

We define the conditional local average treatment effect (CLATE) and the conditional local average treatment effect of the treated (CLATT) parameters as

$$\text{CLATE}(x_1) \equiv E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, X_1 = x_1]$$

$$\text{CLATT}(x_1) \equiv E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, D = 1, X_1 = x_1].$$

Following Donald, Hsu, and Lieli (2014a), we can show that CLATE(x_1) and CLATT(x_1) are identified by

$$\gamma(x_1) = E\left[\frac{ZY}{q(X)} - \frac{(1-Z)Y}{1-q(X)}\middle| X_1 = x_1\right] \Big/ E\left[\frac{ZD}{q(X)} - \frac{(1-Z)D}{1-q(X)}\middle| X_1 = x_1\right],$$

$$\gamma_t(x_1) = E\left[ZY - \frac{q(X)(1-Z)Y}{1-q(X)}\middle| X_1 = x_1\right] \Big/ E\left[ZD - \frac{q(X)(1-Z)D}{1-q(X)}\middle| X_1 = x_1\right],$$

where $q(x) \equiv P[Z = 1|X = x]$. That is, the CLATE (CLATT) is identified by the CATE (CATT) of Z on Y over the CATE (CATT) of Z on D. Therefore, we can use the CATE and CATT estimators developed in previous sections to estimate the numerators and denominators of CLATE and CLATT. Under regularity conditions similar to those in Theorems 1 and 2, one can obtain the asymptotic properties of the CLATE and CLATT estimators by the delta method. We omit the formal statements.

5.3 Multivalued Treatment

In this section, we consider multivalued (rather than binary) treatments. For example, Walker, Tekin, and Wallace (2009) and Cattaneo (2010) further divide the smoking indicator into several groups depending on the intensity of the smoking. In particular, Walker, Tekin, and Wallace (2009) consider four groups: nonsmokers, smokers with tobacco use between 1 and 10 cigarettes per day, smokers with tobacco use between 11 and 20 cigarettes per day, and smokers with tobacco use more than 20 cigarettes per day. On the other hand, Cattaneo (2010) divides smokers into five groups depending on the daily tobacco use: 1–5, 6–10, 11–15, 16–20, and 20+. By doing this, one can study the effect of maternal smoking intensity on birth weight in detail.

We introduce the model and the notation following Cattaneo (2010). Treatment status (categorical or ordinal) is indexed by $t \in \{0, 1, \ldots, J\} \equiv \mathcal{T}$ with $J \in \mathbb{N}$ fixed. For given $t \in \mathcal{T}$, let $Y(t)$ be the potential outcome under treatment level t. Let the random variable $T \in \mathcal{T}$ indicate which of the $J + 1$ potential outcomes is observed and let $D_t \equiv 1(T = t)$ for all $t \in \mathcal{T}$, where $1(\cdot)$ is the indicator function. The observed outcome Y is given by $\sum_{t\in\mathcal{T}} D_t Y(t)$. The parameters considered in Cattaneo (2010) are the marginal average treatment effects (MATE): $E[Y(t) - Y(s)]$ for all $t, s \in \mathcal{T}$. Just as before, we can define the conditional marginal average treatment effect (CMATE) given $X_1 = x_1$ as: $E[Y(t) - Y(s)|X_1 = x_1]$ for all $t, s \in \mathcal{T}$.

Define the generalized propensity score as $p_t(x) \equiv P(D_t = 1|X = x)$ for $t \in \mathcal{T}$. The following assumption is the main assumption we need to identify the MATE or CMATE:

Assumption 11 (i) (Unconfoundedness Assumption): For all $t \in T$, $Y(t) \perp D_t|X$. (ii) (Generalized Propensity Score): For all $t \in T$, $p_t(x) \geq \delta > 0$ on $\mathcal{X}$.

Assumption 11 is similar to the binary treatment case, so we omit the discussion. Under Assumption 11, $E[Y(t)|X_1 = x_1]$ is identified as

$$E[Y(t)|X_1 = x_1] = E\left[\frac{D_t Y}{p_t(X)}\middle| X_1 = x_1\right],$$

and $E[Y(t)|X_1 = x_1] \equiv \lambda_t(x_1)$ can be estimated by

$$\hat{\lambda}_t(x_1) = \frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n \left(\frac{D_{ti}Y_i}{\hat{p}_t(X_i)}\right) K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)},$$

where $\hat{p}_t(X_i)$ is the leave-i-out Nadaraya–Watson estimator for $p_t(X_i)$ as in (3). Under similar regularity conditions as in Theorem 1, one can obtain the asymptotic properties of the CMATE estimator and we omit the formal statements. Alternatively, one can model the generalized propensity score parametrically for each t and extend the semiparametric results in a straightforward manner.

6. CONCLUSIONS

We study the estimation of the conditional average treatment effect (CATE), a functional parameter designed to capture the variation in the average treatment effect conditional on some covariate(s) used in identifying it. We propose inverse probability weighted estimators of this function and provide pointwise first-order asymptotic theory. We also discuss extensions to multivalued treatments, instrumental variable frameworks, etc. Using the semiparametric version of the estimator, we estimate the average effect of a first time mother's smoking on her baby's birth weight conditional on the mother's age. The main qualitative finding is that smoking has a more negative impact at higher ages. Nevertheless, the numerical estimates are rather sensitive to bandwidth choice and, to a variable extent, some aspects of the specification and trimming of the estimated propensity score function. Overall, the sempiarametric CATE estimator appears to be a promising practical tool for exploring the heterogeneous effects of a treatment, but the empirical exercise highlights the need for formal, preferably data-driven, criteria for bandwidth choice. We consider this an important problem for future research.

APPENDIX

A. Deriving Standard Errors for the Parametric CATE Estimator

Define $y_i \equiv (Y_i, (X_i - \bar{X})')'$, $\hat{u}_i \equiv (\hat{\epsilon}_i, \hat{u}_i^{(1)}, \ldots, \hat{u}_i^{(k)})'$, both $(k + 1) \times 1$ vectors, and

$$Z_i' \equiv \begin{pmatrix} (1, D_i, X_i', D_i(X_i - \bar{X})') & 0 & \ldots & 0 \\ 0 & \tilde{X}_{1i}' & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \tilde{X}_{1i}' \end{pmatrix},$$

a $(k + 1) \times (2 + 2k + k(\ell + 1))$ matrix. Furthermore, collect the OLS coefficient estimates in (5) and (6) into the $(2 + 2k + k(\ell + 1)) \times 1$ vector

$$\hat{\theta} \equiv (\hat{c}, \hat{\alpha}, \hat{\beta}', \hat{\delta}', \hat{\gamma}^{(1)\prime}, \ldots, \hat{\gamma}^{(k)\prime})'.$$

With these definitions, (5) and (6) can be represented as the SUR system $y_i = Z_i'\hat{\theta} + \hat{u}_i$. Furthermore, it is easy to see that $\hat{\theta}$ coincides with the system OLS estimator, that is, $\hat{\theta} = (\sum_{i=1}^n Z_i Z_i')^{-1} \sum_{i=1}^n Z_i y_i$.

By standard asymptotic results for iid data (see, e.g., Thm. 7.2 of Wooldridge 2010),

$$(\hat{A}^{-1}\hat{B}\hat{A}^{-1})^{-1/2}\sqrt{n}(\hat{\theta}-\theta) \to_d N(0, I), \tag{A.1}$$

where $\theta \equiv \operatorname{plim}\hat{\theta}$, $\hat{A} \equiv n^{-1}\sum_{i=1}^n Z_i Z_i'$, $\hat{B} \equiv n^{-1}\sum_{i=1}^n Z_i\hat{u}_i\hat{u}_i' Z_i'$, and I is the identity matrix conformable with $\hat{\theta}$. (Note that this result allows for arbitrary correlations between the components of $u_i = \operatorname{plim}\hat{u}_i$ as well as heteroscedasticity.) For x_1 fixed, let $g(\hat{\theta}) \equiv \hat{\alpha} + (\tilde{x}_1'\hat{\gamma})\hat{\delta}$ denote the CATE estimator in (7). It is easy to verify that the gradient of g is given by

$$\nabla g(\hat{\theta}) = (0_{(1\times 1)}, 1, 0_{(1\times k)}, \tilde{x}_1'\hat{\gamma}^{(1)}, \ldots, \tilde{x}_1'\hat{\gamma}^{(k)}, \hat{\delta}_1\tilde{x}_1', \ldots, \hat{\delta}_k\tilde{x}_1')'.$$

It follows from (A.1) and the delta method that for large n, the variance of $\hat{\alpha} + (\tilde{x}_1'\hat{\gamma})$ is approximately $\nabla g(\hat{\theta})'\hat{A}^{-1}\hat{B}\hat{A}^{-1}\nabla g(\hat{\theta})/n$.

B. Properties of $\hat{p}(\cdot)$

First we establish some properties of the proposed propensity score estimator needed to show Theorem 1. For $i = 1, \ldots, n$, write

$$\hat{p}(X_i) = \sum_{j:j\neq i} \omega_{ij} Y_j, \tag{A.2}$$

where

$$\omega_{ij} \equiv \frac{\frac{1}{nh^k}K\left(\frac{X_i - X_j}{h}\right)}{\frac{1}{nh^k}\sum_{t:t\neq i}K\left(\frac{X_i - X_t}{h}\right)}$$

depends on $X_1, \ldots, X_n$ only.

Lemma A.1. Given Assumptions 1 through 8, the propensity score estimator satisfies

$$(a) \quad |\omega_{ij} - \omega_{ji}| \le \frac{C_n}{nh^k}\left|K\left(\frac{X_i - X_j}{h}\right)\right|, \tag{A.3}$$

where $C_n = O_p(h)$ and does not depend on i, j. Furthermore,

$$(b) \quad \sup_{x\in\mathcal{X}}|\hat{p}(x) - p(x)| = O_p\left(h^s + \sqrt{\frac{\log n}{nh^k}}\right), \tag{A.4}$$

and, in particular, $E[\hat{p}(x) \mid X_1, \ldots, X_n] - p(x) = O_p(h^s)$ uniformly in x.

Proof. To see part (a), first note that by Assumption 7(i), $\omega_{ij} = \omega_{ji} = 0$ for $\|X_j - X_i\|_\infty > h$. Now assume $\|X_j - X_i\|_\infty \le h$. For all i, define

$$\hat{f}(X_i) \equiv \frac{1}{nh^k}\sum_{t:t\neq i}K\left(\frac{X_i - X_t}{h}\right).$$

Then

$$\begin{aligned}
|\omega_{ij} - \omega_{ji}| &= \frac{1}{nh^k}\left|K\left(\frac{X_i - X_j}{h}\right)\right| \cdot \left|\hat{f}^{-1}(X_i) - \hat{f}^{-1}(X_j)\right| \\
&\le \frac{1}{nh^k}\left|K\left(\frac{X_i - X_j}{h}\right)\right| \cdot \left\{\left|\hat{f}^{-1}(X_i) - f^{-1}(X_i)\right| \right. \\
&\quad + \left|\hat{f}^{-1}(X_j) - f^{-1}(X_j)\right| \\
&\quad \left. + \left|f^{-1}(X_i) - f^{-1}(X_j)\right|\right\}.
\end{aligned} \tag{A.5}$$

We will bound the three terms in the braces, uniformly in i, j. Using standard arguments in nonparametric estimation (similar to the proof of, e.g., Corollary 1 of Masry 1996), it is possible to show that, under the maintained assumptions,

$$\sup_i|\hat{f}(X_i) - f(X_i)| = O_p\left(h^s + \sqrt{\frac{\log n}{nh^k}}\right). \tag{A.6}$$

As $s \ge k \ge 2$, Assumption 8(i) implies that the quantity in (A.6) is $o_p(h)$. By the mean value theorem,

$$\sup_i|\hat{f}^{-1}(X_i) - f^{-1}(X_i)| \le \sup_i\frac{1}{\tilde{f}_i^2}\sup_i|\hat{f}(X_i) - f(X_i)|, \tag{A.7}$$

where $\tilde{f}_i$ is a quantity between $\hat{f}(X_i)$ and $f(X_i)$. Since f is bounded away from zero, $\sup_i \tilde{f}_i^{-2} = O_p(1)$, and so the rhs of (A.7) is $o_p(h)$. Obviously, the same argument applies to the second term. As for the last term, the mean value theorem and the fact that f is continuously differentiable on its compact support and bounded away from zero implies $|f^{-1}(x_1) - f^{-1}(x_2)| \le M\|x_1 - x_2\|_\infty$ for all $x_1, x_2 \in \mathcal{X}$ and some constant $M > 0$. Hence, $|f^{-1}(X_i) - f^{-1}(X_j)| = O(h)$ given $\|X_j - X_i\|_\infty \le h$. Combining these observations yields (A.3), where C_n can be taken as Mh plus twice the lhs (or rhs) of (A.7).

Part (b) is a standard result in kernel based nonparametric regression theory; see, for example, Pagan and Ullah (1999) and Su (2011). The term h^s is the leading term in the expansion of the bias of $\hat{p}(x)$ and $1/(nh^k)$ is the leading term in the variance expansion (the $\log(n)$ factor arises if one requires uniform convergence). Note also that the bias of the estimator conditional on $X_1, \ldots, X_n$ is also of order $O(h_s)$ uniformly in x. □

C. The Proof of Theorem 1

For purposes of exposition only, we present the proof with $\ell = 1$. All arguments remain valid in the general case with minimal and straightforward modifications. In this section we use the letter M to denote a generic positive constant whose value can change from context to context.

Expanding $\hat{\tau}(x_1)$. We build on the expansion of the Hirano, Imbens, and Ridder (2003) estimator by Ichimura and Linton (2005). We start by establishing notation. Let $w = (y, d, x)$, $\tau = \tau(x_1)$, and

$$\Psi(w, p) \equiv \frac{dy}{p} - \frac{(1-d)y}{1-p}.$$

The first and second partial derivatives of Ψ w.r.t. the argument p are denoted as Ψ_p and Ψ_{pp}, respectively.

We further define

$$\begin{aligned}
S_p(X_i) &\equiv E[\Psi_p(W_i, p(X_i))|X_i] \\
&= -\left(\frac{m_1(X_i)}{p(X_i)} + \frac{m_0(X_i)}{1-p(X_i)}\right), \\
\zeta_i &\equiv \Psi_p(W_i, p(X_i)) - S_p(X_i), \\
\epsilon_i &\equiv D_i - p(X_i), \\
\beta_n(X_i) &\equiv E[\hat{p}(X_i)|X_1, \ldots, X_n] - p(X_i) \\
&= \sum_{j:j\neq i}\omega_{ij}p(X_j) - p(X_i),
\end{aligned}$$

where the last quantity is the bias of the propensity score estimator conditional on $X_1, \ldots, X_n$.

We can write the proposed CATE estimator as

$$\begin{aligned}
&\sqrt{nh_1}(\hat{\tau} - \tau(x_1)) \\
&\quad = \frac{\frac{1}{\sqrt{nh_1}}\sum_{i=1}^n K_1\left(\frac{X_{1i} - x_1}{h_1}\right)[\Psi(W_i, \hat{p}(X_i) - \tau(x_1))]}{\frac{1}{nh_1}\sum_{i=1}^n K_1\left(\frac{X_{1i} - x_1}{h_1}\right)},
\end{aligned}$$

where $W_i = (Y_i, D_i, X_i)$. As

$$\frac{1}{nh_1}\sum_{i=1}^n K_1\left(\frac{X_{1i} - x_1}{h_1}\right) \xrightarrow{p} f_1(x_1)$$

under the stated assumptions, Theorem 1 part (a) will follow from showing that

$$\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h}\right)[\Psi(W_i,\hat{p}(X_i))-\tau(x_1)] = \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h}\right)[\psi(W_i)-\tau(x_1)]+o_p(1). \tag{A.8}$$

By a Taylor series expansion around $p(X_i)$,

$$\begin{aligned}
&\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h}\right)[\Psi(W_i,\hat{p}(X_i))-\tau(x_1)]\\
&= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)[\Psi(W_i,p(X_i))-\tau(x_1)]\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)\Psi_p(W_i,p(X_i))(\hat{p}(X_i)\\
&\quad-p(X_i))\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)\Psi_{pp}(W_i,p^*(X_i))(\hat{p}(X_i)\\
&\quad-p(X_i))^2\\
&= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)[\Psi(W_i,p(X_i))-\tau(x_1)]\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)(\hat{p}(X_i)-p(X_i))\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)\zeta_i(\hat{p}(X_i)-p(X_i))\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)\Psi_{pp}(W_i,p^*(X_i))(\hat{p}(X_i)\\
&\quad-p(X_i))^2\\
&\equiv J_0+J_1+J_2+J_3,
\end{aligned}$$

where $p^*(X_i)$ is a value between $\hat{p}(X_i)$ and $p(X_i)$ for all i, and the J terms are defined line by line.

We can further expand J_1 as

$$\begin{aligned}
J_1 &= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)(\hat{p}(X_i)-p(X_i))\\
&= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h}\right)S_p(X_i)\\
&\quad\left(\sum_{j:j\neq i}\omega_{ij}D_j-p(X_i)\right)\\
&= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\\
&\quad\left(\sum_{j:j\neq i}\omega_{ij}(\epsilon_j+p(X_j))-p(X_i)\right)\\
&= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\epsilon_i\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\left(\sum_{j:j\neq i}\omega_{ij}\epsilon_j-\epsilon_i\right)\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\\
&\quad\times\left(\sum_{j:j\neq i}\omega_{ij}p(X_j)-p(X_i)\right)\\
&= \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\epsilon_i\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n}\epsilon_i\left(\sum_{j:j\neq i}\omega_{ji}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j)\right.\\
&\quad\left.-K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right)\\
&\quad+\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\beta_n(X_i)\\
&\equiv J_{11}+J_{12}+J_{13},
\end{aligned}$$

where the $J_{1\cdot}$ terms are again defined line by line. Note that J_0 and J_{11} can be combined to yield the expression in (A.8). Hence, it is sufficient to show that J_{12}, J_{13}, J_2, and J_3 are all $o_p(1)$.

Bounding J_{12}. We start by writing

$$\begin{aligned}
&\frac{1}{\sqrt{h_1}}\left(\sum_{j:j\neq i}\omega_{ji}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j)\right.\\
&\left.-K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right)\\
&\quad=\frac{1}{\sqrt{h_1}}\sum_{j:j\neq i}(\omega_{ji}-\omega_{ij})K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j) \qquad \text{(A.9)}\\
&\quad+\frac{1}{\sqrt{h_1}}\left(\sum_{j:j\neq i}\omega_{ij}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j)\right.\\
&\quad\left.-K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right) \qquad \text{(A.10)}
\end{aligned}$$

and bounding (A.9) and (A.10) separately.

Turning to (A.9),

$$\begin{aligned}
&\frac{1}{\sqrt{h_1}}\sup_i\left|\sum_{j:j\neq i}(\omega_{ji}-\omega_{ij})K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j)\right|\\
&\quad\le\sup_i\sum_{j:j\neq i}|\omega_{ji}-\omega_{ij}|\left|K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j)\right|\\
&\quad\le\frac{MC_n}{h}\cdot\frac{h}{\sqrt{h_1}}\cdot\sup_i\sum_{j:j\neq i}\frac{1}{nh^k}\left|K\left(\frac{X_j-X_i}{h}\right)\right|\\
&\quad=O_p(1)\cdot o_p(1)\cdot O_p(1)=o_p(1), \qquad \text{(A.11)}
\end{aligned}$$

where the second inequality follows from Lemma 6.1 part (a) and the fact that $K_1(\cdot)S_p(\cdot)$ is bounded on $\mathcal{X}$ by some constant $M>0$ by Assumptions 5, 6, and 7(ii). As for the order of the factors, note that $C_n/h=O_p(1)$ by Lemma 6.1 part (a), $h/\sqrt{h_1}=o_p(1)$ by Assumption 8(iii), and $\sup_i\sum_{j:j\neq i}\frac{1}{nh^k}\left|K\left(\frac{X_j-X_i}{h}\right)\right|=O_p(1)$ by standard nonparametric estimation theory.

Turning to (A.10), first note that $\sum_{j:j\neq i}\omega_{ij}K_{1j}S_p(X_j)$ is an estimator of $K_{1i}S_p(X_i)$, and (A.10) can be regarded as the bias of this estimator conditional on $X_1,\ldots,X_n$. One can use standard arguments in the nonparametric econometrics literature to analyze the order of such conditional bias terms (see, e.g., Pagan and Ullah 1999, pp. 102–103). The novelty is the presence of the factor $K_1(\cdot/h_1)$, which changes somewhat the order of the bias term from the usual $O(h^s)$ (as in say Lemma 6.1 part (b)). More specifically, under Assumptions 8(i) and (ii), we can show that

$$\sup_i\left|\left(\sum_{j:j\neq i}\omega_{ij}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j) - K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right)\right| = O_p\left(\frac{h^s}{h_1^s}\right).$$

Therefore, by Assumption 8(iii),

$$\sup_i\left|\frac{1}{\sqrt{h_1}}\sum_{j:j\neq i}\omega_{ij}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j) - K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right| = o_p(1). \tag{A.12}$$

Combining (A.11) and (A.12) yields

$$\frac{1}{\sqrt{h_1}}\sup_i\left|\sum_{j:j\neq i}\omega_{ji}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j) - K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right| = o_p(1).$$

As the ϵ_i's are mutually independent conditional on the sample path of the X_i's, it further follows that

$$J_{12} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\epsilon_i\left\{\frac{1}{\sqrt{h_1}}\sum_{j:j\neq i}\omega_{ji}K_1\left(\frac{X_{1j}-x_1}{h_1}\right)S_p(X_j) - K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\right\} = o_p(1),$$

conditional on the sample path of the X_i's with probability approaching one.

Bounding J_{13}. Note that $\beta_n(x)$ is the bias of $\hat{p}(x)$, conditional on $X_1,\ldots,X_n$, which is $O_p(h^s)$ uniformly in x (see Lemma 6.1 part (b)). We can then bound J_{13} as follows:

$$\begin{aligned}|J_{13}| &= \left|\frac{1}{\sqrt{nh_1}}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)S_p(X_i)\beta_n(X_i)\right| \\ &\leq \sqrt{nh_1}\sup_{x\in\mathcal{X}}|\beta_n(x)|\frac{1}{nh_1}\cdot\sum_{i=1}^n\left|K_1\left(\frac{X_{1i}-x_1}{h_1}\right)\right||S_p(X_i)| \\ &= \sqrt{nh_1}O_p(h^s)\cdot O_p(1) = o_p(1)\cdot O_p(1) = o_p(1),\end{aligned}$$

where the second part of Assumption 8(iii) is used on the last line.

Bounding J_2. By Lemma 6.1(b) and Assumption 8(i), $\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)| = o_p(h^{s/2})$. Hence, by Assumption 8(iii), $h_1^{-1/2}\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)| = o_p(1)$. As the $\{\zeta_i\}$ are mutually independent conditional on the sample path of the X_i's, we have $J_2 = o_p(1)$ by the same argument used to show $J_{12} = o_p(1)$.

Bounding J_3. Note that $p^*(X_i)$ is between $\hat{p}(X_i)$ and $p(X_i)$, so it is uniformly bounded away from zero and one. Furthermore, by Lemma 6.1(b) and Assumption 8(i), $\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)|^2 = o_p(h^s)$. Therefore, by Assumption 8(iii), $\sqrt{nh_1}h^s\cdot h^{-s}\sup_i(\hat{p}(X_i)-p(X_i))^2 = o_p(1)\cdot o_p(1) = o_p(1)$, and the same argument used to control J_{13} yields $J_3 = o_p(1)$.

Combining the bounds above establishes (A.8) and hence Theorem 1 part (a).

Turning to Theorem 1 part (b), we write

$$\sqrt{nh_1}(\hat{\tau}(x_1)-\tau(x_1)) = \frac{1}{\sqrt{nh_1}}\frac{1}{f_1(x_1)}\sum_{i=1}^n[\psi(X_i,Y_i,D_i)-\tau(X_{1i})]K_1\left(\frac{X_{1i}-x_1}{h_1}\right) \tag{A.13}$$

$$+\frac{1}{\sqrt{nh_1}}\frac{1}{f_1(x_1)}\sum_{i=1}^n[\tau(X_{1i})-\tau(x_1)]K_1\left(\frac{X_{1i}-x_1}{h_1}\right) \tag{A.14}$$

$$+o_p(1).$$

It is straightforward to show that $E([\psi(X_i,Y_i,D_i)-\tau(X_{1i})]K_{1in}) = 0$, where we write $K_{1in} = K_1((X_{1i}-x_1)/h_1)$ to make it explicit that this quantity depends on n through h_1. For each n, the random variables $\{[\psi(X_i,Y_i,D_i)-\tau(X_{1i})]K_{1in}\}_{i=1}^n$ are independent and one can apply Lyapunov's CLT for triangular arrays to (A.13) to obtain the asymptotic distribution shown in part (*b*) of Theorem 1. The verification of the conditions of Lyapunov's CLT mimics exactly the proof of Theorem 3.5 by Pagan and Ullah (1999). The bias term given in (A.14) is $o_p(1)$ under Assumption 8(ii) and therefore has no bearing on the limit distribution. The verification of this claim follows, in turn, the proof of Theorem 3.6 by Pagan and Ullah (1999) coupled with the subsequent discussion about higher order kernels.

D. Handling Discrete Covariates

Let $q_m(x_1)\equiv P(M=1|X_1=x_1)$ and $\tau_m(x_1)\equiv E[Y(1)-Y(0)|X_1=x_1,M=1]$. For the female subpopulation the quantities $q_f(x_1)$ and $\tau_f(x_1)$ are defined analogously. For simplicity only, take again $\ell=1$. Note that the parameter of interest can be written as $\tau(x_1) = q_m(x_1)\tau_m(x_1)+q_f(x_1)\tau_f(x_1)$. Then, we estimate $\tau_m(x_1)$, $\tau_f(x_1)$, $q_m(x_1)$, and $q_m(x_1)$ by

$$\hat{\tau}_m(x_1)\equiv\frac{\frac{1}{nh_1^\ell}\sum_{i:M_i=1}\left(\frac{D_iY_i}{\hat{p}_m(X_i)}-\frac{(1-D_i)Y_i}{1-\hat{p}_m(X_i)}\right)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i:M_i=1}K_1\left(\frac{X_{1i}-x_1}{h_1}\right)},$$

$$\hat{\tau}_f(x_1)\equiv\frac{\frac{1}{nh_1^\ell}\sum_{i:M_i=0}\left(\frac{D_iY_i}{\hat{p}_f(X_i)}-\frac{(1-D_i)Y_i}{1-\hat{p}_f(X_i)}\right)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i:M_i=0}K_1\left(\frac{X_{1i}-x_1}{h_1}\right)},$$

$$\hat{q}_m(x_1)\equiv\frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n M_iK_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)},$$

$$\hat{q}_f(x_1)\equiv 1-\hat{q}_m(x_1) = \frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n(1-M_i)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}. \tag{A.15}$$

Last, $\tau(x_1)$ is estimated by

$$\hat{\tau}(x_1) = \hat{q}_m(x_1)\hat{\tau}_m(x_1)+\hat{q}_f(x_1)\hat{\tau}_f(x_1). \tag{A.16}$$

After we plug in those expressions in (A.15) into (A.16), the expression of $\hat{\tau}(x_1)$ in (A.16) reduces to that in (9).

Next, we verify the influence function representation displayed in (10). Note that the CATE estimator for the male group can be written as

$$\sqrt{n_mh_1}(\hat{\tau}_m(x_1)-\tau_m(x_1)) = \frac{1}{f_m(x_1)\sqrt{n_mh_1}}\sum_{i=1}^n(\psi_i-\tau_m(x_1))M_iK_{1i}+o_p(1),$$

where $K_{1i} = K((X_{1i} - x_1)/h_1)$ and $\psi_i = \psi(M_i, X_i, Y_i, D_i)$. We replace n_m with n in the scaling factor by writing

$$\begin{aligned}\sqrt{nh_1}\big(\hat{\tau}_m(x_1) - \tau_m(x_1)\big) &= \frac{n}{n_m}\frac{1}{f_m(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}(\psi_i \\ &\quad -\tau_m(x_1))M_i K_{1i} + o_p(1) \\ &= \frac{1}{q_m f_m(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}\psi_i M_i K_{1i} \\ &\quad + o_p(1),\end{aligned}$$

where $q_m \equiv P(M = 1)$ and $f_m(x_1)$ is the conditional density of X_1 on $M = 1$. Also, note that

$$P(X_1 = x_1, M = 1) = f_1(x_1)q_m(x_1) = q_m f_m(x_1).$$

Therefore,

$$\begin{aligned}\sqrt{nh_1}(\hat{\tau}_m(x_1) - \tau_m(x_1)) &= \frac{1}{f_1(x_1)q_m(x_1)\sqrt{nh_1}}\sum_{i=1}^{n} \\ &\quad (\psi_i - \tau_m(x_1))M_i K_{1i} + o_p(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\sqrt{nh_1}(\hat{\tau}_f(x_1) - \tau_f(x_1)) &= \frac{1}{f_1(x_1)q_f(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}(\psi_i - \tau_f(x_1)) \\ &\quad (1 - M_i)K_{1i} + o_p(1).\end{aligned}$$

Furthermore, the estimators $\hat{q}_m(x_1)$ and $\hat{q}_f(x_1)$ can be represented as

$$\begin{aligned}\sqrt{nh_1}(\hat{q}_m(x_1) - q_m(x_1)) &= \frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n} \\ &\quad (M_i - q_m(x_1))K_{1i} + o_p(1),\end{aligned}$$

and

$$\begin{aligned}\sqrt{nh_1}(\hat{q}_f(x_1) - q_f(x_1)) &= \frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n} \\ &\quad (1 - M_i - q_f(x_1))K_{1i} + o_p(1).\end{aligned}$$

We can combine the previous four displays to obtain the representation in (10):

$$\begin{aligned}&\sqrt{nh_1}(\hat{\tau}(x_1) - \tau(x_1)) \\ &\quad = \sqrt{nh_1}[\hat{q}_m(x_1)\hat{\tau}_m(x_1) + \hat{q}_f(x_1)\hat{\tau}_f(x_1) \\ &\qquad -q_m(x_1)\tau_m(x_1) - q_f(x_1)\tau_f(x_1)] \\ &\quad = \sqrt{nh_1}q_m(x_1)(\hat{\tau}_m(x_1) - \tau_m(x_1)) + \sqrt{nh_1}(\hat{q}_m(x_1) \\ &\qquad -q_m(x_1))\hat{\tau}_m(x_1) + \sqrt{nh_1}q_f(x_1)\big(\hat{\tau}_f(x_1) - \tau_f(x_1)\big) \\ &\qquad +\sqrt{nh_1}\big(\hat{q}_f(x_1) - q_f(x_1)\big)\hat{\tau}_f(x_1) \\ &\quad = \frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}(\psi_i - \tau_m(x_1))M_i K_{1i} \\ &\qquad +\frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}\big(M_i - q_m(x_1)\big)\tau_m(x_1)K_{1i} \\ &\qquad +\frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}(\psi_i - \tau_f(x_1))(1 - M_i)K_{1i} \\ &\qquad +\frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}\big(1 - M_i - q_f(x_1)\big)\tau_f(x_1)K_{1i} + o_p(1) \\ &\quad = \frac{1}{f_1(x_1)\sqrt{nh_1}}\sum_{i=1}^{n}(\psi_i - q_m(x_1)\tau_m(x_1) \\ &\qquad -q_f(x_1)\tau_f(x_1))K_{1i} + o_p(1) \\ &\quad = (10),\end{aligned}$$

where the last equality follows from that $\tau(x_1) = q_m(x_1)\tau_m(x_1) + q_f(x_1)\tau_f(x_1)$.

E. The Proof of Theorem 2

Using the notation introduced in Appendix C, we again expand the numerator of the CATE estimator around $p(X_i) = p(X_i; \theta_0)$ as

$$\begin{aligned}&\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n}K_1\left(\frac{X_{1i} - x_1}{h}\right)[\Psi(W_i, p(X_i; \hat{\theta}_n)) - \tau(x_1)] \\ &\quad = \frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n}K_1\left(\frac{X_{1i} - x_1}{h_1}\right)[\Psi(W_i, p(X_i)) - \tau(x_1)] \quad \text{(A.17)} \\ &\qquad +\frac{1}{\sqrt{nh_1}}\sum_{i=1}^{n}K_1\left(\frac{X_{1i} - x_1}{h_1}\right)\Psi_p(W_i, p^*(X_i)) \\ &\qquad (\hat{p}(X_i; \hat{\theta}_n) - p(X_i)),\end{aligned}$$

where $p^*(X_i)$ is between $p(X_i; \hat{\theta}_n)$ and $p(X_i)$. We bound the second term as

$$\begin{aligned}&\frac{1}{\sqrt{nh_1}}\left|\sum_{i=1}^{n}K_1\left(\frac{X_{1i} - x_1}{h_1}\right)\Psi_p(W_i, p^*(X_i))\right. \\ &\qquad \left.(p(X_i; \hat{\theta}_n) - p(X_i))\right| \\ &\quad \le \sqrt{nh_1}\sup_{x\in\mathcal{X}}|p(x; \hat{\theta}_n) - p(x)| \cdot \\ &\qquad \frac{1}{nh_1}\sum_{i=1}^{n}\left|K_1\left(\frac{X_{1i} - x_1}{h_1}\right)\Psi_p(W_i, p^*(X_i))\right|, \quad \text{(A.18)}\end{aligned}$$

where the first factor is $o_p(1)$ by Assumption 9 and the second factor is $O_p(1)$ under mild regularity conditions. Since $\Psi(W_i, p(X_i, \theta)) =$

$\psi_\theta(X_i, Y_i, D_i)$, part (a) of Theorem 2 is proven. The proof of part (b) is identical to the the proof of Theorem 1(b) and is therefore omitted.

F. Proof of Theorem 3

The proof is based on establishing the following representations:

$$\sqrt{nh_1^\ell}\left(\frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n\left(D_iY_i-\frac{\hat{p}(X_i)(1-D_i)Y_i}{1-\hat{p}(X_i)}\right)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}-\tau_t(x_1)p_{x_1}\right)$$

$$=\frac{1}{\sqrt{nh_1^\ell}}\frac{1}{f_1(x_1)}\sum_{i=1}^n\left(D_i(Y_i-m_1(X_i))-\frac{p(X_i)(1-D_i)(Y_i-m_0(X_i))}{1-p(X_i)}+D_i(m_1(X_i)-m_0(X_i))-\tau_t(x_1)p_{x_1}\right)\times K_1\left(\frac{X_{1i}-x_1}{h_1}\right)+o_p(1),$$

$$\sqrt{nh_1^\ell}\left(\frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n D_iK_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}-p_{x_1}\right)$$

$$=\frac{1}{\sqrt{nh_1^\ell}}\frac{1}{f_1(x_1)}\sum_{i=1}^n(D_i-p_{x_1})K_1\left(\frac{X_{1i}-x_1}{h_1}\right)+o_p(1).$$

The second equation suggests that we can replace the denominator of $\hat{\tau}_t(x_1)$ in (12) with $\sum_{i=1}^n D_iK_1\big((X_{1i}-x_1)/h_1\big)/(nh_1^\ell)$ without changing the first-order asymptotic of $\hat{\tau}_t(x_1)$.

ACKNOWLEDGMENTS

The authors thank Fabio Canova, Steven Durlauf, Kei Hirano, Gábor Kézdi, Miklós Koren, Botond Kőszegi, Hon Ho Kwok, Francis Vella, the editors, and two anonymous referees for useful comments. All errors are authors' responsibility.

[Received July 2012. Revised March 2014.]

REFERENCES

Abadie, A. (2003), "Semiparametric Instrument Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [499]

Abrevaya, J. (2006), "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach," *Journal of Applied Econometrics*, 21, 489–519. [494,495,496]

Abrevaya, J., and Dahl, C. (2008), "The Effects of Birth Inputs on Birthweight: Evidence From Quantile Estimation on Panel Data," *Journal of Business and Economic Statistics*, 26, 379–397. [486,495]

Almond, D., Chay, K. Y., and Lee, D. S. (2005), "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120, 1031–1083. [494,495]

Cattaneo, M. D. (2010), "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability," *Journal of Econometrics*, 155, 138–154. [500]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [496]

da Veiga, P. V., and Wilder, R. P. (2008), "Maternal Smoking During Pregnancy and Birthweight: A Propensity Score Matching Approach," *Maternal and Child Health Journal*, 12, 194–203. [495,496]

Donald, S. G., Hsu, Y.-C., and Lieli, R. P. (2014a), "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business and Economic Statistics*, 32, 395–415. [492,499]

——— (2014b), "Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion," *Statistics and Probability Letters*, 95, 132–138. [489]

Frölich, M. (2007), "Nonparametric IV Estimation of Local Average Treatment Effects With Covariates," *Journal of Econometrics*, 139, 35–75. [499]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [485,486,488,490,499]

Heckman, J., Ichimura, H., and Todd, P. (1997), "Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Program," *Review of Economic Studies*, 64, 605–654. [485]

——— (1998), "Matching as an Econometric Evaluations Estimator," *Review of Economic Studies*, 65, 261–294. [485]

Heckman, J., and Vytlacil, E. (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. [485]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [486,488,490,496,499,501]

Hong, H., and Nekipelov, D. (2010), "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 1, 279–304. [499]

Hsu, Y.-C. (2012), "Consistent Tests for Conditional Treatment Effects," Working Paper, Institute of Economics, Academia Sinica, Taiwan. [485]

Ichimura, H., and Linton, O. (2005), "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds. D. W. K. Andrews and J. Stock, Cambridge, UK: Cambridge University Press. [488,501]

Imbens, G., and Ridder, G. (2009), "Estimation and Inference for Generalized Full and Partial Means and Derivatives," Working Paper, Department of Economics, Harvard University. [490]

Imbens, G. W., and Wooldridge, J. W. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [487]

Khan, S., and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 6, 2021–2042. [485,488,492]

Lee, S., and Whang, J.-Y. (2009), "Nonparametric Tests of Conditional Treatment Effects," *Cowles Foundation Discussion Papers 1740*, Cowles Foundation, Yale University. [485]

MaCurdy, T., Chen, X., and Hong, H. (2011), "Flexible Estimation of Treatment Effect Parameters," *American Economic Review*, 101, 544–551. [485]

Masry, E. (1996), "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599. [501]

Pagan, A., and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge, UK: Cambridge University Press. [501,503]

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [485,486]

———(1985), "Reducing Bias in Observational Studies Using Subclassfication on the Propensity Score," *Journal of American Statistical Association*, 79, 516–524. [485]

Su, L. (2011), "A Brief Introduction to Nonparametric Econometrics," Lecture Notes, School of Economics, Singapore Management University. [501]

Walker, M. B., Tekin, E., and Wallace, S. (2009), "Teen Smoking and Birth Outcomes," *Southern Economic Journal*, 75, 892–907. [495,496,500]

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data* (2nd ed.), Cambridge, MA: MIT Press. [487]