Journal of Econometrics 151 (2009) 33–46

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A specification test for the propensity score using its distribution conditional on participation☆

Azeem M. Shaikh [a], Marianne Simonsen [b], Edward J. Vytlacil [c,*], Nese Yildiz [d]

[a] Department of Economics, University of Chicago, United States
[b] School of Economics and Management, University of Aarhus, Denmark
[c] Department of Economics, Yale University, United States
[d] Department of Economics, University of Rochester, United States

ARTICLE INFO

Article history:
Received 6 March 2006
Received in revised form
4 August 2008
Accepted 31 January 2009
Available online 9 March 2009

Keywords:
Propensity score matching
Specification test
Nonparametric alternative

ABSTRACT

Propensity score matching has become a popular method for the estimation of average treatment effects. In empirical applications, researchers almost always impose a parametric model for the propensity score. This practice raises the possibility that the model for the propensity score is misspecified and therefore the propensity score matching estimator of the average treatment effect may be inconsistent. We show that the common practice of calculating estimates of the densities of the propensity score conditional on the participation decision provides a means for examining whether the propensity score is misspecified. In particular, we derive a restriction between the density of the propensity score among participants and the density among nonparticipants. We show that this restriction between the two conditional densities is equivalent to a particular orthogonality restriction and derive a formal test based upon it. The resulting test is shown via a simulation study to have dramatically greater power than competing tests for many alternatives. The principal disadvantage of this approach is loss of power against some alternatives.

1. Introduction

Propensity score matching is a widely used approach for the estimation of average treatment effects. See Heckman et al. (1999) for a detailed survey of its use. The method is based on the following well-known result of Rosenbaum and Rubin (1983): if selection into the program is independent of the potential outcomes of interest conditional on a vector of covariates, then selection into the program is also independent of the potential outcomes conditional on the propensity score, where the propensity score is the probability of selection into the program conditional on the same vector of covariates. The high dimensionality of the vector of observed covariates often forces researchers to adopt a parametric model for the propensity score, in lieu of a more flexible nonparametric model. This practice raises the possibility that the model for the propensity score is misspecified. If this is the case, then the estimator of the propensity score will be inconsistent and the resulting propensity score matching estimator of the average treatment effect may be inconsistent as well.

Following Heckman et al. (1998a), it is now common when implementing propensity score matching to calculate estimates of the densities of the propensity score conditional on the participation decision. These estimates are calculated in order to determine the region of common support on which to perform matching. The importance of common support was recognized as early as Rosenbaum and Rubin (1983). We show that these conditional densities provide a means for examining whether the propensity score is misspecified. In particular, we derive a restriction between the density of the propensity score among participants and the density among nonparticipants. Failure of this restriction to hold for the estimated conditional densities provides evidence that the model for the propensity score is misspecified. In this way, it provides a convenient diagnostic tool for detecting misspecification. We show further that this restriction between the two conditional densities is equivalent to a particular orthogonality restriction and derive a formal test based upon it. Unlike other tests for correct specification of the propensity score versus a nonparametric alternative, our test has dramatically greater power than competing tests for many alternatives. The principal drawback of this approach is that our test does not have power against certain alternatives, but we argue that these alternatives are rather exceptional.

The literature on propensity score matching includes instances in which researchers have used conventional tests of a parametric null hypothesis against a parametric alternative hypothesis to

☆ This research was supported by NSF SES-0832845.
* Corresponding author. Tel.: +1 203 4323244.
E-mail addresses: amshaikh@uchicago.edu (A.M. Shaikh), msimonsen@econ.au.dk (M. Simonsen), edward.vytlacil@yale.edu (E.J. Vytlacil), nyildiz@mail.rochester.edu (N. Yildiz).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.01.014

guide specification of the model for the propensity score. See, for example, Lechner (1999, 2000). In contrast to the tests in this literature, our test is a test of a parametric null hypothesis against a nonparametric alternative hypothesis.

The literature on propensity score matching has also used "balancing score" tests to detect misspecification of the model for the propensity score. These tests were first proposed by Rosenbaum and Rubin (1985), who suggest that researchers examine whether the observable characteristics of the population are independent of participation conditional on the propensity score. In practice, this idea is often implemented by researchers by examining whether moments of the observable characteristics for "matched" participant and nonparticipant observations are the same. See Dehejia and Wahba (2002), Lechner (2002), Sianesi (2004), and Smith and Todd (2005) for examples of such tests. As described above, our test differs from these tests in that it examines whether the conditional densities of the propensity score for the "unmatched" participants and nonparticipants differ in a certain way. In this sense, our test is in fact the opposite of a "balancing score" test.

To the best of our knowledge, we are the first to use the conditional distribution of the propensity score in the unmatched sample as a means of testing for misspecification of the model for the propensity score. In this way, our paper complements the literature on propensity score matching that uses the conditional distribution of the propensity score to provide insight into the degree of bias of naive estimators of average treatment effects that do not adjust for covariates.

Our paper proceeds as follows. In Section 2, we first provide a summary of the method of propensity score matching. We then derive in Section 3 the restriction upon which our test is based. In Section 4, we use this result to develop a formal test of misspecification. We examine its finite sample properties via a simulation study in Section 5. We provide an empirical illustration of our procedure in Section 6. Section 7 concludes.

2. Review of matching

Before proceeding, we review in this section matching as a means of program evaluation. There are two groups of individuals, participants and nonparticipants, in a program of interest. Participation in the program of interest is denoted by the dummy variable D, with $D = 1$ if the individual chooses to participate and $D = 0$ otherwise. Individuals in each of these two groups are associated with observed characteristics X. Two commonly used metrics for evaluating the effect of participation in a program are the average treatment effect, given by $E[Y_1 - Y_0]$, and the average treatment effect on the treated, given by $E[Y_1 - Y_0|D = 1]$, where Y_1 is the potential outcome in the case of participation and Y_0 is the potential outcome in the case of nonparticipation. The difficulty with estimation of these objects lies with the following missing data problem: the counterfactual outcome Y_{1-D} is never observed, which precludes direct estimation of $E[Y_0|D = 1]$ and $E[Y_1|D = 0]$.

The method of matching resolves this difficulty by matching each participant with a nonparticipant that is similar in terms of observed characteristics X. As described in Rosenbaum and Rubin (1983), matching formally requires that:

Assumption 2.1. $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.

Assumption 2.2. $0 < P(x) < 1$ where $P(x) = \Pr[D = 1|X = x]$ for all $x \in \text{supp}(X)$.

Note that a consequence of Assumption 2.2 is that $E[Y_0|D = 1, X = x]$ and $E[Y_1|D = 0, X = x]$ are well defined for all $x \in \text{supp}(X)$. Hence, matching suggests estimation of $E[Y_0|D = 1]$ and $E[Y_1|D = 0]$ by a two-step procedure in which $E[Y_0|D = 1, X]$ and $E[Y_1|D = 0, X]$ are first estimated by exploiting Assumption 2.2, and then integrated with respect to the empirical distribution of X in order to obtain an estimate of $E[Y_1 - Y_0]$ or integrated with respect to the empirical distribution of X conditional on $D = 1$ to obtain an estimate of $E[Y_1 - Y_0|D = 1]$. Following Heckman et al. (1998a), it is clear that one can relax the first condition to only require mean independence instead of full independence. Using this procedure, it is in principle possible to construct a $\sqrt{n}$-normal estimator of the parameter of interest without imposing any parametric restrictions. See Heckman et al. (1998b), Hahn (1998) and Abadie and Imbens (2007).

An alternative approach for estimating $E[Y_0|D = 1]$ and $E[Y_1|D = 0]$ is propensity score matching, which relies upon a celebrated result of Rosenbaum and Rubin (1983). They show that the above assumptions imply:

Assumption 2.3. $(Y_0, Y_1) \perp\!\!\!\perp D \mid P(X)$.

Assumption 2.4. $0 < \Pr[D = 1|P(X) = p] < 1$ for all $p \in \text{supp}(P(X))$.

This result can be restated as follows: if matching on X is valid, so is matching based on the propensity score $P(X)$. This result motivates propensity score matching, in which one first estimates the propensity score in a first step and then performs matching, as described above, using the estimated propensity score.

Importantly, both matching on X and matching on $P(X)$ suffer from the so-called "curse of dimensionality". While matching on X requires the researcher to estimate $E[Y_0|D = 1, X]$, matching on $P(X)$ requires the researcher to estimate $E[D|X]$, an equally high dimensional object. Thus, if X has more than a few dimensions, nonparametric procedures for estimating these objects are undesirable due to sizable finite sample bias. This difficulty leads to the common practice of implementing propensity score matching with a parametric model for the propensity score. This raises the possibility that the model for the propensity score is misspecified, in which case we would expect the estimator of the propensity score to be inconsistent. In this case, we would generally expect propensity score matching to lead to inconsistent estimates of both $E[Y_1 - Y_0]$ and $E[Y_1 - Y_0|D = 1]$.

3. Restriction between the conditional distributions

In order to ensure that the appropriate conditional expectations exist, propensity score matching is only valid over the region of common support of the densities of the propensity score conditional on the participation decision. Following the influential work of Heckman et al. (1998a), researchers therefore compute estimates of both of these densities to determine the appropriate region over which to perform matching. The test for misspecification that we develop in the following section will exploit a restriction that must hold between these two densities.

Throughout the following we will assume that the propensity score $P(X)$ has a density with respect to Lebesgue measure. Let $f(p)$ denote this density, $f_1(p)$ the density of the propensity score conditional on participation, and $f_0(p)$ the density of the propensity score conditional on nonparticipation. Using this notation, we have the following result:

Lemma 3.1. *Let* $\alpha = \frac{\Pr\{D=0\}}{\Pr\{D=1\}}$ *and assume* $0 < \Pr\{D = 0\} < 1$. *Then, for all* $0 < p < 1$ *and* $p \in \text{supp}(P)$ *we have that*

$$f_1(p) = \alpha \frac{p}{1-p} f_0(p). \tag{1}$$

Proof. Consider $0 < p < 1$ and $p \in \text{supp}(P)$. For such p, by the Law of Iterated Expectations, we have that

$$\Pr\{D = 1|P(X) = p\} = E[E[D|X]|P(X) = p] = p.$$

Bayes' Theorem implies further that

$$f_1(p)\Pr\{D=1\} = \Pr[D=1|P(X)=p]f(p) = pf(p) > 0\,.$$

Similarly, we have that

$$f_0(p)\Pr\{D=0\} = (1-p)f(p) > 0\,.$$

Combining these two implications, we have that

$$\frac{f_1(p)}{f_0(p)} = \alpha\frac{p}{1-p}\,,$$

from which the asserted conclusion follows immediately. ■

Lemma 3.1 implies that to the extent that the parametric model for $P(X)$ is correctly specified, we would expect (1) to hold approximately when estimated. The property can be easily checked given estimates $\hat{f}_{1,n}(\cdot)$ and $\hat{f}_{0,n}(\cdot)$ by graphing $\hat{f}_{1,n}(p)$ along with the function $\hat{\alpha}_n \frac{p}{1-p}\hat{f}_{0,n}(p)$, where $\hat{\alpha}_n$ is the sample analogue of α. Dramatic departures of one graph from another should be taken as evidence of misspecification of the propensity score. In this way, Lemma 3.1 provides a convenient diagnostic tool for detecting misspecification.

We now provide an illustration of this diagnostic for detecting misspecification of the propensity score. Define

$$X_1 = Z_1 \tag{2a}$$

$$X_2 = \frac{Z_1+Z_2}{\sqrt{2}}, \tag{2b}$$

where Z_1 and Z_2 are independent standard normal variables. Consider the model

$$D^* = 1 + X_1 + X_2 + X_1X_2 - \epsilon, \quad \epsilon \sim N(0,\sigma^2) \tag{3a}$$

$$D = I\{D^* > 0\}, \tag{3b}$$

where $\epsilon \perp\!\!\!\perp (X_1, X_2)$, and $I\{\cdot\}$ is the logical indicator function. We consider fitting a misspecified model

$$Q(X,\theta) = \Phi(\theta_0 + \theta_1X_1 + \theta_2X_2) \tag{4}$$

that only differs from the true model by omitting the quadratic term.

We generate an i.i.d. sequence of random variables $(D_i, X_{1,i}, X_{2,i}, \epsilon_i)$, $i = 1,\ldots,n$ for $n = 10{,}000$ according to the model given by (3a) and (3b). We then estimate the propensity score using the incorrectly specified model (4) and maximum likelihood. Estimates of the densities of the propensity score conditional on the participation decision are then constructed using kernel density estimation. For this purpose, we use the biweight kernel and choose the bandwidth using Silverman's rule of thumb. See Silverman (1986) for further details.

Fig. 1 depicts the resulting estimates of $\hat{f}_{1,n}(p)$ and $\hat{\alpha}_n \frac{p}{1-p}\hat{f}_{0,n}(p)$. We see that graphs of these two objects deviate substantially from each other, which can be taken as evidence of misspecification.

We also consider the case where the true data generating process is given by

$$D^* = 1 + X_1 + X_2 - \epsilon, \quad \epsilon \sim U(-1,1) \tag{5}$$

and we again fit the misspecified model (4). The estimated densities from this exercise are displayed in Fig. 2. We find that the two graphs lie very close to one another, suggesting that the misspecification is not very severe. This feature is perhaps not too surprising when one considers the fact that the only source of misspecification is the distribution for ϵ and the assumed distribution shares many features with the true distribution. In particular, both are symmetric about zero.

This example shows that misspecification of the propensity score may lead to noticeable departures from the restriction (1) when estimated. In order to develop a formal test based on this restriction, it will be useful to restate it in terms of an equivalent orthogonality condition. To this end, we have the following result:

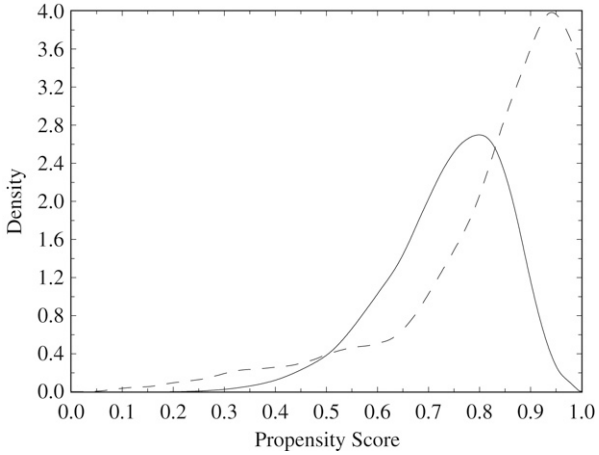

Fig. 1. Estimates of density of propensity score. Note: _____, $\hat{\alpha}_n \frac{p}{1-p}\hat{f}_{0,n}(p)$; – – – – –, $\hat{f}_{1,n}(p)$ DGP: $D^* = 1 + X_1 + X_2 + X_1X_2 - \epsilon_i, \epsilon \sim N(0,\sigma^2)$; estimated model: $Q(X,\theta) = \Phi(\theta_0 + \theta_1X_1 + \theta_2X_2)$. Density estimated using the biweight kernel and Silverman's rule of thumb.

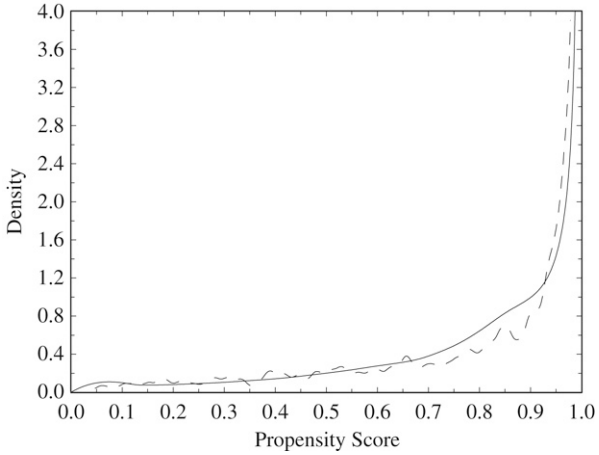

Fig. 2. Estimates of density of propensity score. Note: _____, $\hat{\alpha}_n \frac{p}{1-p}$; – – – – –, $\hat{f}_{1,n}(p)$ DGP: $D^* = 1 + X_1 + X_2 - \epsilon_i, \epsilon \sim U(-1,1)$; estimated model: $Q(X,\theta) = \Phi(\theta_0 + \theta_1X_1 + \theta_2X_2)$. Density estimated using the biweight kernel and Silverman's rule of thumb.

Lemma 3.2. *Let* $\alpha = \frac{\Pr\{D=0\}}{\Pr\{D=1\}}$ *and assume that* $0 < \Pr\{D=0\} < 1$. *Let* Q *be a random variable on the unit interval with density w.r.t. Lebesgue measure. Denote by* $g_1(q)$ *the density of* Q *conditional on* $D=1$ *and by* $g_0(q)$ *the density of* Q *conditional on* $D=0$. *Then,*

$$g_1(q) = \alpha\frac{q}{1-q}g_0(q) \tag{6}$$

for all $q \in \text{supp}(Q)$ *such that* $0 < q < 1$ *if and only if*

$$E[D-Q|Q=q] = 0 \tag{7}$$

for all $q \in \text{supp}(Q)$ *such that* $0 < q < 1$.

Proof. First consider necessity. Consider $q \in \text{supp}(Q)$ such that $0 < q < 1$. The restriction (7) is equivalent to

$$\Pr\{D=1|Q=q\} = E[D|Q=q] = q.$$

For such q, we therefore have by Bayes' Theorem that

$$g_1(q)\Pr\{D=1\} = \Pr\{D=1|Q=q\}g(q) = qg(q) > 0,$$

where $g(q)$ is the density of Q. Similarly, we have that

$$g_0(q)\Pr\{D=0\} = (1-q)g(q) > 0.$$

Combining these two implications, we have that

$$\frac{g_1(q)}{g_0(q)} = \alpha \frac{q}{1-q},$$

from which the asserted conclusion follows.

Now consider sufficiency. Consider $q \in \text{supp}(Q)$ such that $0 < q < 1$. For such q, we have by Bayes' Theorem that the restriction (6) implies that

$$E[D|Q=q] = \frac{g_1(q)\Pr\{D=1\}}{\Pr\{D=0\}g_0(q) + \Pr\{D=1\}g_1(q)} = q,$$

where the last equality follows from plugging in $g_1(q) = \alpha \times \frac{q}{1-q} g_0(q)$. It follows that $E[D-Q|Q=q] = q$, as desired. ■

It is important to observe that this characterization of the restriction only involves low dimensional conditional expectations. We will show in the following section that a consequence of this fact is that, unlike more conventional nonparametric tests for correct specification of the propensity score, our test will not suffer from a curse of dimensionality and will therefore have much greater power in finite samples against many alternatives. The principal drawback of this approach is that there will be certain alternatives for which we will not have power. To examine this issue further, consider the following example. Suppose $X = (X_1, X_2)$, and $E[D|X_1, X_2] \neq E[D|X_1]$. Let $Q(X) = E[D|X_1]$. Thus, $Q(X) \neq P(X)$, and yet $Q(X)$ will satisfy $E[D - Q(X)|Q(X)] = 0$. More generally, this restriction will not be able to detect the omission of covariates provided that the conditional expectation of D given the included covariates is correctly specified. Of course, this example is rather exceptional, since $Q(X)$ is in this case correctly specified for a subvector of X.

Remark 3.1. By integrating (1), we have immediately that for any appropriate functions $g : [0, 1] \to \mathbf{R}$,

$$E[g(P)|D=1] = \alpha E\left[\frac{P}{1-P} g(P)|D=0\right],$$

which implies further that

$$E\left[g(P)D - g(P)\frac{P}{1-P}(1-D)\right] = 0.$$

In particular, we have

$$E[P^k(1-P)D - P^{k+1}(1-D)] = 0$$

for $k = 0, 1, 2, \ldots$. As before, to the extent that the parametric model for $P(X)$ is correctly specified, we would expect these restrictions to hold approximately when estimated. One could, of course, develop a formal test based on these restrictions instead of (7), but we do not pursue that idea here. ■

4. Formal test

In this section we develop a formal test based on the restriction (7) in Lemma 3.2. We adapt the test proposed by Zheng (1996) for testing whether the parametric model for the conditional expectation is correctly specified. The advantages of this approach include the following: (i) the resulting test statistic has an asymptotically normal distribution under the null hypothesis, so the test is easy to implement; (ii) the power of the resulting test under local alternatives is easy to analyze; (iii) the test does not impose strong smoothness conditions on the alternative conditional expectation function; and (iv) the test does not require homoskedasticity of the generalized residual of the regression function. We could also have followed the analysis of Horowitz and Spokoiny (2001), which would result in a test with some desirable theoretical properties, but it would be more difficult to implement. On the other hand, we could not have followed the analysis of Hong and White (1995), since their analysis requires that the generalized residuals are homoskedastic or multiplicatively heteroskedastic, which rules out the case of a binary dependent variable. See Hart (2007) for a survey of these and other alternative approaches for testing a parametric null versus a nonparametric alternative.

Under the null hypothesis of Zheng (1996), the conditional expectation $E[Y|X]$ is assumed to belong to a parametric family of real valued functions $Q(X, \theta)$ on $\mathbf{R}^k \times \Theta$, where $\Theta \subseteq \mathbf{R}^d$. Concretely, his null and alternative hypotheses are given by

$$\begin{aligned} &H_0\text{: } \exists\, \theta_0 \in \Theta \text{ s.t. } \Pr\{E[Y|X] = Q(X, \theta_0)\} = 1. \\ &H_1\text{: } \Pr\{E[Y|X] = Q(X, \theta)\} < 1 \quad \forall\, \theta \in \Theta. \end{aligned}$$

Note that any θ_0 satisfying the null hypothesis also solves $\min_{\theta\in\Theta} E[(Y - Q(X, \theta))^2]$.

Zheng (1996) proposes a test of the above null hypothesis based on the following idea. Let $\epsilon = Y - Q(X, \theta_0)$ and let $f_X(\cdot)$ denote the density of X. Then, under the null hypothesis,

$$E[\epsilon E[\epsilon|X] f_X(X)] = 0, \tag{8}$$

while under the alternative hypothesis,

$$E[\epsilon E[\epsilon|X] f_X(X)] = E[[E[\epsilon|X]]^2 f_X(X)] > 0.$$

The last inequality follows because under the alternative hypothesis $(E[\epsilon|X])^2 > 0$ with positive probability. On the basis of this observation, he uses the sample analogue of the left-hand side of (8) to form his test. The test statistic is given by

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} \frac{1}{h^k} K\left(\frac{X_i - X_j}{h}\right) \hat{\varepsilon}_i \hat{\varepsilon}_j,$$

where $\hat{\varepsilon}_i := Y_i - Q(X_i, \hat{\theta}_n)$, $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of $\arg\min_{\theta\in\Theta} E[(Y - Q(X, \theta))^2]$, h is a smoothing parameter, and $K(\cdot)$ is a kernel.

We would like to test whether there exists $\theta_0 \in \Theta$ such that $E[D|Q(X, \theta_0)] = Q(X, \theta_0)$ with probability 1. Our null and alternative hypotheses are therefore given by

$$\begin{aligned} &H_0\text{: } \exists\, \theta_0 \in \Theta \text{ s.t. } \Pr\{E[D|Q(X, \theta_0)] = Q(X, \theta_0)\} = 1. \\ &H_1\text{: } \Pr\{E[D|Q(X, \theta)] = Q(X, \theta)\} < 1 \quad \forall \theta \in \Theta. \end{aligned}$$

Note that this differs from the framework of Zheng (1996) in that the variable that is conditioned on is not observed. By analogy with the test statistic in Zheng (1996), we consider testing based upon

$$\hat{V}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} \frac{1}{h} K\left(\frac{Q(X_i, \hat{\theta}_n) - Q(X_j, \hat{\theta}_n)}{h}\right) \hat{\varepsilon}_i \hat{\varepsilon}_j \tag{9}$$

where $\hat{\varepsilon}_i = D_i - Q(X_i, \hat{\theta}_n)$ and $\hat{\varepsilon}_i, \hat{\theta}_n, h$ and $K(\cdot)$ are as defined above. Unfortunately, the analysis of Zheng (1996) does not apply directly to this case, and thus we now extend his analysis. We will impose the following conditions:

Assumption 4.1. (D_i, X_i), $i = 1, \ldots, n$, is an i.i.d. sequence of random variables on $\{0, 1\} \times \mathbf{R}^k$.

Assumption 4.2. Θ is a compact subset of $\mathbf{R}^d$.

Assumption 4.3. $Q : \text{supp}(X_i) \times \Theta \to [0, 1]$ satisfies:

(a) $Q(X_i, \theta)$ has a continuous density $f(x, \theta)$ w.r.t. Lebesgue measure for all θ in a neighborhood of θ_0, where

$$\theta_0 = \arg\min_{\theta\in\Theta} E[(D_i - Q(X_i, \theta))^2].$$

(b) $Q(x,\theta)$ is Lipschitz continuous w.r.t. θ in the sense that for all $\theta \in \Theta$ and $\theta' \in \Theta$

$$|Q(x,\theta) - Q(x,\theta')| \leq G(x)\|\theta - \theta'\|,$$

where $E[G^{4+\delta}(X_i)] < \infty$ for some $\delta > 0$.

Assumption 4.4. $K : \mathbf{R} \to \mathbf{R}$ is bounded, Lipschitz continuous, symmetric and satisfies:

(a) $\int K(u)\mathrm{d}u = 1$.
(b) $\int |K(u)|\mathrm{d}u < \infty$.
(c) $\int |uK(u)|\mathrm{d}u < \infty$.
(d) $\int |u^2K(u)|\mathrm{d}u < \infty$.

Assumption 4.5. $\hat{\theta}_n$ satisfies

$$\|\hat{\theta}_n - \theta_0\| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Assumption 4.6. The bandwidth sequence satisfies $0 < h = h_n \to 0$ and $nh^4 \to \infty$.

The following extension of Zheng's analysis describes the behavior of our test statistic under the null hypothesis.

Theorem 4.1. *Suppose Assumptions 4.1–4.6. If $E[D_i|X_i] = Q(X_i,\theta_0)$ with probability 1, then $\hat{V}_n$ defined by (9) satisfies*

$$n\sqrt{h}\hat{V}_n \to N(0,\Sigma),$$

where

$$\Sigma = 2\iint q_1^2(1-q_1)^2K^2(u)f^2(q_1)\mathrm{d}u\mathrm{d}q_1.$$

Moreover, Σ may be consistently estimated by

$$\hat{\Sigma}_n = \sum_{1\leq i,j\leq n: i\neq j} \frac{2\hat{\epsilon}_i^2\hat{\epsilon}_j^2}{n(n-1)h}K^2\left(\frac{Q(X_i,\hat{\theta}_n) - Q(X_j,\hat{\theta}_n)}{h}\right). \tag{10}$$

Proof. See Appendix. ■

The following theorem shows that our test is consistent against any fixed alternative.

Theorem 4.2. *Suppose Assumptions 4.1–4.6. If $E[D_i|X_i] = Q(X_i,\theta_0)$ with probability strictly less than 1, then $\hat{V}_n$ and $\hat{\Sigma}_n$ defined by (9) and (10), respectively, satisfy*

$$\hat{V}_n \xrightarrow{P} \int (r(q_1) - q_1)^2 f^2(q_1)\mathrm{d}q_1 > 0$$

$$\hat{\Sigma}_n \xrightarrow{P} 2\int (r(q_1) - 2r(q_1)q_1 + q_1^2)^2K(u)f^2(q_1)\mathrm{d}q_1.$$

Proof. See Appendix. ■

Finally, in the following theorem, we investigate the power of our test under local alternatives of the form

$$\mathrm{H}_1^n\text{: } \Pr\{E[D|Q(X,\theta_0)] = Q(X,\theta_0) + \tfrac{1}{\sqrt{nh^{1/4}}}\ell(Q(X,\theta_0))\} = 1$$

for some function $\ell : [0,1] \to \mathbf{R}$.

Theorem 4.3. *Suppose Assumptions 4.1–4.6. Let $\ell : [0,1] \to \mathbf{R}$ be a continuous function and $r(q) = E[D_i|Q(X_i,\theta_0) = q]$. If*

$$E[D_i|X_i] = Q(X_i,\theta_0) + \frac{1}{\sqrt{nh^{1/4}}}\ell(Q(X_i,\theta_0))$$

with probability 1, then $\hat{V}_n$ defined by (9) satisfies

$$n\sqrt{h}\hat{V}_n \to N\left(\int \ell^2(q_1)f^2(q_1)\mathrm{d}q_1, \Sigma\right),$$

where

$$\Sigma = 2\iint r(q_1)^2(1-r(q_1))^2K^2(u)f^2(q_1)\mathrm{d}u\mathrm{d}q_1. \tag{11}$$

Proof. See Appendix. ■

Remark 4.1. Assumption 4.4(a) is exploited in the proofs of Theorems 4.2 and 4.3, but is not exploited in the proof of Theorem 4.1. Thus, the stated behavior of our test statistic under the null hypothesis will continue to hold if the kernel does not integrate to 1. Likewise, Assumption 4.3(b) is exploited in the proofs of Theorems 4.1 and 4.3, but the proof of Theorem 4.2 only requires the weaker restriction that $E[G(X_i)] < \infty$. ■

5. Simulation study

As noted earlier, there are alternatives against which the test of Zheng (1996) has power tending to one, but our test does not. Zheng (1996), however, requires estimation of expectations conditional on X, which in practice is high dimensional. Our procedure only requires estimation of expectations conditional on $Q(X,\theta_0)$ and thereby avoids this difficulty. As a result, we expect our test to perform noticeably better in finite samples against many alternatives of interest.

We now shed some light on the finite sample properties of our testing procedure via a simulation study. Our setup will follow Zheng (1996) closely. As before, let X_1 and X_2 be given by (2a) and (2b). We define now several different data generating processes for D^* that will be used at different points in our simulation study.

$$D^* = 1 + X_1 + X_2 - \epsilon, \quad \epsilon \sim N(0,\sigma^2) \tag{12}$$

$$D^* = 1 + X_1 + X_2 + X_1X_2 - \epsilon, \quad \epsilon \sim N(0,\sigma^2) \tag{13}$$

$$D^* = (1 + X_1 + X_2)^2 - \epsilon, \quad \epsilon \sim N(0,\sigma^2) \tag{14}$$

$$D^* = 1 + X_1 + X_2 - \epsilon, \quad \epsilon \sim \chi_1^2 \tag{15}$$

$$D^* = 1 + X_1 + X_2 - \epsilon, \quad \epsilon \sim U(-1,1). \tag{16}$$

For each of these data generating processes, $D = I\{D^* > 0\}$ and $\epsilon \perp\!\!\!\perp (X_1, X_2)$.

Throughout the simulations presented below, we use the normal kernel given by

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right).$$

The bandwidth h is chosen to be equal to $cn^{-1/8}$ for c equal to 0.05, 0.10, and 0.15. We consider samples sizes n equal to 100, 200, 400, 500, 800, and 1000. The number of replications for each simulation is always 1000. Note that the bandwidth satisfies the requirements from Theorem 4.1.

We first examine the size of our test. The data is generated according to (12). The null and alternative hypotheses are as described in Section 4 with $Q(X,\theta) = \Phi(\theta_0 + \theta_1X_1 + \theta_2X_2)$. These results are summarized in Table 1. We find that the actual finite sample size of the test in our Monte Carlo simulations is close to but slightly smaller than the nominal size, so the test is slightly conservative.

We now go on to consider the power of our test against certain misspecifications of the model for the propensity score. In Tables 2–5, we consider four different scenarios in which the true data generating process is given by (13)–(16), respectively. For each scenario, the null and alternative hypotheses are as before. The test performs admirably when the true data generating process is given by (13)–(15), showing high power for even moderately

Table 1
Size.

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.010	0.500	0.113
200	0.008	0.042	0.104
400	0.005	0.032	0.081
500	0.005	0.033	0.080
800	0.010	0.037	0.086
1000	0.007	0.039	0.089
$c = 0.10$			
100	0.008	0.026	0.081
200	0.005	0.012	0.059
400	0.005	0.017	0.052
500	0.005	0.013	0.055
800	0.005	0.011	0.060
1000	0.008	0.020	0.078
$c = 0.15$			
100	0.006	0.010	0.044
200	0.001	0.008	0.026
400	0.003	0.012	0.033
500	0.002	0.011	0.026
800	0.002	0.003	0.021
1000	0.006	0.014	0.045

DGP: $D^* = 1 + X_1 + X_2 - \epsilon_i, \epsilon \sim N(0, \sigma^2)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$.

Table 2
Power.

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.078	0.151	0.216
200	0.354	0.496	0.586
400	0.879	0.929	0.960
500	0.966	0.983	0.989
800	1.000	1.000	1.000
1000	1.000	1.000	1.000
$c = 0.10$			
100	0.108	0.186	0.253
200	0.458	0.592	0.664
400	0.936	0.973	0.978
500	0.987	0.993	0.995
800	1.000	1.000	1.000
1000	1.000	1.000	1.000
$c = 0.15$			
100	0.097	0.168	0.211
200	0.449	0.593	0.658
400	0.947	0.972	0.984
500	0.990	0.993	0.995
800	1.000	1.000	1.000
1000	1.000	1.000	1.000

DGP: $D^* = 1 + X_1 + X_2 + X_1X_2 - \epsilon_i, \epsilon \sim N(0, \sigma^2)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$.

Table 3
Power.

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.357	0.468	0.538
200	0.813	0.868	0.890
400	0.992	0.995	0.996
500	1.000	1.000	1.000
800	1.000	1.000	1.000
1000	1.000	1.000	1.000
$c = 0.10$			
100	0.204	0.308	0.371
200	0.712	0.801	0.844
400	0.989	0.993	0.997
500	1.000	1.000	1.000
800	1.000	1.000	1.000
1000	1.000	1.000	1.000
$c = 0.15$			
100	0.081	0.133	0.179
200	0.455	0.574	0.639
400	0.960	0.983	0.988
500	0.996	0.999	0.999
800	1.000	1.000	1.000
1000	1.000	1.000	1.000

DGP: $D^* = (1 + X_1 + X_2)^2 - \epsilon_i, \epsilon \sim N(0, \sigma^2)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$.

Table 4
Power.

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.029	0.075	0.144
200	0.110	0.208	0.279
400	0.369	0.569	0.651
500	0.570	0.741	0.796
800	0.924	0.967	0.976
1000	0.986	0.994	0.996
$c = 0.10$			
100	0.038	0.077	0.129
200	0.162	0.269	0.330
400	0.528	0.683	0.756
500	0.713	0.806	0.851
800	0.968	0.982	0.988
1000	0.993	0.996	0.998
$c = 0.15$			
100	0.041	0.068	0.111
200	0.171	0.269	0.338
400	0.546	0.702	0.756
500	0.721	0.813	0.860
800	0.967	0.982	0.989
1000	0.994	0.997	0.999

DGP: $D^* = 1 + X_1 + X_2 - \epsilon_i, \epsilon \sim \xi_1^2$, estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$.

sized samples. Recall that (13) is precisely that of our heuristic from Lemma 3.1 presented in Fig. 1. Given the noticeable departure of the two graphs in Fig. 1, it is not surprising that our test performs well. The test performs less well, however, when the true data generating process is given by (16), which is again consonant with our earlier findings in Fig. 2.

Finally, we compare our test with the test proposed by Zheng (1996). First, we consider the same setup of Table 2, where the data generating process is given by (13). The results are presented in Table 6. As noted earlier, the competing test requires estimation of expectations conditional on X, which in practice is high dimensional, and so we might expect that our test would perform better in finite samples for many alternatives. Indeed, this feature is borne out by our simulation: our test rejects the null hypothesis much more frequently for nearly all sample sizes. To investigate the severity of this problem, we consider a further model which includes an additional covariate. Define

$$X_3 = \frac{Z_1 + Z_3}{\sqrt{2}},$$

where Z_3 is a standard normal random variable independent of Z_1 and Z_2. The data generating process for D^* in this instance is given by

$$D^* = 1 + X_1 + X_2 + X_1X_2 + X_3 - \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

where $\epsilon \perp\!\!\!\perp (X_1, X_2, X_3)$. The null and alternative hypotheses are as before, but with $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3)$. The results

Table 5
Power.

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.001	0.060	0.114
200	0.007	0.030	0.075
400	0.006	0.041	0.085
500	0.009	0.034	0.072
800	0.031	0.078	0.126
1000	0.047	0.109	0.177
$c = 0.10$			
100	0.007	0.040	0.106
200	0.007	0.021	0.057
400	0.009	0.034	0.066
500	0.013	0.034	0.070
800	0.037	0.079	0.125
1000	0.054	0.113	0.170
$c = 0.15$			
100	0.003	0.023	0.062
200	0.005	0.016	0.038
400	0.008	0.026	0.050
500	0.013	0.034	0.059
800	0.028	0.072	0.105
1000	0.055	0.106	0.152

DGP: $D^* = 1 + X_1 + X_2 - \epsilon_i, \epsilon \sim U(-1, 1)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$.

Table 6
Power, Zheng (1996).

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.002	0.031	0.075
200	0.001	0.034	0.097
400	0.007	0.048	0.102
500	0.014	0.069	0.125
800	0.031	0.108	0.172
1000	0.035	0.128	0.214
$c = 0.10$			
100	0.003	0.040	0.108
200	0.008	0.061	0.115
400	0.038	0.112	0.188
500	0.049	0.152	0.245
800	0.139	0.315	0.423
1000	0.220	0.455	0.572
$c = 0.15$			
100	0.006	0.047	0.105
200	0.018	0.081	0.144
400	0.090	0.199	0.282
500	0.121	0.291	0.396
800	0.360	0.547	0.663
1000	0.547	0.744	0.822

DGP: $D^* = 1 + X_1 + X_2 + X_1X_2 - \epsilon_i, \epsilon \sim N(0, \sigma^2)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$.

Table 7
Power.

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.036	0.104	0.149
200	0.160	0.266	0.342
400	0.494	0.627	0.695
500	0.670	0.780	0.829
800	0.922	0.962	0.979
1000	0.980	0.991	0.995
$c = 0.10$			
100	0.063	0.118	0.158
200	0.252	0.379	0.437
400	0.628	0.743	0.807
500	0.784	0.870	0.908
800	0.973	0.991	0.994
1000	0.994	0.996	0.998
$c = 0.15$			
100	0.068	0.110	0.153
200	0.279	0.402	0.449
400	0.682	0.791	0.831
500	0.821	0.894	0.922
800	0.987	0.994	0.998
1000	0.996	0.996	1.000

DGP: $D^* = 1 + X_1 + X_2 + X_1X_2 + X_3 - \epsilon_i, \epsilon \sim N(0, \sigma^2)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3)$.

Table 8
Power, Zheng (1996).

n	Proportion of rejections		
	1% level	5% level	10% level
$c = 0.05$			
100	0.001	0.002	0.003
200	0.000	0.004	0.017
400	0.001	0.010	0.050
500	0.001	0.008	0.054
800	0.003	0.024	0.081
1000	0.001	0.033	0.104
$c = 0.10$			
100	0.000	0.008	0.042
200	0.002	0.021	0.073
400	0.004	0.037	0.096
500	0.005	0.038	0.089
800	0.004	0.051	0.110
1000	0.019	0.072	0.135
$c = 0.15$			
100	0.001	0.020	0.067
200	0.006	0.041	0.096
400	0.010	0.048	0.107
500	0.008	0.065	0.113
800	0.019	0.081	0.165
1000	0.044	0.143	0.220

DGP: $D^* = 1 + X_1 + X_2 + X_1X_2 + X_3 - \epsilon_i, \epsilon \sim N(0, \sigma^2)$,
estimated model: $Q(X, \theta) = \Phi(\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3)$.

of this comparison are presented in Tables 7 and 8. In this case, with the additional covariate, we find that our test retains high power at all sample sizes, but the highest power of the competing test is approximately 20% and often much lower. Thus, we believe that our test will be of great use in practice, especially when X is high dimensional. Of course, it should be emphasized again that this advantage of our test comes at the expense of power against certain alternatives.

6. Empirical illustration

In this section, we provide an empirical illustration of our testing procedure. Specifically, we model the probability of motherhood in a subsample of the Danish population and test for misspecification. Simonsen and Skipper (2006) use this as an input into a matching analysis of the effects of motherhood on wages.

The available data set contains information on a representative sample of 5% of all Danish individuals in the 15–74 age bracket. Information stems from several registers all maintained by Statistics Denmark. The registers include variables describing socio-economic status on a yearly basis. In the empirical analysis below, we use a 1997 cross-sectional subsample of about 29,000 women aged 20–40 years, who are employed more than 200 hours per year, who are not self-employed, and not undertaking education.

Table 9
Descriptive statistics, selected variables.

Variables	Women		
	All	Mothers	Non-mothers
Log wages	4.80	4.81	4.79
	(0.28)	(0.26)	(0.30)
Age (years)	29.72	32.87	25.88
	(5.80)	(4.20)	(5.13)
Length of completed education (years)	12.23	12.20	12.26
	(2.45)	(2.42)	(2.48)
Type of highest completed education:			
General (0/1)	0.22	0.19	0.26
Business (0/1)	0.34	0.33	0.35
Industry (0/1)	0.01	0.01	0.01
Construction (0/1)	0.01	0.01	0.01
Graphical (0/1)	0.01	0.01	0.01
Services (0/1)	0.02	0.02	0.02
Food and beverages (0/1)	0.04	0.04	0.03
Agricultural (0/1)	0.01	0.01	0.01
Transportation (0/1)	0.00	0.00	0.00
Health	0.12	0.15	0.07
Pedagogic (0/1)	0.06	0.09	0.04
Humanistic (0/1)	0.03	0.03	0.03
Musical (0/1)	0.00	0.00	0.00
Social (0/1)	0.04	0.03	0.05
Technical (0/1)	0.02	0.01	0.02
Public security (0/1)	0.00	0.00	0.00
Unknown (0/1)	0.08	0.07	0.09
Province (0/1)	0.63	0.68	0.57
# observations	29,006	15,958	13,048

Table 9 shows selected descriptive statistics for the sample used in our analysis along with descriptive statistics of mothers and non-mothers. We classify women as mothers if they have given birth to a child. 54.8% of the women in our sample were mothers in 1996. It is clear that mothers differ significantly from non-mothers in terms of observables: mothers are on average seven years older and are more likely to have an education directed towards the health care sector or the schooling system. Furthermore, they are more often settled outside Greater Copenhagen.

We model the propensity score by a standard probit and consider two specifications:

Specification 1: X includes age dummies (20 or below, 20–22, 22–24, 24–26, 26–28, 28–30, 30–32, 32–34, 34–36, 36–38, above 38), type of education, length of education given type, nine regional dummies indicating place of habitation, and interaction terms between age and type of education. This is the specification used in Simonsen and Skipper (2006).
Specification 2: X is defined as in Specification 1 but interaction terms between age and type of education are excluded.

For both of these specifications, we again test the null and alternative hypotheses described in Section 4 with $Q(X,\theta) = \Phi(X\theta)$. Table 10 shows the test results.

It is worthwhile to point out that the heuristic from Section 3 is not helpful in this case in the sense that the estimate $\hat{\alpha}_n \frac{p}{1-p}\hat{f}_{0,n}(p)$ closely resembles $\hat{f}_{1,n}(p)$ regardless of the specification. For this reason, we must rely upon the formal methodology developed in Section 4. At the 5% significance level, we see that Specification 1 cannot be rejected, whereas Specifications 2 is clearly rejected in most cases. It is reassuring that these conclusions are largely insensitive to the choice of bandwidth, c.

7. Conclusion

In this paper, we have shown that the commonly computed estimates of the densities of the propensity score conditional on participation provide a means of examining whether the parametric model for the propensity score is correctly specified. In particular, correct specification of the propensity score implies that a certain restriction between the estimated conditional densities must hold. We have shown further that this restriction is equivalent to an orthogonality restriction, which can be used as the basis of a formal test for correct specification. While our test does not have power against all forms of misspecification of the propensity score, we argue that for a large class of alternatives our test will perform better in finite samples than existing tests that have power against all forms of misspecification. Our simulation study of the finite sample behavior of our test corroborates this claim. Since our test is also easily implemented, it is our hope that this work will persuade researchers to examine the specification of their model for the propensity score, as its validity is essential for consistency of their estimators.

Appendix

Throughout the following, we will use the notation $a \ll b$ to indicate that $a \leq cb$ for some constant $c > 0$.

A.1. *Proof of* Theorem 4.1

Let

$$\epsilon_i = D_i - Q(X_i, \theta_0)$$

and let $f(x) = f(x, \theta_0)$. Note that

$$\hat{V}_n = V_{1,n} - 2V_{2,n} + V_{3,n} + V_{4,n} + V_{5,n},$$

where

$$V_{1,n} = \sum_{1\leq i,j\leq n: i\neq j} \frac{\epsilon_i\epsilon_j}{n(n-1)h} K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right) \tag{17}$$

$$V_{2,n} = \sum_{1\leq i,j\leq n: i\neq j} \frac{\epsilon_i(Q(X_j,\hat{\theta}_n) - Q(X_j,\theta_0))}{n(n-1)h} \times K\left(\frac{Q(X_i,\hat{\theta}_n) - Q(X_j,\hat{\theta}_n)}{h}\right) \tag{18}$$

$$V_{3,n} = \sum_{1\leq i,j\leq n: i\neq j} \frac{(Q(X_i,\hat{\theta}_n) - Q(X_i,\theta_0))(Q(X_j,\hat{\theta}_n) - Q(X_j,\theta_0))}{n(n-1)h} \times K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right) \tag{19}$$

$$V_{4,n} = \sum_{1\leq i,j\leq n: i\neq j} \frac{\epsilon_i\epsilon_j}{n(n-1)h}\left[K\left(\frac{Q(X_i,\hat{\theta}_n) - Q(X_j,\hat{\theta}_n)}{h}\right) - K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right)\right] \tag{20}$$

$$V_{5,n} = \sum_{1\leq i,j\leq n: i\neq j} \frac{(Q(X_i,\hat{\theta}_n) - Q(X_i,\theta_0))(Q(X_j,\hat{\theta}_n) - Q(X_j,\theta_0))}{n(n-1)h} \times \left[K\left(\frac{Q(X_i,\hat{\theta}_n) - Q(X_j,\hat{\theta}_n)}{h}\right) - K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right)\right]. \tag{21}$$

We now analyze each of these terms separately.

Analysis of $V_{1,n}$:

Let $Z_i = (\epsilon_i, X_i)$ and note that $V_{1,n}$ may be written as a U-statistic with kernel

$$H_n(Z_i, Z_j) = \frac{\epsilon_i\epsilon_j}{h} K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right).$$

Table 10
Results from specification tests.

	Bandwidth $c = 0.05$		Bandwidth $c = 0.10$		Bandwidth $c = 0.15$		Share of correct predictions	Pseudo-R^2
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value		
1. Full model	1.73	0.08	1.58	0.11	1.24	0.21	0.80	0.36
2. No cross-terms	3.61	0.00	4.58	0.00	5.34	0.00	0.80	0.35

In order to establish the asymptotic properties of $V_{1,n}$, we apply Lemma 3.2 of Zheng (1996). To this end, first note that $E[H_n(Z_i, Z_j)|Z_i] = 0$. Note further that

$$\begin{aligned} E[H_n^2(Z_i, Z_j)] &\ll \iint \frac{1}{h^2} K^2\left(\frac{q_1 - q_2}{h}\right) f(q_1) f(q_2) \mathrm{d}q_1 \mathrm{d}q_2 \\ &\ll \iint \frac{1}{h} |K(u)| f(q) f(q - uh) \mathrm{d}q \mathrm{d}u \\ &\ll \iint \frac{1}{h} |K(u)| f(q) \mathrm{d}q \mathrm{d}u = O\left(\frac{1}{h}\right) \end{aligned}$$

where the first inequality follows from the boundedness of ϵ_i, the second follows from the boundedness of $K(\cdot)$, and the third follows from the boundedness of $f(\cdot)$. Similarly, we have that

$$E[H_n^4(Z_i, Z_j)] = O\left(\frac{1}{h^3}\right).$$

Let

$$G_n(Z_i, Z_j) = E[H_n(Z_k, Z_i) H_n(Z_k, Z_j) | Z_i, Z_j].$$

Note that

$$\begin{aligned} E[G_n^2(Z_i, Z_j)] &\ll E\Bigg[E\Bigg[\frac{1}{h^2}\left|K\left(\frac{Q(X_k, \theta_0) - Q(X_i, \theta_0)}{h}\right)\right| \\ &\times \left|K\left(\frac{Q(X_k, \theta_0) - Q(X_j, \theta_0)}{h}\right)\right| \Bigg| Z_i, Z_j\Bigg]^2\Bigg] \\ &= \iint \left[\int \frac{1}{h^2}\left|K\left(\frac{q_3 - q_1}{h}\right)\right| \left|K\left(\frac{q_3 - q_2}{h}\right)\right| f(q_3) \mathrm{d}q_3\right]^2 \\ &\times f(q_1) f(q_2) \mathrm{d}q_1 \mathrm{d}q_2 \\ &= \frac{1}{h} \iint \left[\int |K(u)||K(u + v)| f(q_1 + uh) \mathrm{d}u\right]^2 \\ &\times f(q_1 - vh) f(q_1) \mathrm{d}v \mathrm{d}q_1 \\ &\ll \frac{1}{h} \iint \left[\int |K(u)||K(u + v)| \mathrm{d}u\right]^2 f(q_1) \mathrm{d}v \mathrm{d}q_1, \end{aligned}$$

where the first inequality follows from the boundedness of ϵ_i and the second inequality follows from the boundedness of $f(\cdot)$. Note further that

$$\begin{aligned} &\int \left[\int |K(u)||K(u + v)| \mathrm{d}u\right]^2 \mathrm{d}v \le \iint K^2(u) K^2(u + v) \mathrm{d}v \mathrm{d}u \\ &\ll \int |K(u)| \left[\int |K(u + v)| \mathrm{d}v\right] \mathrm{d}u \\ &\le \int |K(u)| \left[\int |K(z)| \mathrm{d}z\right] \mathrm{d}u < \infty \end{aligned}$$

where the first inequality follows from Jensen's inequality and Tonelli's Theorem and the second inequality follows from the boundedness of $K(\cdot)$. Thus,

$$E[G_n^2(Z_i, Z_j)] = O\left(\frac{1}{h}\right).$$

It follows that

$$\frac{E[G_n^2(Z_i, Z_j)] + E[H_n^4(Z_i, Z_j)]/n}{E[H_n^2(Z_i, Z_j)]^2} \to 0,$$

so by Lemma 3.2 of Zheng (1996),

$$n\sqrt{h} V_{1,n} \xrightarrow{d} N(0, \Sigma),$$

where

$$\Sigma = \lim_{n \to \infty} 2hE[H_n^2(Z_1, Z_2)].$$

Note that

$$\begin{aligned} 2hE[H_n^2(Z_i, Z_j)] &= 2 \iint \frac{q_1(1 - q_1) q_2(1 - q_2)}{h} K^2\left(\frac{q_1 - q_2}{h}\right) \\ &\times f(q_1) f(q_2) \mathrm{d}q_1 \mathrm{d}q_2 \\ &= 2 \iint q_1(1 - q_1)(q_1 - uh)(1 - q_1 + uh) K^2(u) f(q_1) \\ &\times f(q_1 - uh) \mathrm{d}u \mathrm{d}q_1 \\ &= 2 \iint q_1^2 (1 - q_1)^2 K^2(u) f(q_1) f(q_1 - uh) \mathrm{d}u \mathrm{d}q_1 \\ &- h \iint q_1(1 - q_1)(1 - 2q_1) u K^2(u) f(q_1) f(q_1 - uh) \mathrm{d}u \mathrm{d}q_1 \\ &- h^2 \iint q_1(1 - q_1) u^2 K^2(u) f(q_1) f(q_1 - uh) \mathrm{d}u \mathrm{d}q_1. \end{aligned}$$

We may apply the Dominated Convergence Theorem to conclude that

$$\lim_{n \to \infty} 2hE[H_n^2(Z_i, Z_j)] = 2 \iint q_1^2 (1 - q_1)^2 K^2(u) f^2(q_1) du \mathrm{d}q_1.$$

Analysis of $V_{2,n}$:

Let

$$g_n(Z_i, Z_j, \theta) = \frac{\epsilon_i(Q(X_j, \theta) - Q(X_j, \theta_0)) + \epsilon_j(Q(X_i, \theta) - Q(X_i, \theta_0))}{2(n - 1)\sqrt{h}} \times K\left(\frac{Q(X_i, \theta) - Q(X_j, \theta)}{h}\right).$$

Using this notation, we may write

$$n\sqrt{h} V_{2,n} = \sum_{1 \le i,j \le n : i \ne j} \psi_{1,n}(Z_i, Z_j, \hat{\theta}_n) + \sum_{1 \le i \le n} \psi_{2,n}(Z_i, \hat{\theta}_n), \tag{22}$$

where

$$\begin{aligned} \psi_{1,n}(Z_i, Z_j, \theta) &= g_n(Z_i, Z_j, \theta) - E[g_n(Z_i, Z_j, \theta)|Z_i] \\ &\quad - E[g_n(Z_i, Z_j, \theta)|Z_j] \\ \psi_{2,n}(Z_i, \theta) &= E[2(n - 1) g_n(Z_i, Z_j, \theta)|Z_i]. \end{aligned}$$

Note that the sums in (22) are degenerate U-statistics of degree 2 and degree 1, respectively. We will use Lemma 3 of Heckman et al. (1998b) to show that both of these sums tend to zero in probability. To this end, let $0 < \delta_n \to \infty$, but so slowly that $\sqrt{h}\delta_n^2 \to 0$. Define

$$\begin{aligned} \Psi_{1,n} &= \{\psi_{1,n}(Z_i, Z_j, \theta) : \theta \in \Theta_n\} \\ \Psi_{2,n} &= \{\psi_{2,n}(Z_i, \theta) : \theta \in \Theta_n\}, \end{aligned}$$

where

$$\Theta_n = \{\theta \in \Theta : \sqrt{n}\|\theta - \theta_0\| \le \delta_n\}.$$

In order to apply the lemma to the first sum, first note that for $\theta \in \Theta_n$ and $\theta' \in \Theta_n$

$$|\psi_{1,n}(Z_i,Z_j,\theta)-\psi_{1,n}(Z_i,Z_j,\theta')| \ll \frac{\|\theta-\theta'\|}{(n-1)\sqrt{h}}(G(X_i)+G(X_j)+1)$$

$$\leq \frac{\delta_n}{(n-1)\sqrt{nh}}(G(X_i)+G(X_j)+1)=F_{1,n}(Z_i,Z_j), \tag{23}$$

where the first inequality follows from the boundedness of ϵ_i, the boundedness of $K(\cdot)$, and the Lipschitz continuity of $Q(x,\theta)$ w.r.t. θ and the final equality is a definition. Hence,

$$\sum_{1\leq i,j\leq n:i\neq j} E[F_{1,n}(Z_i,Z_j)^2] \ll \frac{\delta_n^2}{(n-1)h}=o(1),$$

where the inequality follows from the assumption on the moments of $G(X_i)$. Next, note that the required condition on the covering numbers of $\Psi_{1,n}$ follows from (23) and Theorem 2.7.11 of van der Vaart and Wellner (1996). Finally, note that

$$E[(\psi_{1,n}(Z_i,Z_j,\theta)-\psi_{1,n}(Z_i,Z_j,\theta'))^2] \ll \frac{\delta_n^2}{(n-1)^2nh}=o(1).$$

Since $\hat{\theta}_n \in \Theta_n$ with probability tending to one, it follows from the lemma that the first sum in (22) tends to zero in probability.

In order to apply the lemma to the second sum, let $p=4$ and $q=4/3$ and note that for $\theta\in\Theta_n$ and $\theta'\in\Theta_n$

$$|\psi_{2,n}(Z_i,\theta)-\psi_{2,n}(Z_i,\theta')| \ll \frac{\|\theta-\theta'\|}{\sqrt{h}}$$
$$\times E\left[G(X_j)\left|K\left(\frac{Q(X_i,\theta)-Q(X_j,\theta)}{h}\right)\right|Z_i\right]$$
$$\leq \frac{\|\theta-\theta'\|}{\sqrt{h}}E[G(X_j)^p|Z_i]^{1/p}E\left[\left|K\left(\frac{Q(X_i,\theta)-Q(X_j,\theta)}{h}\right)\right|^q\middle|Z_i\right]^{1/q}$$
$$\ll \|\theta-\theta'\|h^{1/4}$$
$$\leq \frac{\delta_n h^{1/4}}{\sqrt{n}}=F_{2,n}(Z_i), \tag{24}$$

where the first inequality follows from the boundedness of ϵ_i and the Lipschitz continuity of $Q(x,\theta)$ w.r.t. θ, the second inequality follows from Holder's inequality, the third inequality follows from the assumptions on the moments of $G(X_i)$ and earlier arguments, and the equality is a definition. Hence,

$$\sum_{1\leq i\leq n} E[F_{2,n}(Z_i)^2]=\sqrt{h}\delta_n^2=o(1).$$

Next, note that the required condition on the covering numbers of $\Psi_{2,n}$ follows from (24) and Theorem 2.7.11 of van der Vaart and Wellner (1996). Finally, note that

$$E[(\psi_{2,n}(Z_i,\theta)-\psi_{2,n}(Z_i,\theta'))^2] \ll \frac{\sqrt{h}\delta_n^2}{n}=o(1).$$

Since $\hat{\theta}_n\in\Theta_n$ with probability tending to one, it follows from the lemma that the second sum in (22) tends to zero in probability.

Analysis of $V_{3,n}$:

Note that

$$n\sqrt{h}|V_{3,n}| \ll n\|\hat{\theta}_n-\theta_0\|^2 S_n,$$

where

$$S_n=\sum_{1\leq i,j\leq n:i\neq j}\frac{G(X_i)G(X_j)}{n(n-1)\sqrt{h}}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|.$$

Let $p=(4+\delta)/2$ and $1/q=1-1/p$ and note that

$$E[S_n]=E\left[\frac{G(X_i)G(X_j)}{\sqrt{h}}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|\right]$$
$$\leq \frac{1}{\sqrt{h}}E[G^p(X_i)G^p(X_j)]^{1/p}E\left[\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|^q\right]^{1/q}$$
$$\leq \frac{h^{1/q}}{\sqrt{h}}E[G^{2p}(X_i)]^{1/(2p)}E[G^{2p}(X_j)]^{1/(2p)}$$
$$\times E\left[\frac{1}{h}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|^q\right]^{1/q}$$
$$\ll \frac{h^{1/q}}{\sqrt{h}}$$

where the first inequality follows from Holder's inequality, the second inequality follows from the Cauchy–Schwartz inequality, and the third inequality follows from the assumptions on the moments of $G(X_i)$ and earlier arguments. Thus, by Markov's inequality, $S_n=O_P(h^{1/q}/\sqrt{h})=o_P(1)$. Hence,

$$n\sqrt{h}|V_{3,n}|=o_P(1).$$

Analysis of $V_{4,n}$:

Write

$$n\sqrt{h}V_{4,n}=\sum_{1\leq i,j\leq n:i\neq j}\psi_n(Z_i,Z_j,\hat{\theta}_n),$$

where

$$\psi_n(Z_i,Z_j,\theta)=\frac{\epsilon_i\epsilon_j}{(n-1)\sqrt{h}}\left[K\left(\frac{Q(X_i,\theta)-Q(X_j,\theta)}{h}\right)-K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right].$$

This is a degenerate U-statistic of degree 2. We will use Lemma 3 of Heckman et al. (1998b) to show that this sum tends to zero in probability. To this end, let $0<\delta_n\to\infty$, but so slowly that $\delta_n^2/(nh^3)\to 0$. Define

$$\Psi_n=\{\psi_n(Z_i,Z_j,\theta):\theta\in\Theta_n\},$$

where

$$\Theta_n=\{\theta\in\Theta:\sqrt{n}\|\theta-\theta_0\|\leq\delta_n\}.$$

In order to apply the lemma, first note that for $\theta\in\Theta_n$ and $\theta'\in\Theta_n$

$$|\psi_n(Z_i,Z_j,\theta)-\psi_n(Z_i,Z_j,\theta')| \ll \frac{\|\theta-\theta'\|}{(n-1)h^{3/2}}(G(X_i)+G(X_j))$$

$$\leq \frac{\delta_n}{(n-1)\sqrt{nh^3}}(G(X_i)+G(X_j))=F_n(Z_i,Z_j), \tag{25}$$

where the first inequality follows from the boundedness of ϵ_i, the Lipschitz continuity of $K(\cdot)$ and the Lipschitz continuity of $Q(x,\theta)$ w.r.t. θ and the final equality is a definition. Hence,

$$\sum_{1\leq i,j\leq n:i\neq j}E[F_n(Z_i,Z_j)^2] \ll \frac{\delta_n^2}{(n-1)h^3}=o(1),$$

where the first inequality follows from the assumption on the moments of $G(X_i)$. Next, note that the required condition on the covering numbers of Ψ_n follows from (25) and Theorem 2.7.11 of van der Vaart and Wellner (1996). Finally, note that

$$E[(\psi_n(Z_i,Z_j\theta)-\psi_n(Z_i,Z_j,\theta'))^2] \ll \frac{\delta_n^2}{(n-1)^2nh^3}=o(1).$$

Since $\hat{\theta}_n\in\Theta_n$ with probability tending to one, it follows from the lemma that the sum tends to zero in probability.

Analysis of $V_{5,n}$:

Note that

$$\left|K\left(\frac{Q(X_i,\hat{\theta}_n)-Q(X_j,\hat{\theta}_n)}{h}\right)-K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|$$
$$\ll \frac{\|\hat{\theta}_n-\theta_0\|}{h}(G(X_i)+G(X_j)).$$

Hence,

$$|V_{5,n}| \leq \|\hat{\theta}_n-\theta_0\|^3 S_n,$$

where

$$S_n = \sum_{i \neq j} \frac{G(X_i)G(X_j)(G(X_i)+G(X_j))}{n(n-1)h^2}.$$

Note that

$$E[S_n] = E\left[\frac{G(X_i)G(X_j)(G(X_i)+G(X_j))}{h^2}\right] = O\left(\frac{1}{h^2}\right),$$

where the second equality follows from the Cauchy–Schwartz inequality. Hence, by Markov's inequality,

$$n\sqrt{h}|V_{5,n}| = O_P\left(\frac{1}{\sqrt{nh^3}}\right) = o_P(1).$$

Consistency of $\hat{\Sigma}_n$:

We now argue that

$$\hat{\Sigma}_n \xrightarrow{P} \Sigma.$$

To this end, let

$$\Sigma_{n,0} = \sum_{1 \leq i,j \leq n: i \neq j} \frac{2\epsilon_i^2\epsilon_j^2}{n(n-1)h} K^2\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right)$$

and note that

$$|\hat{\Sigma}_n - \Sigma_{n,0}| \leq S_{1,n} + S_{2,n},$$

where

$$S_{1,n} = \sum_{1 \leq i,j \leq n: i \neq j} \left| \frac{2(\hat{\epsilon}_i^2\hat{\epsilon}_j^2 - \epsilon_i^2\epsilon_j^2)}{n(n-1)h} K^2\left(\frac{Q(X_i,\hat{\theta}_n) - Q(X_j,\hat{\theta}_n)}{h}\right)\right|$$

$$S_{2,n} = \sum_{1 \leq i,j \leq n: i \neq j} \left| \frac{2\epsilon_i^2\epsilon_j^2}{n(n-1)h}\left[K^2\left(\frac{Q(X_i,\hat{\theta}_n) - Q(X_j,\hat{\theta}_n)}{h}\right) - K^2\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right)\right]\right|.$$

We will now argue that both of these sums tend to zero in probability.

Consider first $S_{1,n}$. Note that

$$\hat{\epsilon}_i\hat{\epsilon}_j - \epsilon_i\epsilon_j = (Q(X_i,\theta_0) - Q(X_i,\hat{\theta}_n))\epsilon_j + (Q(X_j,\theta_0) - Q(X_j,\hat{\theta}_n))\hat{\epsilon}_i. \tag{26}$$

Hence,

$$S_{1,n} \ll \frac{\|\hat{\theta}_n - \theta_0\|}{h} \sum_{1 \leq i,j \leq n: i \neq j} \frac{G(X_i)+G(X_j)}{n(n-1)}.$$

Note that

$$\frac{\|\hat{\theta}_n - \theta_0\|}{h} = O_P\left(\frac{1}{\sqrt{nh}}\right) = o_P(1).$$

Since

$$E\left[\sum_{1 \leq i,j \leq n: i \neq j} \frac{G(X_i)+G(X_j)}{n(n-1)}\right] < \infty,$$

it follows by Markov's inequality that

$$\sum_{1 \leq i,j \leq n: i \neq j} \frac{G(X_i)+G(X_j)}{n(n-1)} = O_P(1).$$

Thus, $S_{1,n} = o_P(1)$.

Now consider $S_{2,n}$. Note that

$$S_{2,n} \ll \frac{\|\hat{\theta}_n - \theta_0\|}{h^2} \sum_{1 \leq i,j \leq n: i \neq j} \frac{G(X_i)+G(X_j)}{n(n-1)}.$$

Since

$$\frac{\|\hat{\theta}_n - \theta_0\|}{h^2} = O_P\left(\frac{1}{\sqrt{nh^2}}\right) = o_P(1),$$

we have as before that $S_{2,n} = o_P(1)$.

Hence,

$$\hat{\Sigma}_n = \Sigma_{n,0} + o_P(1).$$

We now analyze the asymptotic behavior of $\Sigma_{n,0}$. Note that $\Sigma_{n,0}$ may be written as a U-statistic with kernel

$$H_n(Z_i,Z_j) = \frac{2\epsilon_i^2\epsilon_j^2}{h} K^2\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right).$$

Moreover,

$$E[\|H_n(Z_i,Z_j)\|^2] \ll E\left[\frac{1}{h^2}K^4\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right)\right] = O\left(\frac{1}{h}\right).$$

Hence, by Lemma 3.1 of Powell et al. (1989), we have that

$$\Sigma_{n,0} = E[H_n(Z_i,Z_j)] + \frac{2}{n}\sum_{1 \leq i \leq n}(E[H_n(Z_i,Z_j)|Z_i] - E[H_n(Z_i,Z_j)]) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Since

$$|E[H_n(Z_i,Z_j)|Z_i]| \ll E\left[\frac{1}{h}K^2\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right)\right] < \infty,$$

we have by Chebychev's inequality that

$$\Sigma_{n,0} = E[H_n(Z_i,Z_j)] + o_P(1).$$

We may rewrite $E[H_n(Z_i,Z_j)]$ as

$$\begin{aligned}
&\iint \frac{2q_1(1-q_1)q_2(1-q_2)}{h} K^2\left(\frac{q_1-q_2}{h}\right) f(q_1)f(q_2)\mathrm{d}q_1\mathrm{d}q_2 \\
&= \iint 2q_1(1-q_1)(q_1-uh)(1-q_1+uh)K^2(u) \\
&\times f(q_1)f(q_1-uh)\mathrm{d}q_1\mathrm{d}u \\
&= \iint 2q_1^2(1-q_1)^2K^2(u)f(q_1)f(q_1-uh)\mathrm{d}q_1\mathrm{d}u \\
&\quad + h\iint 2q_1(1-q_1)(1-2q_1)uK^2(u)f(q_1)f(q_1-uh)\mathrm{d}q_1\mathrm{d}u \\
&\quad + h^2\iint 2q_1(1-q_1)u^2K^2(u)f(q_1)f(q_1-uh)\mathrm{d}q_1\mathrm{d}u.
\end{aligned}$$

We may apply the Dominated Convergence Theorem to conclude that

$$\lim_{n\to\infty} E[H_n(Z_i,Z_j)] = 2\iint q_1^2(1-q_1)^2K^2(u)f^2(q_1)\mathrm{d}q_1\mathrm{d}u = \Sigma,$$

as desired.

A.2. *Proof of* Theorem 4.2

Consider first $\hat{V}_n$. Let $Z_i = (\epsilon_i, X_i)$ and

$$\epsilon_i = D_i - Q(X_i,\theta_0).$$

Define

$$V_{n,0} = \sum_{1 \leq i,j \leq n: i \neq j} \frac{\epsilon_i\epsilon_j}{n(n-1)h} K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right).$$

Using arguments similar to those used to establish consistency of $\hat{\Sigma}_n$ in the proof of Theorem 4.1, we can show that

$$\hat{V}_n = V_{n,0} + o_P(1)$$

and, using Lemma 3.1 of Powell et al. (1989), we can show that

$V_{n,0} = E[H_{1,n}(Z_i, Z_j)] + o_P(1)$,

where

$$H_{1,n}(Z_i, Z_j) = \frac{\epsilon_i \epsilon_j}{h} K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right).$$

We may rewrite $E[H_{1,n}(Z_i, Z_j)]$ as

$$\begin{aligned}
&\iint \frac{(r(q_1) - q_1)(r(q_2) - q_2)}{h} K\left(\frac{q_1 - q_2}{h}\right) f(q_1) f(q_2) \mathrm{d}q_1 \mathrm{d}q_2 \\
&= \iint (r(q_1) - q_1)(r(q_1 - uh) - q_1 + uh) K(u) \\
&\quad \times f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u \\
&= \iint (r(q_1) - q_1)(r(q_1 - uh) - q_1) K(u) f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u \\
&\quad + h \iint (r(q_1) - q_1) u K(u) f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u.
\end{aligned}$$

We may apply the Dominated Convergence Theorem to conclude that

$$\lim_{n\to\infty} E[H_{1,n}(Z_i, Z_j)] = \int (h(q_1) - q_1)^2 f^2(q_1) \mathrm{d}q_1 > 0,$$

where the final inequality follows from the assumption that $E[D_i|Q(X_i, \theta_0)] = Q(X_i, \theta_0)$ with probability strictly less than 1.

Now consider $\hat{\Sigma}_n$. Let

$$\begin{aligned}
\Sigma_{n,0} = &\sum_{1\le i,j\le n: i\neq j} \frac{2\epsilon_i^2 \epsilon_j^2}{n(n-1)h} (E[D_i|Q(X_i, \theta_0)] - 2E[D_i|Q(X_i, \theta_0)] \\
&+ Q(X_i, \theta_0)^2).
\end{aligned}$$

Using arguments similar to those used to establish consistency of $\hat{\Sigma}_n$ in the proof of Theorem 4.1, we can show that

$\hat{\Sigma}_n = \Sigma_{n,0} + o_P(1)$

and, using Lemma 3.1 of Powell et al. (1989), that

$\Sigma_{n,0} = E[H_{2,n}(Z_i, Z_j)] + o_P(1)$,

where

$$H_{2,n}(Z_i, Z_j) = \frac{2\epsilon_i^2 \epsilon_j^2}{h} K^2\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right).$$

We may rewrite $E[H_{2,n}(Z_i, Z_j)]$ as

$$\begin{aligned}
&\iint \frac{2(r(q_1) - 2r(q_1)q_1 + q_1^2)(r(q_2) - 2r(q_2)q_2 + q_2^2)}{h} \\
&\quad \times K^2\left(\frac{q_1 - q_2}{h}\right) f(q_1) f(q_2) \mathrm{d}q_1 \mathrm{d}q_2 \\
&= \iint 2(r(q_1) - 2r(q_1)q_1 + q_1^2)(r(q_1 - uh) - 2r(q_1 - uh) \\
&\quad \times (q_1 - uh) + (q_1 - uh)^2) K^2(u) f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u \\
&= \iint 2(r(q_1) - 2r(q_1)q_1 + q_1^2)(r(q_1 - uh) \\
&\quad - 2r(q_1 - uh)q_1 + q_1^2) K^2(u) f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u \\
&\quad - h \iint 4(r(q_1) - 2r(q_1)q_1 + q_1^2)(r(q_1 - uh) - q_1) u K^2(u) \\
&\quad \times f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u - h^2 \iint 2(r(q_1) \\
&\quad - 2r(q_1)q_1 + q_1^2) u^2 K^2(u) f(q_1) f(q_1 - uh) \mathrm{d}q_1 \mathrm{d}u.
\end{aligned}$$

We may apply the Dominated Convergence Theorem to conclude that

$$\lim_{n\to\infty} E[H_{2,n}(Z_i, Z_j)] = 2 \int (h(q_1) - 2h(q_1)q_1 + q_1^2)^2 f^2(q_1) \mathrm{d}q_1,$$

as desired.

A.3. Proof of Theorem 4.3

Let

$\epsilon_i = D_i - Q(X_i, \theta_0)$

and let $f(x) = f(x, \theta_0)$. As in the proof of Theorem 4.1, we may write

$\hat{V}_n = V_{1,n} - 2V_{2,n} + V_{3,n} + V_{4,n} + V_{5,n}$,

where $V_{1,n}, \ldots, V_{5,n}$ are defined by (17)–(21). We now analyze each of these terms separately.

Analysis of $V_{1,n}$:

Let $\alpha_n = 1/(\sqrt{n}h^{1/4})$ and define

$u_i = \epsilon_i - \alpha_n \ell(Q(X_i, \theta_0)) = D_i - E[D_i|Q(X_i, \theta_0)]$.

Using this notation, we may write

$V_{1,n} = V_{1,n}^A + \alpha_n V_{2,n}^B + \alpha_n^2 V_{3,n}^C$,

where

$$\begin{aligned}
V_{1,n}^A &= \sum_{1\le i,j\le n: i\neq j} \frac{u_i u_j}{n(n-1)h} K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right) \\
V_{2,n}^B &= \sum_{1\le i,j\le n: i\neq j} \frac{u_i \ell(Q(X_j, \theta_0)) + u_j \ell(Q(X_i, \theta_0))}{n(n-1)h} \\
&\quad \times K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right) \\
V_{3,n}^C &= \sum_{1\le i,j\le n: i\neq j} \frac{\ell(Q(X_j, \theta_0))\ell(Q(X_i, \theta_0))}{n(n-1)h} \\
&\quad \times K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right).
\end{aligned}$$

It follows from the analysis of $V_{1,n}$ in the proof of Theorem 4.1 that

$n\sqrt{h} V_{1,n}^A \xrightarrow{d} N(0, \Sigma)$,

where Σ is given by (11). We now analyze the remaining two terms.

Consider first $V_{1,n}^B$. We may write $V_{2,n}^B$ as a U-statistic with kernel

$$\begin{aligned}
H_{1,n}(Z_i, Z_j) = &\frac{u_i \ell(Q(X_j, \theta_0)) + u_j \ell(Q(X_i, \theta_0))}{h} \\
&\times K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right),
\end{aligned}$$

where $Z_i = (u_i, X_i)$. Note that

$$\begin{aligned}
E[\|H_{1,n}(Z_i, Z_j)\|^2] &\ll E\left[\frac{1}{h^2}\left|K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right)\right|\right] \\
&= O\left(\frac{1}{h}\right) = o(n),
\end{aligned}$$

where the inequality follows from the boundedness of u_i and $\ell(Q(X_i, \theta_0))$. Hence, by Lemma 3.1 of Powell et al. (1989), we have that

$$\begin{aligned}
\sqrt{n} V_{1,n}^B = &\sqrt{n} E[H_{1,n}(Z_i, Z_j)] + \frac{1}{\sqrt{n}} \sum_{1\le i\le n} (E[H_{1,n}(Z_i, Z_j)|Z_i] \\
&- E[H_{1,n}(Z_i, Z_j)]) + o_P(1).
\end{aligned}$$

Since $E[H_{1,n}(Z_i, Z_j)] = 0$ and

$$|E[H_{1,n}(Z_i, Z_j)|Z_i]| \ll E\left[\frac{1}{h}\left|K\left(\frac{Q(X_i, \theta_0) - Q(X_j, \theta_0)}{h}\right)\right|\right] < \infty,$$

we have by Chebychev's inequality that

$\sqrt{n} V_{1,n}^B = O_P(1)$.

It follows that

$$n\sqrt{h}\alpha_n V_{2,n}^B = h^{1/4}\sqrt{n}V_{1,n}^B = o_P(1).$$

Now consider $V_{1,n}^C$. We may write $n\sqrt{h}\alpha_n^2 V_{1,n}^C = V_{1,n}^C$ as a U-statistic with kernel

$$H_{2,n}(Z_i, Z_j) = \frac{\ell(Q(X_j,\theta_0))\ell(Q(X_i,\theta_0))}{h} K\left(\frac{Q(X_i,\theta_0) - Q(X_j,\theta_0)}{h}\right).$$

Since

$$\begin{aligned} E[\|H_{1,n}(Z_i,Z_j)\|^2] &\ll E\left[\frac{1}{h^2}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|\right] \\ &= O\left(\frac{1}{h}\right) = o(n), \end{aligned}$$

we have by Lemma 3.1 of Powell et al. (1989) that

$$\begin{aligned} V_{1,n}^C &= E[H_{2,n}(Z_i,Z_j)] + \frac{1}{n}\sum_{1\le i\le n}(E[H_{2,n}(Z_i,Z_j)|Z_i] \\ &\quad - E[H_{2,n}(Z_i,Z_j)]) + o_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Since

$$|E[H_{2,n}(Z_i,Z_j)|Z_i]| \ll E\left[\frac{1}{h}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|\right] < \infty,$$

we have by Chebychev's inequality that

$$V_{1,n}^C = E[H_{2,n}(Z_i,Z_j)] + o_P(1).$$

To complete the argument note that

$$\begin{aligned} E[H_{2,n}(Z_i,Z_j)] &= \iint \frac{\ell(q_1)\ell(q_2)}{h} K\left(\frac{q_1-q_2}{h}\right) f(q_1)f(q_2)\mathrm{d}q_1\mathrm{d}q_2 \\ &= \iint \ell(q_1)\ell(q_1-uh)K(u)f(q_1)f(q_1-uh)\mathrm{d}q_1\mathrm{d}u. \end{aligned}$$

We may apply the Dominated Convergence Theorem to conclude that

$$\lim_{n\to\infty} E[H_{2,n}(Z_i,Z_j)] = \int \ell^2(q_1)f^2(q_1)\mathrm{d}q_1.$$

It follows that

$$n\sqrt{h}V_{1,n} \xrightarrow{d} N\left(\int \ell^2(q_1)f^2(q_1)\mathrm{d}q_1,\ \Sigma\right).$$

Analysis of $V_{2,n}$:

We may write

$$V_{2,n} = V_{2,n}^A + \alpha_n V_{2,n}^B,$$

where

$$\begin{aligned} V_{2,n}^A &= \sum_{1\le i,j\le n: i\ne j} \frac{u_i(Q(X_j,\hat\theta_n) - Q(X_j,\theta_0))}{n(n-1)h} \\ &\quad \times K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right) \\ V_{2,n}^B &= \sum_{1\le i,j\le n: i\ne j} \frac{\ell(Q(X_i,\theta_0))(Q(X_j,\hat\theta_n) - Q(X_j,\theta_0))}{n(n-1)h} \\ &\quad \times K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right). \end{aligned}$$

It follows from the analysis of $V_{2,n}$ in the proof of Theorem 4.1 that

$$n\sqrt{h}V_{2,n}^A = o_P(1).$$

We now analyze the remaining term.

Note that

$$|n\sqrt{h}\alpha_n V_{2,n}^B| \le \sqrt{n}\|\hat\theta_n - \theta_0\| h^{1/4} S_n,$$

where

$$S_n = \sum_{1\le i,j\le n: i\ne j} \frac{G(X_j)}{n(n-1)h}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|.$$

Furthermore,

$$\begin{aligned} E[S_n] &= E\left[\frac{G(X_j)}{h}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|\right] \\ &= E\left[G(X_j)E\left[\frac{1}{h}\left|K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right|\,\middle|\, Z_j\right]\right] \\ &\ll E[G(X_j)] < \infty. \end{aligned}$$

Thus, by Markov's inequality, $S_n = O_P(1)$, which implies that

$$n\sqrt{h}\alpha_n V_{2,n}^B = o_P(1).$$

Analysis of $V_{3,n}$:

Since $V_{3,n}$ does not depend on ϵ_i, the proof of Theorem 4.1 shows that

$$n\sqrt{h}V_{3,n} = o_P(1).$$

Analysis of $V_{4,n}$:

We may write

$$V_{4,n} = V_{4,n}^A + \alpha_n V_{4,n}^B + \alpha_n^2 V_{4,n}^C,$$

where

$$\begin{aligned} V_{4,n}^A &= \sum_{1\le i,j\le n: i\ne j} \frac{u_i u_j}{n(n-1)h}\left[K\left(\frac{Q(X_i,\hat\theta_n)-Q(X_j,\hat\theta_n)}{h}\right)\right. \\ &\quad \left. - K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right] \\ V_{4,n}^B &= \sum_{1\le i,j\le n: i\ne j} \frac{u_i\ell(Q(X_j,\theta_0)) + u_j\ell(Q(X_i,\theta_0))}{n(n-1)h} \\ &\quad \times\left[K\left(\frac{Q(X_i,\hat\theta_n)-Q(X_j,\hat\theta_n)}{h}\right)\right. \\ &\quad \left. - K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right] \\ V_{4,n}^C &= \sum_{1\le i,j\le n: i\ne j} \frac{\ell(Q(X_j,\theta_0))\ell(Q(X_i,\theta_0))}{n(n-1)h} \\ &\quad \times\left[K\left(\frac{Q(X_i,\hat\theta_n)-Q(X_j,\hat\theta_n)}{h}\right)\right. \\ &\quad \left. - K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right]. \end{aligned}$$

It follows from the analysis of $V_{4,n}$ in the proof of Theorem 4.1 that

$$n\sqrt{h}V_{4,n}^A = o_P(1).$$

We now analyze the remaining two terms.

Consider first $V_{4,n}^B$. Let

$$\begin{aligned} g_n(Z_i,Z_j,\theta) &= \alpha_n \frac{u_i\ell(Q(X_j,\theta_0)) + u_j\ell(Q(X_i,\theta_0))}{(n-1)\sqrt{h}} \\ &\quad \times\left[K\left(\frac{Q(X_i,\hat\theta_n)-Q(X_j,\hat\theta_n)}{h}\right)\right. \\ &\quad \left. - K\left(\frac{Q(X_i,\theta_0)-Q(X_j,\theta_0)}{h}\right)\right]. \end{aligned}$$

Using this notation, we may write

$$n\sqrt{h}\alpha_n V_{4,n}^B = \sum_{1\le i,j\le n: i\neq j} \psi_{1,n}(Z_i, Z_j, \hat{\theta}_n) + \sum_{1\le i\le n} \psi_{2,n}(Z_i, \hat{\theta}_n)\,, \quad (27)$$

where

$$\psi_{1,n}(Z_i, Z_j, \theta) = g_n(Z_i, Z_j, \theta) - E[g_n(Z_i, Z_j, \theta)|Z_i] - E[g_n(Z_i, Z_j, \theta)|Z_j]$$

$$\psi_{2,n}(Z_i, \theta) = E[2(n-1)g_n(Z_i, Z_j, \theta)|Z_i].$$

Note that the sums in (27) are degenerate U-statistics of degree 2 and degree 1, respectively. We will use Lemma 3 of Heckman et al. (1998b) to show that both of these sums tend to zero in probability. To this end, let $0 < \delta_n \to \infty$, but so slowly that $\sqrt{h}\delta_n^2 \to 0$. Define

$$\Psi_{1,n} = \{\psi_{1,n}(Z_i, Z_j, \theta) : \theta \in \Theta_n\}$$

$$\Psi_{2,n} = \{\psi_{2,n}(Z_i, \theta) : \theta \in \Theta_n\},$$

where

$$\Theta_n = \{\theta \in \Theta : \sqrt{n}\|\theta - \theta_0\| \le \delta_n\}.$$

In order to apply the lemma to the first sum, first note that for $\theta \in \Theta_n$ and $\theta' \in \Theta_n$

$$|\psi_{1,n}(Z_i, Z_j, \theta) - \psi_{1,n}(Z_i, Z_j, \theta')| \ll \frac{\alpha_n\|\theta - \theta'\|}{(n-1)h^{3/2}}(G(X_i) + G(X_j))$$

$$\le \frac{\alpha_n\delta_n}{\sqrt{n}(n-1)h^{3/2}}(G(X_i) + G(X_j)) = F_{1,n}(Z_i, Z_j), \quad (28)$$

where the first inequality follows from the boundedness of u_i, the Lipschitz continuity of $K(\cdot)$, and the Lipschitz continuity of $Q(x, \theta)$ w.r.t. θ and the final equality is a definition. Hence,

$$\sum_{1\le i,j\le n: i\neq j} E[F_{1,n}(Z_i, Z_j)^2] \ll \frac{\delta_n^2}{n(n-1)h^{7/2}} = o(1),$$

where the inequality follows from the assumption on the moments of $G(X_i)$. Next, note that the required condition on the covering numbers of $\Psi_{1,n}$ follows from (28) and Theorem 2.7.11 of van der Vaart and Wellner (1996). Finally, note that

$$E[(\psi_{1,n}(Z_i, Z_j, \theta) - \psi_{1,n}(Z_i, Z_j, \theta'))^2] \ll \frac{\delta_n^2}{n^2(n-1)^2h^{7/2}} = o(1).$$

Since $\hat{\theta}_n \in \Theta_n$ with probability tending to one, it follows from the lemma that the first sum in (27) tends to zero in probability.

In order to apply the lemma to the second sum, note that for $\theta \in \Theta_n$ and $\theta' \in \Theta_n$

$$|\psi_{2,n}(Z_i, \theta) - \psi_{2,n}(Z_i, \theta')| \ll \frac{\alpha_n\|\theta - \theta'\|}{h^{3/2}}(G(X_i) + 1)$$

$$\le \frac{\alpha_n\delta_n}{\sqrt{n}h^{3/2}} = F_{2,n}(Z_i), \quad (29)$$

where the first inequality follows from the boundedness of u_i, the Lipschitz continuity of $K(\cdot)$, and the Lipschitz continuity of $Q(x, \theta)$ w.r.t. θ, and the equality is a definition. Hence,

$$\sum_{1\le i\le n} E[F_{2,n}(Z_i)^2] = \frac{\delta_n^2}{nh^2} = o(1).$$

Next, note that the required condition on the covering numbers of $\Psi_{2,n}$ follows from (29) and Theorem 2.7.11 of van der Vaart and Wellner (1996). Finally, note that

$$E[(\psi_{2,n}(Z_i, \theta) - \psi_{2,n}(Z_i, \theta'))^2] \ll \frac{\delta_n^2}{n^2h^2} = o(1).$$

Since $\hat{\theta}_n \in \Theta_n$ with probability tending to one, it follows from the lemma that the second sum in (27) tends to zero in probability.

Now consider $V_{4,n}^C$. Note that

$$n\sqrt{h}\alpha_n^2 V_{4,n}^C = V_{4,n}^C \ll \frac{\|\hat{\theta}_n - \theta_0\|}{h^2} S_n,$$

where

$$S_n = \sum_{1\le i,j\le n: i\neq j} \frac{G(X_i) + G(X_j)}{n(n-1)}.$$

Since

$$E[S_n] < \infty,$$

we have by Markov's inequality that

$$S_n = O_P(1)\,.$$

Hence,

$$n\sqrt{h}\alpha_n^2 V_{4,n}^C = O_P\left(\frac{1}{\sqrt{n}h^2}\right) = o_P(1).$$

Analysis of $V_{5,n}$:

Since $V_{5,n}$ does not depend on ϵ_i, the proof of Theorem 4.1 shows that

$$n\sqrt{h}V_{5,n} = o_P(1),$$

which completes the proof.

References

Abadie, A., Imbens, G., 2007. Bias corrected matching estimators for average treatment effects. Working paper, University of California-Berkeley and Harvard University.

Dehejia, R.H., Wahba, S., 2002. Propensity score matching methods for nonexperimental causal studies. The Review of Economics and Statistics 84, 151–161.

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

Hart, J.D., 2007. Nonparametric Smoothing and Lack-of-Fit Tests. In: Springer Series in Statistics.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998a. Characterizing selection bias using experimental data. Econometrica 66 (5), 1017–1098.

Heckman, J., Ichimura, H., Todd, P., 1998b. Matching as an econometric evaluation estimator. Review of Economic Studies 65 (2), 261–294.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Orley, A., Card, D. (Eds.), Handbook of Labor Economics. Elsevier Science, North Holland, Amsterdam, New York, Oxford.

Hong, Y., White, H., 1995. Consistent specification testing via nonparametric series regression. Econometrica 63 (5), 1133–1159.

Horowitz, J.L., Spokoiny, V.G., 2001. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica 69 (3), 599–631.

Lechner, M., 1999. Earnings and employment effects of continuous off-the-job training in East Germany after unification. Journal of Business and Economic Statistics 17, 74–90.

Lechner, M., 2000. An evaluation of public sector sponsored continuous vocational training programs in East Germany. Journal of Human Resources 35, 347–375.

Lechner, M., 2002. Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. The Review of Economics and Statistics 84, 205–220.

Powell, J.L., Stock, J.H., Sotker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 39, 33–38.

Rosenbaum, R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Sianesi, B., 2004. An evaluation of the Swedish system of active labor market programs in the 1990s. The Review of Economics and Statistics 86, 133–155.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC.

Simonsen, M., Skipper, L., 2006. The costs of motherhood: An analysis using matching estimators. Journal of Applied Econometrics 21, 919–934.

Smith, J.A., Todd, P.E., 2005. Rejoinder. Journal of Econometrics 125, 365–375.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York.

Zheng, J.X., 1996. A consistent test of functional form via nonparametric estimation techniques. Journal of Econometrics 75 (2), 263–289.