Estimating Conditional Average Treatment Effects*

Jason Abrevaya† Yu-Chin Hsu‡ Robert P. Lieli§

July, 2012

Abstract

We consider a functional parameter called the conditional average treatment effect (CATE), designed to capture 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 the unconfoundedness assumption and propose inverse probability weighted estimators for it. Under regularity conditions, some of which are standard and some of which 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 per capita income in the mother's zip code. For non-white mothers, the average effect of smoking is predicted to become stronger (more negative) as a function of income.

Keywords: conditional average treatment effect, inverse probability weighted estimation, treatment effect heterogeneity, nonparametric methods, birth weight

*We thank Fabio Canova, Kei Hirano, Gábor Kézdi, Miklós Koren and Botond Kőszegi for useful comments. All errors are our responsibility.

†Department of Economics, University of Texas, Austin.

‡Department of Economics, University of Missouri, Columbia.

§Department of Economics, Central European University, Budapest and the National Bank of Hungary. Email: rlieli@gmail.com.

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) discuss the identification and estimation of CATE(x), which they call ATE(x), in terms of the marginal treatment effect in a general structural model. Khan and Tamer (2010) mention CATE(x) explicitly, but their focus is on ATE. Lee and Whang (2009) and Hsu (2012) consider estimating and testing hypotheses about CATE(x) when X is absolutely continuous, and provide detailed asymptotic theory. MaCurdy, Chen and Hong (2012) also discuss the identification and estimation of CATE(x).

In this paper 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 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 paper 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.[1] 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

[1] For simplicity, some of the formal results stated in the text will assume that the full vector X has continuous components only. Nevertheless, in applications X will typically contain discrete as well as continuous covariates. We will discuss how to deal with the mixed case in comments following the formal theorems.

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 paper we restrict attention to the first approach.

We apply our methods to a dataset on births in North Carolina recorded between 1988 to 2002. More specifically, we estimate the expected effect of a first-time mother's smoking during pregnancy on the birth weight of her child as a function of per capita income in the mother's zip code (intended to be a proxy for per capita family income). In accordance with previous studies, we distinguish mothers by race, and focus on the subsample classified as non-white. The contribution of this exercise to the pertaining empirical literature consists of exploring 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) estimate 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, *ibid.* report 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 paper 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 income levels. This finding is robust across various specifications of X, smoothing parameters, and the type of estimator used. There is however a great deal of uncertainty about the extent of this variation.

The rest of the paper is organized as follows. In Section 2 we introduce the CATE parameter, and discuss its identification and estimation. The first order asymptotic properties of the proposed estimators are developed. In Section 3 we present the empirical application. Section 4 outlines possible extensions of the basic framework, including multivalued treatments and instrumental variables. Section 5 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 paper 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 $X \in \mathbb{R}^k$, $1 \leq \ell < k$, 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

$$\tau(x_1) = E\Big[E[Y|D=1,X] - E[Y|D=0,X]\Big| X_1 = x_1\Big] \quad \text{or} \tag{1}$$

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

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,

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

where $K(\cdot)$ is a kernel function and h is a smoothing parameter (bandwidth). Option (ii) is a parametric estimate of $p(x)$, e.g., 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, i.e., 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).[2]

[2]Note that the second stage calls for estimating the propensity score at the sample observations on X. In the fully nonparametric case we use the "leave-one-out" version of (3) to estimate these quantities.

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 paper. 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 in the last fifteen years or so. 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 paper 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

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

We examine two benchmark cases.

Case 1. If one assumes $r_0(X) = r_1(X) = [X - E(X)]'\beta$, then $\text{CATE}(x_1) = E[Y(1) - Y(0) \mid X_1 = x_1]$ is a constant function whose value is equal to ATE$= \mu_1 - \mu_0$ everywhere. ATE itself can be estimated as the regression coefficient on D from a linear regression of Y on a constant, D and X.

Case 2. If one assumes $r_0(X) = [X - E(X)]'\beta$ and $r_1(X) = [X - E(X)]'(\beta + \delta)$, then ATE can be estimated similarly to Case 1—one only needs to include interaction terms between D and $X - \bar{X}$ as additional explanatory variables in the regression described above. However, $\text{CATE}(X_1)$ is no longer a trivial function of X_1; rather, it is given by

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

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

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

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

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

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

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

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

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.

In most real world applications the assumptions underlying the two benchmark cases will hold, at best, as approximations. Therefore, a nonparametric (or semiparameric) CATE estimator may well tease out useful information from the data that would otherwise be masked by these narrow functional form assumptions. We do not mean to imply that nonparametric methods are inherently superior—often the curse of dimensionality can only be overcome by using tightly parameterized models. It is however important to be aware of the limitations of such models, especially if they are being used for analytical convenience only.

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 $\mathcal{X}$.*

Assumption 7 (Kernels): *(i) $K(u)$ is a kernel of order s, is symmetric around zero, is equal to zero outside $\prod_{i=1}^{k}[-1, 1]$, integrates to 1 and is continuously differentiable.*[3]

[3] $K : \mathbb{R}^k \to \mathbb{R}$ is a kernel of order s if $\int u_1^{p_1} \cdot \ldots \cdot u_k^{p_k} K(u)du = 0$ for all nonnegative integers $p_1, \ldots, p_k$ such that $1 \leq \sum_i p_i < s$.

(ii) $K_1(u)$ *is a kernel of order* s_1*, is symmetric around zero, integrates to 1, 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*,*

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

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

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 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, e.g., 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 in order 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 (2011). 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 non-treated subpopulations, respectively. Then, we estimate $\sigma^2_\psi(x_1)$ by

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

The consistency of $\hat{\sigma}^2_\psi(x_1)$ can be shown as follows. First, let $\tilde{\sigma}^2_\psi(x_1)$ be the (infeasible) estimator for $\sigma^2_\psi(x_1)$ 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 $\tilde{\sigma}^2_\psi(x_1)$ is a consistent estimator for $\sigma^2_\psi(x_1)$. 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.[4]

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

[4]Note that $\hat{\psi}$ is averaged over data points while $\hat{\tau}$ is evaluated at a fixed point. This is why the first estimator needs to be uniformly consistent, and it is enough for the second to be pointwise consistent.

applies to $\hat{\tau}(x_1)$:

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

where

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

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:*

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

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

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, i.e., 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 semiparametric 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

$$\sigma^2_{\psi}(x_1) = E\Big[\big(m_1(X) - m_0(X) - \tau(x_1)\big)^2 + \frac{\sigma^2_1(X)}{p(X)} + \frac{\sigma^2_0(X)}{1-p(X)}\Big| X_1 = x_1\Big],$$
$$\sigma^2_{\psi_\theta}(x_1) = \sigma^2_{\psi}(x_1) + E\Big[p(X)(1-p(X))\Big(\frac{m_1(X)}{p(X)} + \frac{m_0(X)}{1-p(X)}\Big)^2\Big| X_1 = x_1\Big],$$

where $\sigma^2_d(x) = Var(Y(d)|X = x)$ for $d = 0$ and 1. Hence, clearly $\sigma^2_{\psi_\theta}(x_1) \geq \sigma^2_{\psi}(x_1)$. 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 (29) in the expansion of the estimator in Appendix E, and (ii) that the remainder term is well behaved in that the second factor in (30) is $O_p(1)$. A set of primitive sufficient conditions could be obtained by suitable changes to Assumptions 4, 5, 6, and 7(ii).

6. 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 Empirical application

3.1 The data set and the identification strategy

We apply our method to study the effect of maternal smoking on birthweight. 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, e.g., Abrevaya (2006) or Almond, Chay and Lee (2005) for a set of references. It is generally accepted that smoking is one of the major modifiable risk factors for low birth weight, and there are many studies that attempt to estimate its average causal effect. As we point out in the introduction, our intended contribution in this exercise 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.

We use a data set of yearly vital statistics from the North Carolina State Center Health Services, accessible through the Odum Institute at the University of North Carolina. We focus on first-time black mothers between 1988 and 2002; these restrictions, which we will motivate shortly, yield 157,989 observations. An attractive feature of the North Carolina vital statistics data is that it records the mother's zip code. The availability of this information allows us to assign to each mother some zip code level characteristics, such as per capita income, population density and geographical location (latitude and longitude).[5] Per capita income in the mother's zip code can be taken as a proxy for family income, a potentially important covariate typically not recorded in vital statistics data.[6] Population density can help capture further differences between urban, suburban and rural communities and, at least in the fully nonparametric setting, latitude and longitude can be used as a general control for residual zip-code level unobserved heterogeneity ('zip code fixed effects').[7] The

[5]The zip code level data is from the Census 2000 U.S. Gazetteer files.

[6]da Veiga and Wilder (2008, p. 197) explicitly state that they consider the lack of observations on income to be a "main limitation" of their data set.

[7]Using latitude and longitude in the semiparametric setting is somewhat problematic, as it is clearly not realistic to restrict the effect of these variables on the likelihood of smoking to be monotonic. Higher powers as well as interaction terms with other variables are essential. For simplicity, we omit latitude and longitude in the semiparametric analysis but use it for robustness checks in the fully nonparametric estimation.

use of zip code level covariates is, to our knowledge, new in the literature.

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 use variants of the unconfoundedness assumption to identify the average effect of smoking on birth weight. By contrast, Abrevaya (2006) and Abrevaya and Dahl (2008) attempt to control 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. While we have data on how many previous deliveries a mother had, we do not have information on the birthweight of these children. (We cannot identify mothers giving birth in multiple years.) This unobserved factor is likely to be related to the decision whether or not to smoke during the current pregnancy as well as the potential birth weight outcomes for the current pregnancy even after conditioning on a vector X of observables. For example, previous birth weights can contain information about birthweight-relevant genetic factors that are not "picked up" by any component of X. Our focus on first births is an attempt to deal with this problem. In contrast, the decision to restrict attention to non-white mothers is mostly arbitrary. In estimating the effect of smoking on birth weight, the literature routinely breaks down results by race. The number of first time white mothers in our sample is on the order of half a million; we save significant computer time by working with a smaller number of observations. Of course, we could also use a random subsample of white mothers, but we do not pursue that exercise here.

The precise selection of the components of X depends on the estimation method employed. In particular, the fully nonparametric approach limits the number of discrete variables we can use, as the propensity score needs to be estimated separately for each possible configuration of these variables (see comment 8 after Theorem 1). Some of these variables in

question are indicators for relatively rare medical complications or behaviors such as gestational diabetes, hypertension, alcohol consumption during pregnancy, etc. The intersection of these categories is rarer still, making nonparametric estimation of the propensity score conditional on the continuous components of X practically infeasible. The semiparametric CATE estimator is however able accommodate a large number of such covariates by parameterizing the propensity score; specifically, we use a logit model based on a linear index. (The tradeoff, as discussed in Section 2.4, is that the functional form through which the components of X are allowed to affect the probability of smoking is restricted.) We then conduct a sensitivity analysis to determine whether dropping some of the discrete covariates affects the semiparametric estimation results in a substantial way. We are able to identify a smaller set of discrete variables that produces very similar results to the larger baseline specification, and use this smaller set in the fully nonparametric exercise.

More concretely, the baseline specification of X for the semiparametric estimator consists of the mother's age, education, month of first prenatal visit[8], number of prenatal visits, per capita income in the mother's zip code, population density in the mother's zip code, 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. The reduced set of covariates used in the fully nonparametric estimation contains the first six variables listed above (treated as continuous) and two indicators (one for the baby's gender and one for the mother's marital status). Admittedly, some of the variables designated as 'continuous' stretch the limits of the definition. For example, month of first prenatal visit is clearly not continuous, but it is more convenient to treat it as such instead of further partitioning the sample according to the ten possible values this variable can take on.[9]

To avoid scaling issues arising from using the same bandwidth for all continuous covariates, we standardize them before estimation.

[8]We set month of first visit to 10 if prenatal care is foregone.

[9]To avoid technical complications, for the nonparametric estimations we add uniform $[-.5, .5]$ random numbers to mother's age, education, month of fist prenatal visit and number of visits. This has little impact on the results, though it is noticable that there are small differences between replications.

3.2 Empirical implementation

We estimate the conditional average treatment effect of a mother's smoking on her firstborn's birth weight as a function of (zip-code level) per capita annual income. We evaluate this function at a grid of income levels, where the lowest gridpoint is one standard deviation below the mean, the largest is two standard deviations above it, and the step size is .1 standard deviations (the approximate mean and standard deviation are $\$18,000$ and $\$5,000$, respectively).

As is common for nonparametric and semiparametric estimation procedures, the choice of smoothing parameter(s) turns out to have a substantial impact on the results. Unfortunately, the first order asymptotic theory presented in Section 2 only pins down convergence rates and leaves open the question of how to choose the bandwidths for a given sample size. We use several rules of thumb combined with experimentation to calibrate the bandwidths in a way that the resulting CATE estimates are 'reasonable'. The two main criteria for reasonableness are: (i) The ATE value implied by the CATE estimate should not deviate too much from previous estimates in the literature[10]; (ii) the range of the CATE function should not include values that appear extreme in light of prior knowledge. E.g., an estimated average smoking effect of, say, -1500 grams is simply not credible even in the tails of X_1; similarly, a significantly positive smoking effect for some values of X_1 would also contradict previously accumulated evidence and knowledge.

Bandwidth choice is somewhat simpler for the semiparametric estimator as only h_1 needs to be specified. Furthermore, since logit regressions are quick to run even for very large samples, we are able to do a direct grid search over various values of h_1 and examine the resulting CATE functions. Not surprisingly, smaller values of h_1 produce more variable (non-monotonic) CATE estimates with wider range, while the implied ATE tends to stay roughly

[10]By the law of iterated expectations, the average of CATE estimates computed at each sample observation yields an estimate of ATE. (Given about 160,000 observations, the estimation of the implied ATE is rather time consuming.) Typical ATE estimates obtained under some form of the unconfoundedness assumption (OLS, propensity score matching) range from about -200 to -250 grams; see, e.g., Abrevaya (2006), da Veiga and Wilder (2008). Note, however, that no such study known to us restricts attention to first time mothers.

constant. In the fully nonparametric case, we search over combinations of h and h_1. Since the propensity score needs to be estimated nonparametrically at each sample observation, the search can become very time consuming if the full sample is used. We therefore take random subsamples of size 30,000 to calibrate bandwidths, and then downscale them to the full sample size using the bandwidth formulas shown in comment 6 after Theorem 1.[11] Similar to the semiparametric case, smaller values of h_1 produce more variable estimates with a wider range, while leaving the implied ATE roughly constant. On the other hand, decreasing the value of h below a certain level tends to shift the entire function upward in a roughly parallel manner.

Another implementation issue is the specification of the kernels K and K_1. Our estimation results are more robust to bandwidth choice if we use higher order product kernels derived from the normal kernel (a kernel with unbounded support) rather than from the Epanechnikov kernel (a kernel with bounded support), and therefore opt for the former. While our proofs assume that K is zero outside a compact set for expositional convenience, the theory extends easily to kernels with exponential tails (cf. comment 9 after Theorem 1). We use a regular kernel K_1 in implementing the semiparametric estimator.

3.3 Semiparametric results

Figure 1 depicts a range of point estimates for bandwidth levels $h_1 = 0.25, 0.5, 1, 1.5, 6.5$. As the bandwidth increases, the estimated CATE function becomes flatter, less variable, and its range shrinks. The largest bandwidth oversmoothes to such an extent that it forces the CATE function to be essentially constant; the value of this constant (approx. -210 grams) can be taken as an estimate of ATE. The implied ATE levels associated with the other CATE estimates are very similar; specifically, they are equal to -209, -195, -193 and -199 grams

[11] Let h^* and h_1^* denote the bandwidth values picked in a subsample of size n. In the baseline specification for X the number of continuous variables is $k = 6$, so $s = 6$ and $s_1 = 8$. Furthermore, $\ell = 1$. We embed h^* and h_1^* into a bandwidth sequence suggested by the asymptotic theory. By comment 6 after Theorem 1 we write $h^* = an^{-1/(12+\delta)}$ and $h_1^* = a_1 n^{-1/(17-\delta_1)}$ for some (positive) numbers a, a_1, δ and δ_1. As the theory allows δ and δ_1 to be arbitrarily small, we simply set them to zero, so that a and a_1 are pinned down unambiguously. In the full sample with N observations we then set $h = aN^{-1/12}$ and $h_1 = a_1 N^{-1/17}$.

in order of increasing bandwidth. Qualitatively, all the non-constant CATE estimators tell a similar story—the average causal effect of smoking is predicted to become stronger (more negative) for per capita income levels one to two standard deviations above the mean (with a possible turnaround for very high incomes). Furthermore, the effect appears to be lower than the overall ATE at average income levels. A speculative story consistent with the slope of the CATE estimates is that higher income smoking mothers are more intense smokers (i.e., cigarettes are a normal good). An alternative explanation is that the marginal smoking effect is smaller at the lower end of the birth weight distribution, which is where low income families are likely to fall. The Abrevaya and Dahl (2008) quantile regression study is consistent with this story.

Unfortunately, the numerical CATE estimates vary a lot with h_1, especially at higher incomes. The change in the estimated CATE as one moves from the mean income to 1.5 standard deviations above the mean ranges from under -50 grams to over -500 grams, a factor of 10. In addition to the uncertainty generated by bandwidth choice, there is further estimation uncertainty for a given bandwidth. Figures 2 and 3 show the ± 2 standard error bounds for the two 'extreme' CATE estimators with $h_1 = 0.25$ and $h_1 = 6.5$, respectively. In the latter case the bound is roughly constant at about ± 60 grams, while in the former it increases with the level of income from roughly ± 100 grams to ± 300 grams.

We perform several robustness checks in addition to varying h_1. A potential concern is the decline in smoking incidence during the sample period (more exactly, the fact that this decline might take place in a non-random fashion even conditional on X). Adding year fixed effects to X preserves the results in Figure 1 almost exactly; the only noteworthy difference is that the implied ATEs shift downwards to about -240 grams. Including birth year as a raw control (to mimic a trend) produces virtually no change relative to baseline.

The Kessner index of prenatal care utilization is another frequently used control in the literature; nevertheless, including it has no visible effect on the results.[12] In fact, dropping all indicators except baby's gender and marital status from X has little effect on the CATE

[12]While the Kessner index is based entirely on the month of the first prenatal visit and the number of visits, in the semiparametric case these variables enter the propensity score through a restricted functional form, so the addition of this control is not a priori redundant.

estimates; most notably, the minimum (most negative) CATE values attained at higher income levels move slightly upward. (As discussed above, we will employ this smaller set of covariates to produce fully nonparametric estimates of the smoking effect.) Finally, using the log of income instead of levels in the baseline specification does not appreciably change the results.

3.4 Nonparametric results

Based on preliminary bandwidth searches described in Section 3.2, we set up the following grid for full sample estimation: $h \in \{3.7, 4.8\}$ and $h_1 \in \{3.0, 3.7, 5.0, 6.5\}$.[13] In Figure 4 we present results for all combinations of h and h_1 chosen from these sets.

A number of aspects of the bandwidth choice are worth highlighting. As can be seen, we must employ relatively large bandwidths in estimating the propensity score to obtain sensible nonparametric CATE estimates. Even for h values in the 2 to 3 range, one frequently obtains $\hat{p}(X_i)$ values outside of the interval $[0, 1]$ due to K being a higher order kernel. While we trim the estimates at 0.005 and 0.995, a large proportion of $\hat{p}(X_i)$ concentrated around 0 and 1 can shift the CATE function up to clearly unreasonable levels with implied ATEs on the order of *plus* several hundred grams. Larger values of h draw the $\hat{p}(X_i)$ away from the boundaries at the cost of reducing their variation across observations to a few percentage points. Similarly to the semiparametric estimator, increasing h_1 makes the estimated CATE function flatter and its range smaller. Nevertheless, the estimate still has a slight negative slope for h_1 as large as 6.5, unlike in the semiparametric case. The asymptotic theory for the nonparametric estimator seems to put limits on one's ability to decrease h_1 in order to bring out more detail in the CATE estimates. In particular, the theory requires $h/h_1 \to 0$, which suggests restricting attention to (h, h_1) pairs with $h < h_1$, though it is not clear how strictly one should enforce this in practice. Therefore, in Figure 4 we also display some estimates with $h > h_1$; these are indeed the estimates with the largest range.

Qualitatively, the nonparametric estimates are consistent with the main features of the

[13] More specifically, we set $h = 10N^{-1/12}$, $13.1N^{-1/12}$ and $h_1 = 6N^{-1/17}$, $7.5N^{-1/17}$, $10N^{-1/17}$, $13.1N^{-1/17}$, where N is the sample size.

semiparamateric results; namely, that the effect of smoking is predicted to become more negative at higher income levels and is below the overall ATE at average incomes. The most important quantitative difference is that the implied ATEs are clearly smaller in absolute value than in the semiparametric case. The extent to which the smoking effect becomes stronger as one moves from average income ($\approx 18,000$) to, say, 1.5 standard devations above average ($\approx 25,500$) is comparable with the lower end of the semiparametric estimates. Regarding estimation uncertainty for a given bandwidth, Figures 5 and 6 display the estimated ± 2 standard error bounds for two of the estimates. In light of comment 3 after Theorem 2, it is not surprising that the nonparametric standard errors are smaller than the semiparametric ones; nonetheless, barely ± 15 grams around the point estimate feels like a rather too optimistic assessment of the estimator's accuracy.

As in the semiparametric case, we perform a number of robustness checks. Figure 7 shows that disregarding the asymptotic constraints on the ratio h/h_1 and further decreasing h_1 can produce CATE estimates that are qualitatively similar to the undersmoothed semiparametric estimates.[14] As the distribution of births is fairly even across years, we can implement fully nonparametric year fixed effects estimation by splitting the sample by birth year (in addition to the gender and marriage indicators). The estimation results, displayed in Figure 8, are qualitatively very similar to the baseline; there is only a small upward shift in some of the estimates. Replacing zip code population density with 'zip code fixed effects' (more precisely, latitude and longitude), has a more substantial effect on the estimation results; see Figure 9. (The upward adjustment in the smoothing parameters is due to the fact that X has one more component than in the baseline case.) While these estimates look similar in shape and range to Figure 4, they all shift downward, increasing the magnitude of the implied ATEs by about 25 grams. Finally, using the log of income in place of income does not cause noteworthy changes in the results.

[14]We note that some of the subpopulation-specific CATE estimates used in constructing Figure 7 no longer pass sensibility criteria. For example, for married mothers with baby girls we get CATE$\approx$+400 grams (siginificant) at mean income minus one standard deviation, and CATE$\approx$-1400 grams at mean plus two standard deviations.

4 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 multi-valued and unconfoundedness holds, we extend our theory to the conditional marginal average treatment effect (CMATE).

4.1 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

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

which suggests the estimator

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

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 asymptotics 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, e.g., 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) derive for their ATT estimators in the case when $p(x)$ is known.

4.2 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, e.g., Abadie (2003), Frolich (2007), Hong and Nekipelov (2010) and Donald, Hsu and Lieli (2011). 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 (2011) 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

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

Following Donald, Hsu and Lieli (2011), we can show that $CLATE(x_1)$ and $CLATT(x_1)$ are identified by

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

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 Theorem 1 and 2, one can obtain the asymptotic properties of the CLATE and CLATT estimators by the delta method. We omit the formal statements.

4.3 Multi-valued Treatment

In this section, we consider multi-valued (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: non-smokers, 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 5 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

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

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.

5 Conclusions

In this paper we study the estimation of conditional average treatment effects (CATE). This functional parameter is designed to capture the variation in the average treatment effect conditional on some covariate(s) used in identifying the unconditional average. We propose inverse probability weighted estimators of this function and provide pointwise first order asymptotic theory. We also discuss a number of straightforward extensions. The application consists of estimating the average effect of a first time mother's smoking on birth weight conditional on per capita income in the mother's zip code (intended as a proxy for per capita family income); in this paper we further restrict attention to non-white mothers. The main qualitative finding is that smoking has a larger impact (in absolute value) at higher income levels. Nevertheless, the numerical estimates are rather sensitive to bandwidth choice; the estimated difference between CATE at mean income and CATE at 1.5 standard deviations above the mean varies quite a lot. Developing formal criteria for bandwidth choice should enhance applicability of these methods and it is 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, i.e., $\hat{\theta} = (\sum_{i=1}^n Z_i Z_i')^{-1} \sum_{i=1}^n Z_i y_i$. By standard asymptotic results for i.i.d. data (see, e.g., Thm. 7.2 of Wooldridge 2010),

$$(\hat{A}^{-1}\hat{B}\hat{A}^{-1})^{-1/2}\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, I), \tag{13}$$

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 heteroskedasticity.) 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 (13) 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

$$\hat{p}(X_i) = \sum_{j:j\neq i} \omega_{ij} Y_j, \tag{14}$$

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 1 *Given Assumptions 1 through 8, the propensity score estimator satisfies:*

$$(a)\ \ |\omega_{ij} - \omega_{ji}| \leq \frac{C_n}{nh^k}\left|K\Big(\frac{X_i - X_j}{h}\Big)\right|, \tag{15}$$

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

$$(b)\ \ \sup_{x\in\mathcal{X}} |\hat{p}(x) - p(x)| = O_p\left(h^s + \sqrt{\frac{\log n}{nh^k}}\right), \tag{16}$$

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 $\|X_j - X_i\|_\infty \leq h$. For all i, define

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

Then

$$\begin{aligned}
|\omega_{ij} - \omega_{ji}| &= \frac{1}{nh^k}\left|K\left(\frac{X_i - X_j}{h}\right)\right| \cdot \left|\hat{f}^{-1}(X_i) - \hat{f}^{-1}(X_j)\right| \\
\leq &\frac{1}{nh^k}\left|K\left(\frac{X_i - X_j}{h}\right)\right| \cdot \Big\{\left|\hat{f}^{-1}(X_i) - f^{-1}(X_i)\right| \\
&\qquad + \left|\hat{f}^{-1}(X_j) - f^{-1}(X_j)\right| + \left|f^{-1}(X_i) - f^{-1}(X_j)\right|\Big\}.
\end{aligned} \tag{17}$$

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,

$$\sup_i |\hat{f}(X_i) - f(X_i)| = O_p\left(h^s + \sqrt{\frac{\log n}{nh^k}}\right). \tag{18}$$

As $s \geq k \geq 2$, Assumption 8(i) implies that the quantity in (18) is $o_p(h)$. By the mean value theorem,

$$\sup_i |\hat{f}^{-1}(X_i) - f^{-1}(X_i)| \leq \sup_i \frac{1}{\tilde{f}_i^2} \sup_i |\hat{f}(X_i) - f(X_i)|, \tag{19}$$

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 (19) 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 $|f^{-1}(x_1) - f^{-1}(x_2)| \leq M\|x_1 - x_2\|_\infty$ 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 $\|X_j - X_i\|_\infty \leq h$. Combining these observations yields (15), where C_n can be taken as Mh plus twice the lhs (or rhs) of (19).

Part (b) is a standard result in kernel based nonparametric regression theory; see, e.g., 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

$$S_p(X_i) \equiv E[\Psi_p(W_i, p(X_i))|X_i] = -\left(\frac{m_1(X_i)}{p(X_i)} + \frac{m_0(X_i)}{1-p(X_i)}\right),$$

$$\zeta_i \equiv \Psi_p(W_i, p(X_i)) - S_p(X_i),$$

$$\epsilon_i \equiv D_i - p(X_i),$$

$$\beta_n(X_i) \equiv E[\hat{p}(X_i)|X_1, ..., X_n] - p(X_i) = \sum_{j:j\neq i} \omega_{ij} p(X_j) - p(X_i),$$

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

$$\sqrt{nh_1}(\hat{\tau} - \tau(x_1)) = \frac{\frac{1}{\sqrt{nh_1}} \sum_{i=1}^n K_1\big(\frac{X_{1i}-x_1}{h_1}\big)[\Psi(W_i, \hat{p}(X_i) - \tau(x_1))]}{\frac{1}{nh_1} \sum_{i=1}^n K_1\big(\frac{X_{1i}-x_1}{h_1}\big)},$$

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

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

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

$$\begin{aligned} &\frac{1}{\sqrt{nh_1}} \sum_{i=1}^n K_1\Big(\frac{X_{1i} - x_1}{h}\Big)[\Psi(W_i, \hat{p}(X_i)) - \tau(x_1)] \\ =&\frac{1}{\sqrt{nh_1}} \sum_{i=1}^n K_1\Big(\frac{X_{1i} - x_1}{h}\Big)[\psi(W_i) - \tau(x_1)] + o_p(1). \end{aligned} \tag{20}$$

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

$$\frac{1}{\sqrt{nh_1}} \sum_{i=1}^n K_1\Big(\frac{X_{1i} - x_1}{h}\Big)[\Psi(W_i, \hat{p}(X_i)) - \tau(x_1)]$$

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

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

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

where the J_1. terms are again defined line by line. Note that J_0 and J_{11} can be combined to yield the expression in (20). 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

$$\begin{aligned}
&\frac{1}{\sqrt{h_1}}\left(\sum_{j:j\neq i}\omega_{ji}K_1\Big(\frac{X_{1j}-x_1}{h_1}\Big)S_p(X_j)-K_1\Big(\frac{X_{1i}-x_1}{h_1}\Big)S_p(X_i)\right)\\
&=\frac{1}{\sqrt{h_1}}\sum_{j:j\neq i}(\omega_{ji}-\omega_{ij})K_1\Big(\frac{X_{1j}-x_1}{h_1}\Big)S_p(X_j) \qquad (21)\\
&\quad+\frac{1}{\sqrt{h_1}}\left(\sum_{j:j\neq i}\omega_{ij}K_1\Big(\frac{X_{1j}-x_1}{h_1}\Big)S_p(X_j)-K_1\Big(\frac{X_{1i}-x_1}{h_1}\Big)S_p(X_i)\right) \qquad (22)
\end{aligned}$$

and bounding (21) and (22) separately.

Turning to (21),

$$\begin{aligned}
&\frac{1}{\sqrt{h_1}}\sup_i\left|\sum_{j:j\neq i}(\omega_{ji}-\omega_{ij})K_1\Big(\frac{X_{1j}-x_1}{h_1}\Big)S_p(X_j)\right|\\
\leq&\sup_i\sum_{j:j\neq i}|\omega_{ji}-\omega_{ij}|\left|K_1\Big(\frac{X_{1j}-x_1}{h_1}\Big)S_p(X_j)\right|\\
\leq&\frac{MC_n}{h}\cdot\frac{h}{\sqrt{h_1}}\cdot\sup_i\sum_{j:j\neq i}\frac{1}{nh^k}\left|K\Big(\frac{X_j-X_i}{h}\Big)\right|=O_p(1)\cdot o_p(1)\cdot O_p(1)=o_p(1), \qquad (23)
\end{aligned}$$

where the second inequality follows from Lemma 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 1 part (a), $h/\sqrt{h_1}=o_p(1)$ by Assumption 8(iii), and $\sup_i\sum_{j:j\neq i}\frac{1}{nh^k}\big|K\big(\frac{X_j-X_i}{h}\big)\big|=O_p(1)$ by standard nonparametric estimation theory.

Turning to (22), 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 (22) 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), p. 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 1 part (b)). More specifically, under Assumptions 8(i) and (ii), we can show that

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

Therefore, by Assumption 8(iii),

$$\sup_i\left|\frac{1}{\sqrt{h_1}}\sum_{j:j\neq i}\omega_{ij}K_1\Big(\frac{X_{1j}-x_1}{h_1}\Big)S_p(X_j)-K_1\Big(\frac{X_{1i}-x_1}{h_1}\Big)S_p(X_i)\right|=o_p(1). \qquad (24)$$

Combining (23) and (24) yields

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

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

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

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 (cf. Lemma 1 part (b)). We can then bound J_{13} as follows:

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

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

Bounding J_2: By Lemma 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 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 (20) and hence Theorem 1 part (a).

Turning to Theorem 1 part (b), we write

$$\begin{aligned}
&\sqrt{nh_1}(\hat{\tau}(x_1) - \tau(x_1)) \\
&= \frac{1}{\sqrt{nh_1}} \frac{1}{f_1(x_1)} \sum_{i=1}^{n} [\psi(X_i, Y_i, D_i) - \tau(X_{1i})] K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big) \qquad (25) \\
&\quad + \frac{1}{\sqrt{nh_1}} \frac{1}{f_1(x_1)} \sum_{i=1}^{n} [\tau(X_{1i}) - \tau(x_1)] K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big) \qquad (26) \\
&\quad + o_p(1)
\end{aligned}$$

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 (25) 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 (26) 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

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

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

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

$$\hat{q}_f(x_1) \equiv 1 - \hat{q}_m(x_1) = \frac{\frac{1}{nh_1^\ell}\sum_{i=1}^n (1-M_i)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh_1^\ell}\sum_{i=1}^n K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}. \tag{27}$$

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

$$\hat{\tau}(x_1) = \hat{q}_m(x_1)\hat{\tau}_m(x_1) + \hat{q}_f(x_1)\hat{\tau}_f(x_1). \tag{28}$$

After we plug in those expressions in (27) into (28), the expression of $\hat{\tau}(x_1)$ in (28) 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

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

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

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

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,

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

Similarly,

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

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

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

and

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

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

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

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

$$\begin{aligned}
&\frac{1}{\sqrt{nh_1}} \sum_{i=1}^{n} K_1\Big(\frac{X_{1i} - x_1}{h}\Big)[\Psi(W_i, p(X_i; \hat{\theta}_n)) - \tau(x_1)] \\
=&\frac{1}{\sqrt{nh_1}} \sum_{i=1}^{n} K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big)[\Psi(W_i, p(X_i)) - \tau(x_1)] \qquad (29) \\
&+ \frac{1}{\sqrt{nh_1}} \sum_{i=1}^{n} K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big)\Psi_p(W_i, p^*(X_i))(\hat{p}(X_i; \hat{\theta}_n) - p(X_i)),
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{\sqrt{nh_1}} \left| \sum_{i=1}^{n} K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big)\Psi_p(W_i, p^*(X_i))(p(X_i; \hat{\theta}_n) - p(X_i)) \right| \\
&\leq \sqrt{nh_1} \sup_{x \in \mathcal{X}} |p(x; \hat{\theta}_n) - p(x)| \cdot \frac{1}{nh_1} \sum_{i=1}^{n} \left| K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big)\Psi_p(W_i, p^*(X_i)) \right|, \qquad (30)
\end{aligned}$$

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:

$$\begin{aligned}
&\sqrt{nh_1^\ell} \left(\frac{\frac{1}{nh_1^\ell} \sum_{i=1}^{n} \Big(D_i Y_i - \frac{\hat{p}(X_i)(1-D_i)Y_i}{1-\hat{p}(X_i)} \Big) K_1\big(\frac{X_{1i}-x_1}{h_1}\big)}{\frac{1}{nh_1^\ell} \sum_{i=1}^{n} K_1\big(\frac{X_{1i}-x_1}{h_1}\big)} - \tau_t(x_1) p_{x_1} \right) \\
=&\frac{1}{\sqrt{nh_1^\ell}} \frac{1}{f_1(x_1)} \sum_{i=1}^{n} \Bigg(D_i(Y_i - m_1(X_i)) - \frac{p(X_i)(1-D_i)(Y_i - m_0(X_i))}{1 - p(X_i)} \\
&\qquad + D_i(m_1(X_i) - m_0(X_i)) - \tau_t(x_1) p_{x_1} \Bigg) K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big) + o_p(1), \\
&\sqrt{nh_1^\ell} \left(\frac{\frac{1}{nh_1^\ell} \sum_{i=1}^{n} D_i K_1\big(\frac{X_{1i}-x_1}{h_1}\big)}{\frac{1}{nh_1^\ell} \sum_{i=1}^{n} K_1\big(\frac{X_{1i}-x_1}{h_1}\big)} - p_{x_1} \right) \\
=&\frac{1}{\sqrt{nh_1^\ell}} \frac{1}{f_1(x_1)} \sum_{i=1}^{n} (D_i - p_{x_1}) K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big) + o_p(1).
\end{aligned}$$

The second equation suggests that we can replace the denominator of $\hat{\tau}_t(x_1)$ in (12) with $\sum_{i=1}^{n} D_i K_1\big((X_{1i} - x_1)/h_1\big)/(nh_1^\ell)$ without changing the first order asymptotics of $\hat{\tau}_t(x_1)$.

References

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

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

Abrevaya, J., and C. Dahl (2008). "The effects of birth inputs on birthweight: evidence from quantile estimation on panel data," *Journal of Business and Economic Statistics*, **26**, 379-397.

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

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

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

Donald, S. G., Y.-C. Hsu and R. P. Lieli (2011). "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," Working Paper.

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

Hahn, J. (1988). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, **66**, 315–331.

Heckman, J. and J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, **84**, 862–874.

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

Heckman, J., H. Ichimura and P. Todd (1998). "Matching as an Econometric Evaluations Estima-

tor," *Review of Economic Studies*, **65**, 261–294.

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

Hirano, K., G. W. Imbens and G. Ridder (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, **71**, 1161–1189.

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

Hsu, Y-C. (2012). "Consistent Tests for Conditional Treatment Effects". Working Paper, Department of Economics, University of Missouri at Columbia.

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

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

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

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

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

MaCurdy, T., X. Chen, and H. Hong (2012). "Flexible Estimation of Treatment Effect Parameters". Working Paper, Department of Economics, Stanford University.

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

Pagan, A. and A. Ullah (1999). "Nonparametric Econometrics," Cambridge University Press.

Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational

Studies for Causal Effects," *Biometrika*, **70**, 41–55.

Rosenbaum, P. and D. Rubin (1985). "Reducing Bias in Observational Studies Using Subclassfication on the Propensity Score," *Journal of American Statistical Association*, **79**, 516–524.

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

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

Wooldridge, J. M. (2010). "Inverse Probability Weighted M-estimators for Sample Selection, Attrition, and Stratification," *Portuguese Economic Journal*, **1**, 117–139

Wooldridge, J. M. (2010). "Econometric Analysis of Cross Section and Panel Data", 2nd edition, Cambridge, MA: MIT Press.

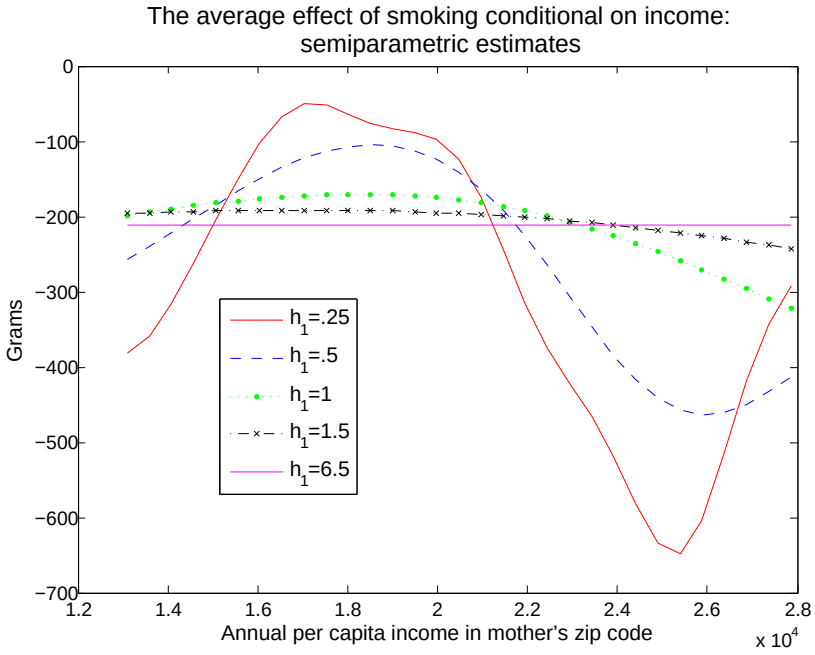

Figure 1: CATE as a function of per capita income: semiparametric estimates

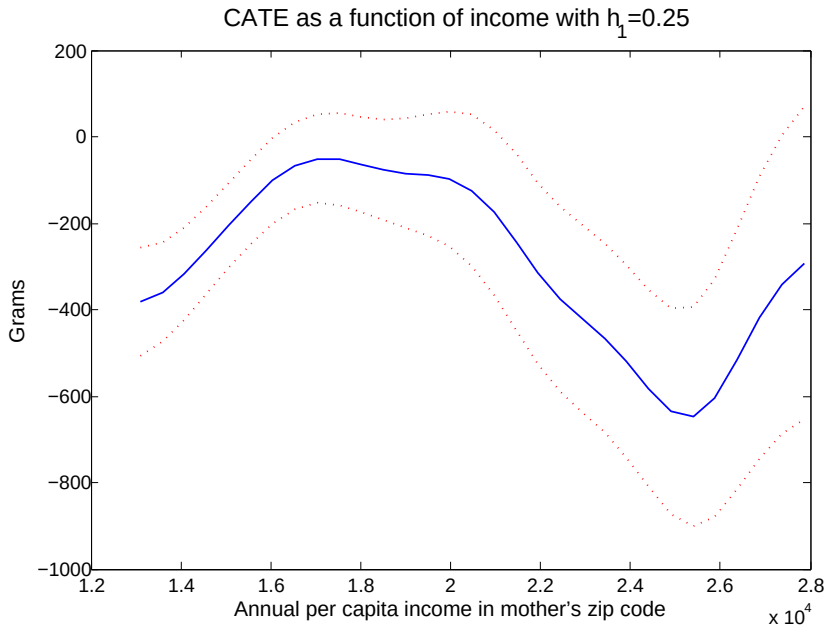

Figure 2: CATE as a function of per capita income: semiparametric estimate with $h_1 = 0.25$ and ± 2 standard errors

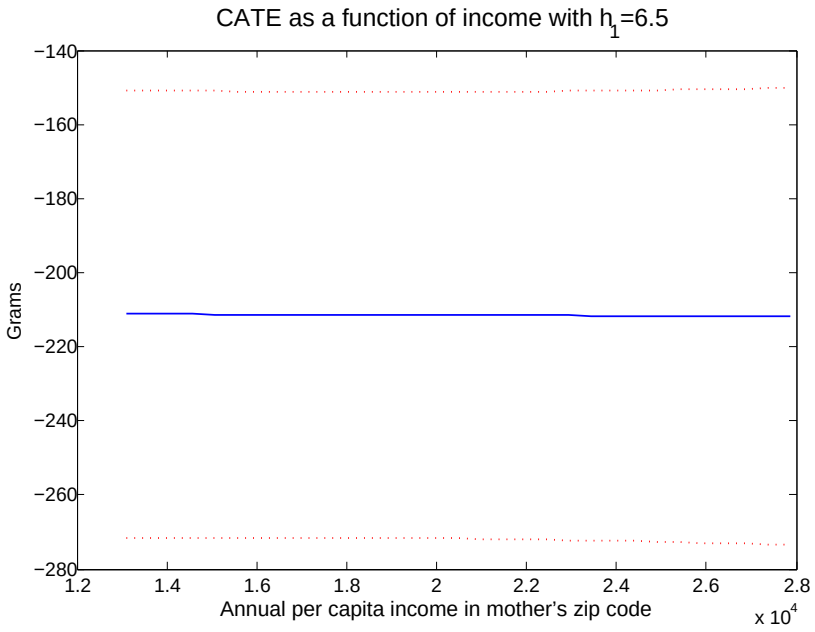

Figure 3: CATE as a function of per capita income: semiparametric estimate with $h_1 = 6.5$ and ± 2 standard errors

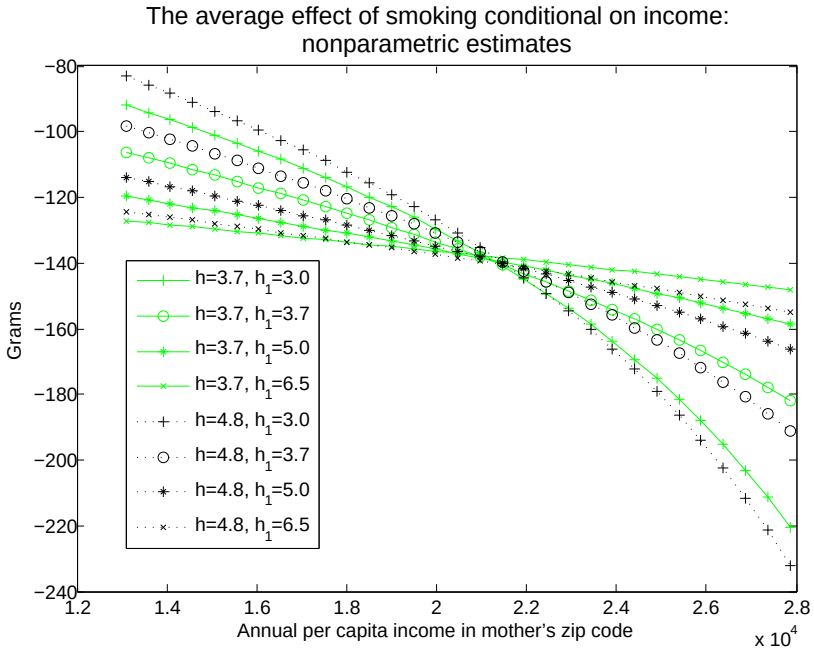

Figure 4: CATE as a function of per capita income: fully nonparametric estimates

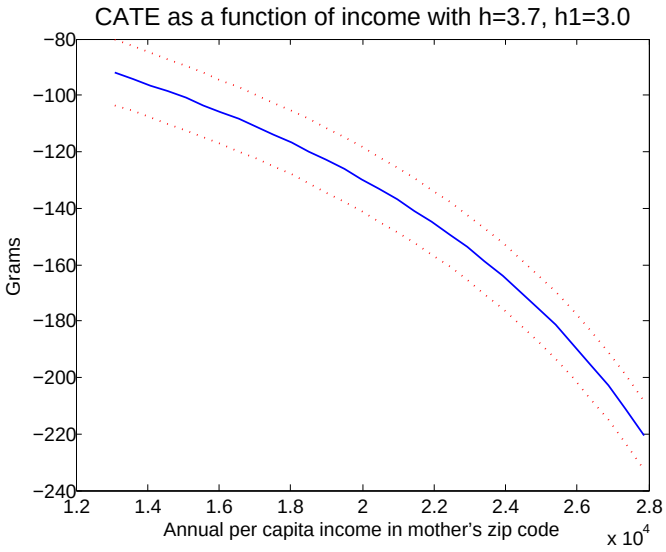

Figure 5: CATE as a function of per capita income: nonparametric estimate with $h = 3.7$, $h_1 = 3$ and ± 2 standard errors

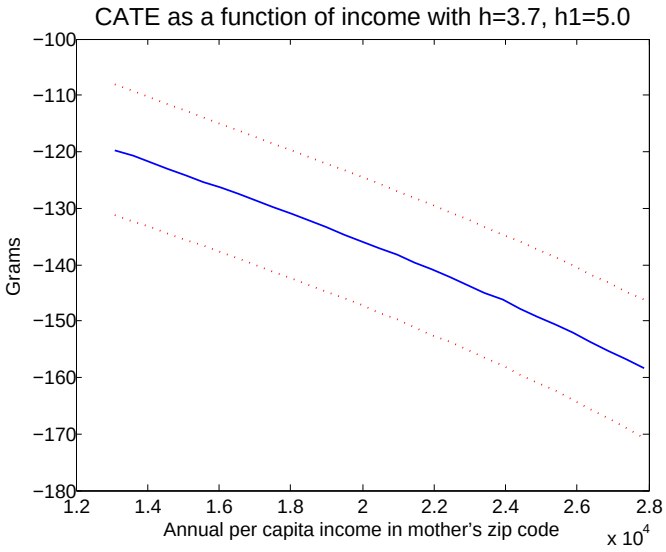

Figure 6: CATE as a function of per capita income: nonparametric estimate with $h = 3.7$, $h_1 = 5$ and ± 2 standard errors

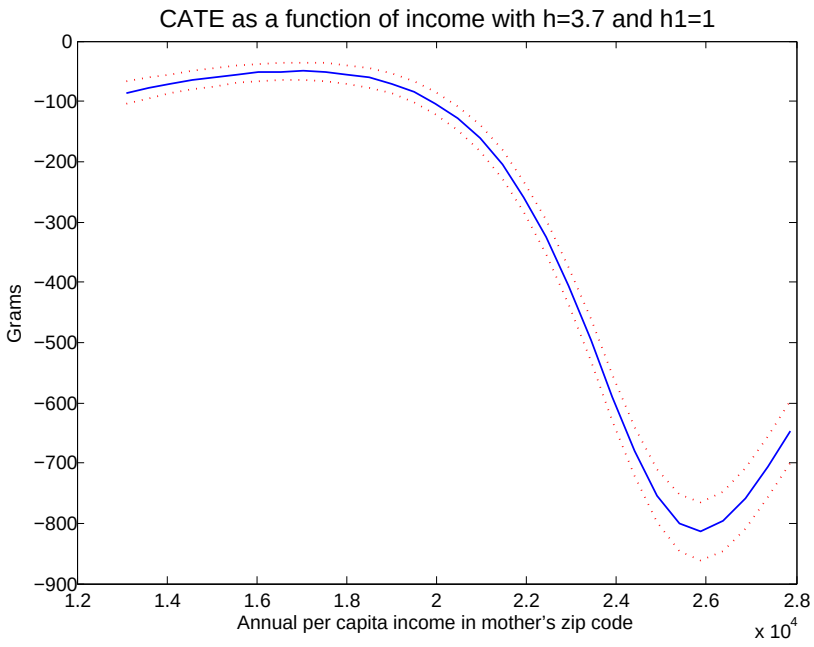

Figure 7: CATE as a function of per capita income: nonparametric estimate with $h = 3.7$, $h_1 = 1$ and ± 2 standard errors

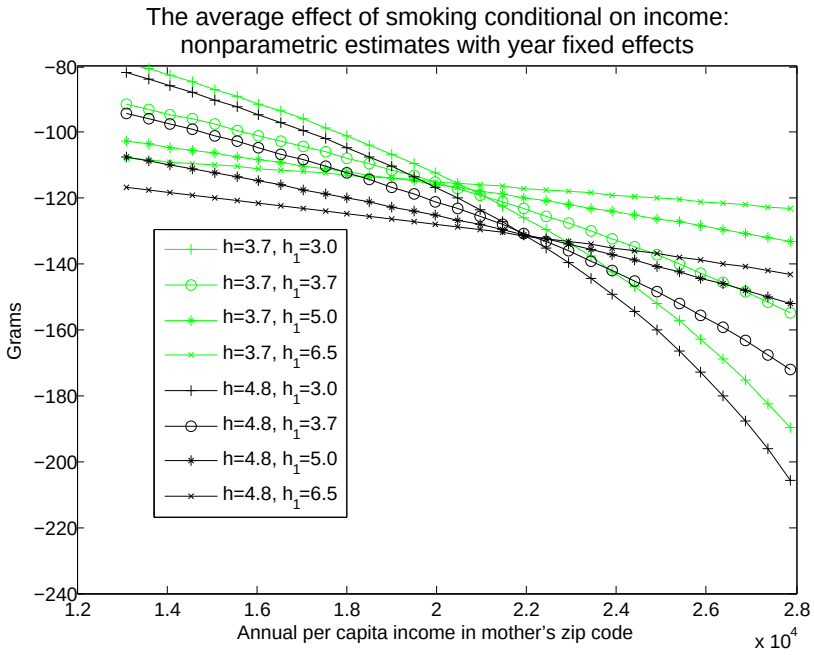

Figure 8: CATE as a function of per capita income: nonparametric estimates with year fixed effects

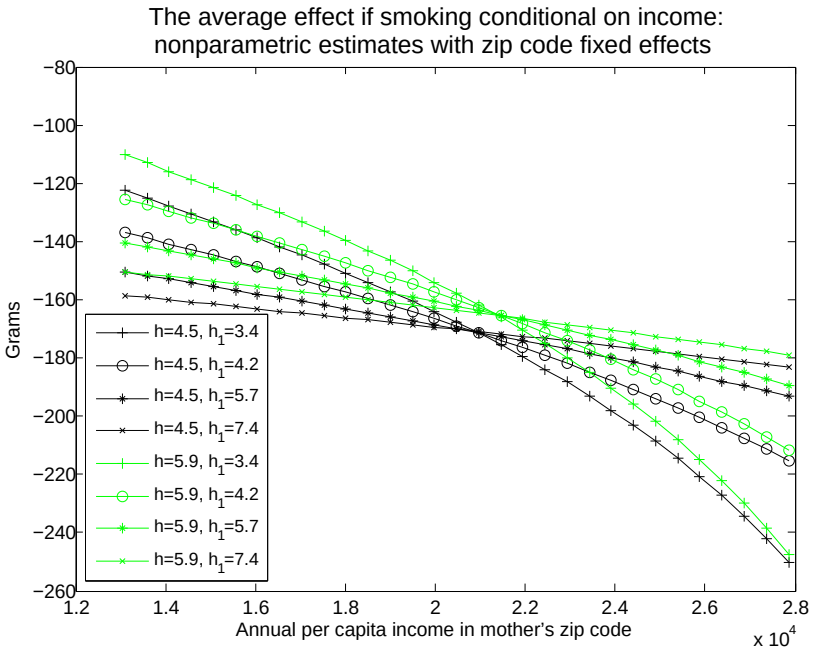

Figure 9: CATE as a function of per capita income: nonparametric estimates with zip code fixed effects