IZA DP No. 2091

Nonparametric Tests for Treatment Effect Heterogeneity

Richard K. Crump
V. Joseph Hotz
Guido W. Imbens
Oscar A. Mitnik

April 2006

Nonparametric Tests
for Treatment Effect Heterogeneity

Richard K. Crump
University of California at Berkeley

V. Joseph Hotz
University of California at Los Angeles

Guido W. Imbens
University of California at Berkeley

Oscar A. Mitnik
University of Miami
and IZA Bonn

ABSTRACT

Nonparametric Tests for Treatment Effect Heterogeneity[*]

A large part of the recent literature on program evaluation has focused on estimation of the average effect of the treatment under assumptions of unconfoundedness or ignorability following the seminal work by Rubin (1974) and Rosenbaum and Rubin (1983). In many cases however, researchers are interested in the effects of programs beyond estimates of the overall average or the average for the subpopulation of treated individuals. It may be of substantive interest to investigate whether there is any subpopulation for which a program or treatment has a nonzero average effect, or whether there is heterogeneity in the effect of the treatment. The hypothesis that the average effect of the treatment is zero for all subpopulations is also important for researchers interested in assessing assumptions concerning the selection mechanism. In this paper we develop two nonparametric tests. The first test is for the null hypothesis that the treatment has a zero average effect for any subpopulation defined by covariates. The second test is for the null hypothesis that the average effect conditional on the covariates is identical for all subpopulations, in other words, that there is no heterogeneity in average treatment effects by covariates. Sacrificing some generality by focusing on these two specific null hypotheses we derive tests that are straightforward to implement.

Corresponding author:

Guido W. Imbens
University of California at Berkeley
Department of Economics
661 Evans Hall
Berkeley, CA 94720-3880
USA
Email: imbens@econ.berkeley.edu

---

1 Introduction

A large part of the recent literature on program evaluation focuses on estimation of the average effect of the treatment under assumptions of unconfoundedness or ignorability following the seminal work by Rubin (1974) and Rosenbaum and Rubin (1983).[1] This literature has typically allowed for general heterogeneity in the effect of the treatment. The literature on testing for the presence of treatment effects in this context is much smaller. An exception is the paper by Abadie (2002) in the context of instrumental variables models.[2] In many cases however, researchers are interested in the effects of programs beyond point estimates of the overall average or the average for the subpopulation of treated individuals. For example, it may be of substantive interest to investigate whether there is any subpopulation for which a program or treatment has a nonzero average effect, or whether there is heterogeneity in the effect of the treatment. Such questions are particularly relevant for policy makers interested in extending the program or treatment to other populations. Some of this interest in treatment effect heterogeneity has motivated the development of estimators for quantile treatment effects in various settings.[3]

The hypothesis that the average effect of the treatment is zero for all subpopulations is also important for researchers interested in assessing assumptions concerning selection mechanisms. In their discussion of specification tests as a tool to obtain better estimators for average treatment effects, Heckman and Hotz (1989) introduced an important class of specification tests. These tests can be interpreted as tests of the null hypothesis of zero causal effects on lagged outcomes. Heckman and Hotz focused on methods that specifically test the hypothesis of a zero effect under the maintained assumption that the effect is constant. However, the motivation for these tests suggests that the fundamental null hypotheses of interest are ones of zero average effects for all subpopulations. Similarly, Rosenbaum (1997) discusses the use of multiple control groups to investigate the plausibility of unconfoundedness. He shows that if both control groups satisfy an unconfoundedness or exogeneity assumption, differences in average outcomes between the control groups, adjusted for differences in covariates, should be zero in expectation. Again the hypothesis of interest can be formulated as one of zero causal effects for all subpopulations, not just a zero average effect.

In this paper we develop two nonparametric tests. The first test is for the null hypothesis that the treatment has a zero average effect for any subpopulation defined by covariates. The second test is for the null hypothesis that the average effect conditional on the covariates is identical for all subpopulations, in other words, that there is no heterogeneity in average treatment effects by covariates. Sacrificing some generality by focusing on these two specific null hypotheses, we derive tests that are straightforward to implement. They are based on a series or sieve approach to nonparametric estimation for average treatment effects (e.g., Hahn,

---

[1]See Angrist and Krueger (2000), Heckman and Robb (1984), Heckman, Lalonde and Smith (2000), Rosenbaum (2001), Wooldridge (2002), Imbens (2004), Lechner (2002) and Lee (2005) for surveys of this literature.

[2]There is also a large literature on testing in the context of randomized experiments using the randomization distribution. See Rosenbaum (2001).

[3]See, for example, Lehmann (1974) Doksum (1974), Firpo (2004), Abadie, Angrist and Imbens (2002), Chernozhukov and Hansen (2005), Bitler, Gelbach and Hoynes (2002).

1998; Imbens, Newey and Ridder, 2006; Chen, Hong, and Tarozzi, 2004; Chen 2005). Given the particular choice of the sieve, the null hypotheses of interest can be formulated as equality restrictions on subsets of the (expanding set of) parameters. The tests can then be implemented using standard parametric methods. In particular, the test statistics are quadratic forms in the differences in the parameter estimates with critical values from a chi-squared distribution. We provide conditions on the sieves that guarantee that in large samples the tests are valid without the parametric assumptions.

There is a large literature on the related problem of testing parametric restrictions on regression functions against nonparametric alternatives. Eubank and Spiegelman (1990), Härdle and Mammen (1993), Bierens (1982, 1990), Hong and White (1995), and Horowitz and Spokoiny (2001), among others, focus on tests of parametric models for regression functions against nonparametric alternatives. However, the focus in this paper is on two specific tests, zero and constant conditional average treatment effects, rather than on general parametric restrictions. As a result, the proposed tests are particularly easy to implement compared to the Härdle-Mammen and Horowitz-Spokoiny tests. For example, p-values for our proposed tests can be obtained from chi-squared or normal tables, whereas Härdle and Mammen (1993) require the use of a variation of the bootstrap they call the wild bootstrap, and Horowitz and Spokoiny (2001) require simulation to calculate the p-value. Our proposed tests are closer in spirit to those suggested by Eubank and Spiegelman (1990) and Hong and White (1995), who also use series estimation for the unknown regression function, and who obtain a test statistic with a standard normal distribution. In particular, Eubank and Spiegelman (1990) also base their test statistic on the estimated coefficients in the series regression. The general approach behind our testing procedure is also related to the strategy of testing conditional moment restrictions by using an expanding set of marginal moment conditions. See, for example, Bierens (1990), De Jong and Bierens (1994). In those papers, as in the Eubank and Spiegelman (1990) paper, the testing procedures are standard given the number of moment conditions or terms in the series, but remain valid as the moment conditions or number of terms in the series increase with the sample size. In contrast, the validity of our tests require that the number of terms of the series increases with the sample size.

The closest papers in terms of focus to the current paper are those by Härdle and Marron (1990), Neumeyer and Dette (2003) and Pinkse and Robinson (1995). Härdle and Marron study tests of parametric restrictions on comparisons of two regression functions. Their formal analysis is restricted to the case with a single regressor, although it is likely that their kernel methods can be adapted (in particular by using higher order kernels) to extend to the case with multivariate covariates. Their proposed testing procedure leads to a test statistic with a bias term involving the form of the kernel. In contrast, the tests proposed here have a standard asymptotic distribution. Neumeyer and Dette (2003) use empirical process methods to test equality of two regression functions, again in the context of a single regressor. Pinske and Robinson focus on efficient estimation of the nonparametric functions and investigate the efficiency gains from pooling the two data sets in settings where the two regression functions differ by a transformation indexed by a finite number of parameters.

We apply these tests to two sets of experimental evaluations of the effects of welfare-to-work

[2]

programs. In both cases the new tests lead to substantively different conclusions regarding the effect of the programs than has been found in previous analyses of these data that focused solely on average treatment effects. We first analyze data from the MDRC experimental evaluation of California's Greater Avenues for INdependence (GAIN) program that was conducted during the 1990s. These welfare-to-work programs were designed to assist welfare recipients in finding employment and improving their labor market earnings. The programs were implemented at the county level and counties had a great deal of discretion in the designs of their programs. We analyze data for four of these counties. We find that the tests we develop in this paper suggest a very different picture of the efficacy of the programs in these counties compared to conclusions drawn from standard tests of zero average treatment effects. In particular, tests that the average effect of the program on labor market earnings is equal to zero are rejected in only one of the four counties. However, using the tests developed in this paper, we find that for three out of the four counties we can decisively reject the hypothesis of a zero average effect on earnings for all subpopulations of program participants, where subpopulations are defined by covariates. We also reject the hypothesis of a constant average treatment effect across these subpopulations. Taken together, the results using these new tests strongly suggest that, in general, these programs were effective in changing the earnings of participants in these programs, even though it may have not improved or even lowered the earnings of some in the programs. Second, we analyze data from the MDRC experimental evaluations of Work INcentive (WIN) programs in Arkansas, Baltimore, Virginia and San Diego. Again, we find that we cannot reject the null hypothesis of a zero average effect for two out of the four locations. At the same time, we can clearly reject the null hypothesis of a zero average effect for all values of the covariates.

The remainder of the paper is organized as follows. In Section 2, we lay out the framework for analyzing treatment effects and characterize the alternative sets of hypotheses we consider in this paper. We also provide a detailed motivation for conducting tests of average treatment effects being zero and for constant treatment effects. In Section 3, we characterize the latter tests in parametric and nonparametric regression settings. We then lay out the conditions required for the validity of both the zero conditional and the constant treatment effect tests in the nonparametric setting. In Section 4, we apply these tests to the GAIN and WIN data and report on our findings, contrasting the results of our nonparametric tests of zero and constant conditional average treatment effects for these programs on labor market earnings. Finally, we offer some concluding remarks.

2 Framework and Motivation

2.1 Set Up

Our basic framework uses the motivating example of testing zero conditional average treatment effects in a program evaluation setting. We note, however, that our tests can be used more generally to test the hypotheses of constant or zero differences between regression functions estimated on separate samples. The set up we use is standard in the program evaluation literature and based on the potential outcome notation popularized by Rubin (1974). See

Angrist and Krueger (2000), Heckman, Lalonde and Smith (2000), Blundell and Costa-Dias (2002), and Imbens (2004) for general surveys of this literature. We have a random sample of size N from a large population. For each unit i in the sample, let W_i indicate whether the active treatment was received, with $W_i = 1$ if unit i receives the active treatment, and $W_i = 0$ if unit i receives the control treatment. Let $Y_i(0)$ denote the outcome for unit i under control and $Y_i(1)$ the outcome under treatment. We observe W_i and Y_i, where Y_i is the realized outcome:

$$Y_i = Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0).$$

In addition, we observe a vector of pre-treatment variables, or covariates, denoted by X_i. Define the two conditional means, $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$, the two conditional variances, $\sigma_w^2(x) = \mathrm{Var}(Y(w)|X = x)$, the conditional average treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x)$, the propensity score, the conditional probability of receiving the treatment $e(x) = \Pr(W = 1|X = x) = \mathbb{E}[W|X = x]$, and the marginal treatment probability $c = \Pr(W = 1) = \mathbb{E}[e(X)]$.

Assumption 2.1 (INDEPENDENT RANDOM SAMPLE)
(Y_i, W_i, X_i), $i = 1, 2, \ldots, N$ *is an independent random sample.*

To solve the identification problem, we maintain throughout the paper the unconfoundedness assumption (Rosenbaum and Rubin, 1983), which asserts that conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes. Formally:

Assumption 2.2 (UNCONFOUNDEDNESS)

$$W \perp (Y(0), Y(1)) \,\Big|\, X. \tag{2.1}$$

In addition we assume there is overlap in the covariate distributions:

Assumption 2.3 (OVERLAP)
For some $\eta > 0$,

$$\eta \leq e(x) \leq 1 - \eta.$$

Later we also impose smoothness conditions on the two regression functions $\mu_w(x)$ and the conditional variances $\sigma_w^2(x)$.

Various estimators have been proposed for the average treatment effect in this setting, e.g., Hahn (1998), Heckman, Ichimura and Todd (1998), Hirano, Imbens and Ridder (2003), Chen, Hong, and Tarozzi (2004), and Abadie and Imbens (2006).

2.2 Hypotheses

In this paper we focus on two null hypotheses concerning the conditional average treatment effect $\tau(x)$. The first pair of hypotheses we consider

$$H_0: \ \forall \ x \in \mathbb{X}, \ \tau(x) = 0, \qquad H_a: \ \exists \ x \in \mathbb{X}, \ \text{s.t.} \ \tau(x) \neq 0. \qquad (2.2)$$

Under the null hypothesis the average effect of the treatment is zero for all values of the covariates, whereas under the alternative there are some values of the covariates for which the effect of the treatment differs from zero.

The second pair of hypotheses is

$$H_0': \ \exists \ \tau \ \text{s.t.} \ \forall \ x \in \mathbb{X}, \ \tau(x) = \tau, \qquad H_a': \ \forall \tau, \exists \ x \in \mathbb{X}, \ \text{s.t.} \ \tau(x) \neq \tau. \qquad (2.3)$$

We refer to this pair as the null hypothesis of no treatment effect heterogeneity. Strictly speaking this is not correct, as we only require the average effect of the treatment to be equal to τ for all values of the covariates, allowing for distributional effects that average out to zero.

We want to contrast these hypotheses with the pair of hypotheses corresponding to zero average effect,

$$H_0'': \ \mathbb{E}[\tau(X)] = 0, \qquad H_a'': \ \mathbb{E}[\tau(X)] \neq 0. \qquad (2.4)$$

Tests of the null hypothesis of a zero average effect are more commonly carried out, either explicitly, or implicitly through estimating the average treatment effect and its standard error. It is obviously much less restrictive than the null hypothesis of a zero conditional average effect.

To clarify the relation between these hypotheses and the hypotheses typically considered in the nonparametric testing literature it is useful to write the former in terms of restrictions on the conditional mean of Y given X and W. Because W is binary we can write this conditional expectation as

$$\mathbb{E}[Y|X = x, W = w] = h_0(x) + w \cdot h_1(x),$$

where $h_0(x) = \mu_0(x)$ and $h_1(x) = \mu_1(x) - \mu_0(x)$. The nonparametric testing literature has largely focused on hypotheses that restrict both $h_0(x)$ and $h_1(x)$ to parametric forms (e.g., Eubank and Spiegelman, 1990; Härdle and Marron, 1990; Hong and White, 1995; Horowitz and Spokoiny, 2001). In contrast, the first null hypothesis we are interested in is $h_1(x) = 0$ for all x, with no restriction on $h_0(x)$. The second null hypothesis is in this representation $h_1(x) = \tau$ for some τ and all x, and again no restriction on $h_0(x)$. This illustrates that the hypotheses in (2.2) and (2.3) generalize the setting considered in the nonparametric testing literature to a setting where we allow for nuisance functions in the regression function under the null hypothesis.

2.3 Motivation

The motivation for considering the two pairs of hypotheses beyond the hypothesis of a zero average effect consists of three parts. The first is substantive. In many cases the primary

interest of the researcher may be in establishing whether the average effect of the program differs from zero. However, even if it is zero on average, there may well be subpopulations for which the effect is substantively and statistically significant. As a first step towards establishing such a conclusion, it would be useful to test whether there is any statistical evidence against the hypothesis that the effect of the program is zero on average for all subpopulations (the pair of hypotheses H_0 and H_a). If one finds that there is compelling evidence that the program has nonzero effect for some subpopulations, one may then further investigate which subpopulations these are, and whether the effects for these subpopulations are substantively important. As an alternative strategy one could directly estimate average effects for substantively interesting subpopulations. However, there may be many such subpopulations and it can be difficult to control size when testing many null hypotheses. Our proposed strategy of an initial single test for zero conditional average treatment effects avoids such problems.

Second, irrespective of whether one finds evidence in favor or against a zero average treatment effect, one may be concerned with the question of whether there is heterogeneity in the average effect conditional on the observed covariates. If there is strong evidence in favor of heterogeneous effects, one may be more reluctant to recommend extending the program to populations with different distributions of the covariates.

The third motivation is very different. In much of the economic literature on program evaluation, there is concern about the validity of the unconfoundedness assumption. If individuals choose whether or not to participate in the program based on information that is not all observed by the researcher, it may well be that conditional on observed covariates there is some remaining correlation between potential outcomes and the treatment indicator. Such correlation is ruled out by the unconfoundedness assumption. The unconfoundedness assumption is not directly testable. Nevertheless, there are two specific sets of tests available that are suggestive of the plausibility of this assumption. Both are based on testing the effect of a pseudo treatment which is known to have no effect. The first set of tests was originally suggested by Heckman and Hotz (1989). See also the discussion in Imbens (2004). Let us partition the vector of covariates X into two parts, a scalar V and the remainder Z, so that $X = (V, Z')'$. The idea is to take the data $(\mathbf{V}, \mathbf{W}, \mathbf{Z})$ and analyze them as if V is the outcome, W is the treatment indicator, and as if unconfoundedness holds conditional on Z. Since V is a pretreatment variable or covariate, we are certain that the effect of the treatment on V is zero for all units. If we find statistical evidence in favor of an effect of the treatment on V it must therefore be the case that the assumption of unconfoundedness conditional on Z is incorrect. Of course, this is not direct evidence against unconfoundedness conditional on $X = (V, Z')'$. But, at the very least, it suggests that unconfoundedness is a delicate assumption in this case with the presence of V essential. Moreover, such tests can be particularly effective if the researcher has data on a number of lagged values of the outcome. In that case one can choose V to be the one-period lagged value of the outcome. If conditional on further lags and individual characteristics one finds differences in lagged outcome distributions for those who will be treated in the future and those who will not be, it calls into question whether conditioning on all lagged outcome values will be sufficient to eliminate differences between control and treatment groups. Heckman and Hotz (1989) implement these tests by testing whether the average effect of the treatment

is equal to zero, testing the pair of hypotheses in (2.4). Clearly, in this setting it would be stronger evidence in support of the unconfoundedness assumption to find that the effect of the treatment on the lagged outcome is zero for all values of Z. This corresponds to implementing tests of the pairs of hypotheses (2.2).

A similar set of issues comes up in Rosenbaum's (1997) discussion of the use of multiple controls groups. Rosenbaum considers a setting with two distinct potential control groups. He suggests that if biases one may be concerned with would likely be different for both groups, then evidence that the two control groups lead to similar estimates is suggestive that unconfoundedness may be appropriate. One can implement this idea by comparing the two control groups. Let $W_i = 1$ if unit i is from the treatment group, $W_i = 0$ if unit i is from the first control group and $W_i = -1$ if unit i is from the second control group. Suppose unconfoundedness holds for both control groups. Formally, $(Y_i(0), Y_i(1)) \perp W_i | X_i, W_i \in \{0, 1\}$ (unconfoundedness relative to first control group) and $(Y_i(0), Y_i(1)) \perp W_i | X_i, W_i \in \{-1, 1\}$ (unconfoundedness relative to second control group). Then it is likely that in fact $(Y_i(0), Y_i(1)) \perp W_i | X_i$. This implies that $Y_i(0) \perp W_i | X_i, W_i \in \{-1, 0\}$ and thus $Y_i \perp W_i | X_i, W_i \in \{-1, 0\}$. This last conditional independence relation is directly testable. To carry out the test, one can analyze the subsample with $W_i \in \{-1, 0\}$ as if $D_i = 1\{W_i = 0\}$ is a treatment indicator. If we find evidence that this pseudo treatment has a systematic effect on the outcome, it must be that for at least one of the two control groups unconfoundedness is violated. As in the Heckman-Hotz setting, the pair of hypotheses to test is that of a zero conditional average treatment effect, (2.2).

In the next section we discuss implementing the two tests in a parametric framework. In Section 3.2, we then provide conditions under which these tests can be interpreted as nonparametric tests.

3 Testing

3.1 Tests in Parametric Models

Here we discuss parametric versions of the tests in (2.2) and (2.3). For notational convenience we assume here that $N_0 = N_1 = N$. This can be relaxed easily, as we will do in the nonparametric case. Suppose the regression functions are specified as

$$\mu_w(x) = \alpha_w + \beta_w' h(x),$$

for some vector of functions of the covariates $h(x)$, with dimension $K - 1$. The simplest case is $h(x) = x$ where we just estimate a linear model. We can estimate α_w and β_w using least squares:

$$(\hat{\alpha}_w, \hat{\beta}_w) = \arg\min \sum_{i | W_i = w} \left(Y_i - \alpha_w - \beta_w' h(X_i) \right)^2. \tag{3.5}$$

Under general heteroskedasticity, with $V(Y(w)|X) = \sigma_w^2(X)$, the normalized covariance matrix of $(\hat{\alpha}_w, \hat{\beta}_w')'$ is

$$\Omega_w = N \cdot \left(\sum_{i=1}^N h(X_i) h(X_i)' \right)^{-1} \sum_{i=1}^N \sigma_w^2(X_i) h(X_i) h(X_i)' \left(\sum_{i=1}^N h(X_i) h(X_i)' \right)^{-1}. \tag{3.6}$$

In large samples,

$$\sqrt{N} \begin{pmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\beta}_0 - \beta_0 \\ \hat{\alpha}_1 - \alpha_1 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega_0 & 0 \\ 0 & \Omega_1 \end{pmatrix}\right) \tag{3.7}$$

Let $\hat{\Omega}_0$ and $\hat{\Omega}_1$ be consistent estimators for Ω_0 and Ω_1. In this parametric setting the first pair of null and alternative hypotheses is

$$H_0: \ (\alpha_0, \beta_0') = (\alpha_1, \beta_1'), \qquad \text{and} \ \ H_a: \ (\alpha_0, \beta_0') \neq (\alpha_1, \beta_1').$$

This can be tested using the quadratic form

$$T = N \cdot \begin{pmatrix} \hat{\alpha}_0 - \hat{\alpha}_1 \\ \hat{\beta}_0 - \hat{\beta}_1 \end{pmatrix}' (\hat{\Omega}_0 + \hat{\Omega}_1)^{-1} \begin{pmatrix} \hat{\alpha}_0 - \hat{\alpha}_1 \\ \hat{\beta}_0 - \hat{\beta}_1 \end{pmatrix}. \tag{3.8}$$

Under the null hypothesis this test statistic has in large samples a chi-squared distribution with K degrees of freedom:

$$T \xrightarrow{d} \mathcal{X}^2(K). \tag{3.9}$$

The second test is similar. The original null and alternative hypothesis in (2.3) translate into

$$H_0': \ \beta_0 = \beta_1, \qquad \text{and} \ \ H_a': \ \beta_0 \neq \beta_1.$$

Partition Ω_w into the part corresponding to the variance for $\hat{\alpha}_w$ and the part corresponding to the variance for $\hat{\beta}_w$:

$$\Omega_w = \begin{pmatrix} \Omega_{w,00} & \Omega_{w,01} \\ \Omega_{w,10} & \Omega_{w,11} \end{pmatrix},$$

and partition $\hat{\Omega}_0$ and $\hat{\Omega}_1$ similarly. The test statistic is now

$$T' = N \cdot \left(\hat{\beta}_0 - \hat{\beta}_1 \right)' (\hat{\Omega}_{0,11} + \hat{\Omega}_{1,11})^{-1} \left(\hat{\beta}_0 - \hat{\beta}_1 \right). \tag{3.10}$$

Under the null hypothesis this test statistic has in large samples a chi-squared distribution with $K - 1$ degrees of freedom:

$$T' \xrightarrow{d} \mathcal{X}^2(K - 1). \tag{3.11}$$

Both these tests are standard in this parametric setting. The next section shows how these testing procedures can be used to do nonparametric tests.

3.2 Nonparametric Estimation of Regression Functions

In order to develop nonparametric extensions of the tests developed in Section 3.1, we need nonparametric estimators for the two regression functions. We use the particular series estimator for the regression function $\mu_w(x)$ developed by Imbens, Newey and Ridder (2006) and Chen, Hong and Tarozzi (2004). See Chen (2005) for a general discussion of sieve methods. Let K denote the number of terms in the series. As the basis we use power series. Let $\lambda = (\lambda_1, ..., \lambda_d)$ be a multi-index of dimension d, that is, a d-dimensional vector of non-negative integers, with norm $|\lambda| = \sum_{k=1}^d \lambda_k$, and let $x^\lambda = x_1^{\lambda_1} \ldots x_d^{\lambda_d}$. Consider a series $\{\lambda(r)\}_{r=1}^\infty$ containing all distinct vectors such that $|\lambda(r)|$ is nondecreasing. Let $p_r(x) = x^{\lambda(r)}$, where $P_r(x) = (p_1(x), ..., p_r(x))'$. Given Assumption 3.1 the expectation $\Omega_K = \mathbb{E}[P_K(X)P_K(X)'|W = 1]$ is nonsingular for all K. Hence we can construct a sequence $R_K(x) = \Omega_K^{-1/2} P_K(x)$ with $\mathbb{E}[R_K(X)R_K(X)'|W = 1] = I_K$. Let $R_{kK}(x)$ be the kth element of the vector $R_K(x)$. It will be convenient to work with this sequence of basis function $R_K(x)$. The nonparametric series estimator of the regression function $\mu_w(x)$, given K terms in the series, is given by:

$$\hat{\mu}_{w,K}(x) = R_K(x)' \left(\sum_{i|W_i=w} R_K(X_i)R_K(X_i)' \right)^{-} \sum_{i|W_i=w} R_K(X_i)Y_i = R_K(x)'\hat{\gamma}_{w,K},$$

where A^- denotes a generalized inverse of A, and

$$\hat{\gamma}_{w,K} = \left(\sum_{i|W_i=w} R_K(X_i)R_K(X_i)' \right)^{-} \sum_{i|W_i=w} R_K(X_i)Y_i.$$

Define the $N_w \times K$ matrix $R_{w,K}$ with rows equal to $R_K(X_i)'$ for units with $W_i = w$, and Y_w to be the N_w vector with elements equal to Y_i for the same units, so that $\hat{\gamma}_{w,K} = (R'_{w,K}R_{w,K})^-(R'_{w,K}Y_w)$.

Given the estimator $\hat{\mu}_{w,K}(x)$ we estimate the error variance σ_w^2 as

$$\hat{\sigma}_{w,K}^2 = \frac{1}{N_w} \sum_{i|W_i=w} (Y_i - \hat{\mu}_{w,K}(X_i))^2.$$

Let

$$\Omega_{w,K} \equiv \mathbb{E}\left[R_K(X)R_K(X)'|W = w \right]$$

so that the limiting variance of $\sqrt{N}\hat{\gamma}_{w,K}$ is $\sigma_{w,K}^2 \cdot \Omega_{w,K}^{-1}$ as the sample size increases for fixed K. We estimate this variance as

$$\hat{\sigma}_{w,K}^2 \cdot (R'_{w,K}R_{w,K}/N)^{-1}.$$

In addition to Assumptions 2.2 and 2.3 we make the following assumptions.

Assumption 3.1 (DISTRIBUTION OF COVARIATES)
$X \in \mathbb{X} \subset \mathbb{R}^d$, where $\mathbb{X}$ is the Cartesian product of intervals $[x_{jL}, x_{jU}]$, $j = 1, \ldots, d$, with $x_{jL} < x_{jU}$. The density of X is bounded away from zero on $\mathbb{X}$.

Assumption 3.2 (CONDITIONAL OUTCOME DISTRIBUTIONS)
(i) *The two regression functions $\mu_w(x)$ are s times continuously differentiable, with $\frac{s}{d} > 25/4$.*
(ii) *for $\varepsilon_{w,i} = Y_i(w) - \mu_w(x_i)$*
  (a) $\mathbb{E}\left[\varepsilon_{w,i}|X = x\right] = 0$
  (b) $\mathbb{E}\left[\varepsilon_{w,i}^2|X = x\right] = \sigma_w^2$ *where $\sigma_w^2 \in (0, \infty)$*
  (c) $\mathbb{E}[|\varepsilon_{w,i}|^3] < \infty.$

Assumption 3.3 (RATES FOR SERIES ESTIMATORS)
$K = N^\nu$, *with $d/(2s + 3d) < \nu < 2/19$.*

We assume homoskedasticity, although this assumption is not essential and can be relaxed to allow the conditional variance to depend on x, as long as it is bounded from above and below.

3.3 Nonparametric Tests: Zero Conditional Average Treatment Effect

In this section, we show how the tests discussed in Section 3.1 based on parametric regression functions can be used to test the null hypothesis against the alternative hypothesis given in (2.2) without the parametric model. Essentially, we are going to provide conditions under which we can apply a sequence of parametric tests identical to those discussed in Section 3.1 and obtain a test that is valid without the parametric specification.

First, we focus on tests of the null hypothesis that the conditional average treatment effect $\tau(x)$ is zero for all values of the covariates, (2.2). To test this hypothesis, we compare estimators for $\mu_1(x)$ and $\mu_0(x)$. Given our use of series estimators, we can compare the estimated parameters $\hat{\gamma}_{0,K}$ and $\hat{\gamma}_{1,K}$. Specifically, we use as the test statistic for the test of the null hypothesis H_0 the normalized quadratic form

$$T = \left(N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K\right)/\sqrt{2K} \tag{3.12}$$

where

$$\hat{V} = \frac{\hat{\sigma}_{0,K}^2}{1 - \hat{c}} \cdot \hat{\Omega}_{0,K}^{-1} + \frac{\hat{\sigma}_{1,K}^2}{\hat{c}} \cdot \hat{\Omega}_{1,K}^{-1} \tag{3.13}$$

with

$$\hat{c} = \frac{N_1}{N} \qquad 1 - \hat{c} = \frac{N_0}{N}. \tag{3.14}$$

Theorem 3.1 *Suppose Assumptions 2.1-2.3 and 3.1-3.3 hold. Then if $\tau(x) = 0$ for all $x \in \mathbb{X}$,*

$$T \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof: See Appendix.

To gain some intuition for this result, it is useful to decompose the difference $\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}$ into three parts. Define the pseudo-true values $\gamma_{w,K}^*$, for $w = 0, 1$, $K = 1, 2, \ldots$, as

$$\gamma_{w,K}^* = \arg\min_\gamma \mathbb{E}\left[\left(\mu(X) - R_K(X)'\gamma\right)^2 |W = w\right]$$

$$= \left(\mathbb{E}\left[R_K(X)R_K(X)'|W = w\right]\right)^{-1}\mathbb{E}\left[R_K(X)Y|W = w\right],\tag{3.15}$$

so that for fixed K, as $N \to \infty$, $\hat{\gamma}_{w,K} \to \gamma_{w,K}^*$. Then

$$\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} = \left(\gamma_{1,K}^* - \gamma_{0,K}^*\right) + \left(\hat{\gamma}_{1,K} - \gamma_{1,K}^*\right) - \left(\hat{\gamma}_{0,K} - \gamma_{0,K}^*\right).$$

For fixed K, in large samples, the last two terms are normally distributed, and centered around zero. The asymptotic distribution of T is based on this approximate normality. This approximation ignores the first term, the difference $\left(\gamma_{1,K}^* - \gamma_{0,K}^*\right)$. For fixed K this difference is not equal to zero even if $\mu_0(x) = \mu_1(x)$ because the covariate distributions differ in the two treatment groups. In large samples, however, with large K, we can ignore this difference. Recall that under the null hypothesis $\mu_0(x) = \mu_1(x)$ for all x. For large enough K, it must be that $\mu_w(x)$ is close to $R_K(x)'\gamma_{w,K}^*$ for all x. Hence, it follows that, for large enough K, it must be that for all x, $R_K(x)'(\gamma_{1,K}^* - \gamma_{0,K}^*)$ is close to zero, implying $\gamma_{0,K}^*$ and $\gamma_{1,K}^*$ are close. The formal result then shows that we can increase K fast enough to make this difference small, while at the same time increasing K slowly enough to maintain the close approximation of the distribution of $\hat{\gamma}_{w,K} - \gamma_{w,K}^*$ by a normal one. A key result here is Theorem 1.1 in Bentkus (2005) that ensures that convergence to multivariate normality is fast enough to hold even with the dimension of the vector increasing.

In large samples, the test statistic has a standard normal distribution if the null hypothesis is correct. However, we would only reject the null hypothesis if the two regression functions are far apart, which corresponds to large positive values of the test statistic. Hence, we recommend using critical values for the test based on one-sided tests, like De Jong and Bierens (1994).

In practice, we may wish to modify the testing procedure slightly. Instead of calculating T we can calculate the quadratic form

$$Q = N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'\hat{V}^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) = \sqrt{2K} \cdot T + K,$$

and compare this to the critical values of a chi-squared distribution with K degrees of freedom. In large samples this would lead to approximately the same decision rule since $(Q - K)/\sqrt{2K}$ is approximately standard normal if Q has a chi-squared distribution with degrees of freedom equal to K for large K. This modification would make the testing procedure identical to the one discussed in Section 3.1, which is what one would do if the parametric model

$$\mu_w(x) = R_K(x)'\gamma_{w,K},$$

is correctly specified. This makes the tests particularly simple to apply. However, in large samples the tests do not rely on the correct specification, instead relying on the increasingly flexible specification as K increases with the sample size.

Next, we analyze the properties of the test when the null hypothesis is false. We consider local alternatives. For the test of the null hypothesis of a zero conditional average treatment effect, the alternative is

$$\mu_1(x) - \mu_0(x) = \rho_N \cdot \Delta(x),$$

for some sequence of $\rho_N \to 0$, and any function $\Delta(x)$, such that $|\Delta(x_0)| > 0$ for some x_0.

Theorem 3.2 (CONSISTENCY OF TEST)

Suppose Assumptions 2.1-2.3, 3.1–3.3 hold. Suppose also that under the alternative hypothesis $\mu_1(x) - \mu_0(x) = \rho_N \cdot \Delta(x)$ with $\Delta(x)$ s times continuously differentiable, and $|\Delta(x_0)| = C_0 > 0$ for some x_0, and $\rho_N^{-1} = O(N^{1/2-3\nu/2-\epsilon})$ for some $\epsilon > 0$. Then $\Pr(T \geq M) \to 1$ for all M.

Proof: See Appendix.

The theorem implies that we cannot necessarily detect alternatives to the null hypothesis that are $N^{-1/2}$ from the null hypothesis. We can, however, detect alternatives whose distance to the null hypothesis is arbitrarily close to $N^{-1/2}$ given sufficient smoothness relative to the dimension of the covariates (so that ν can be close to zero).

3.4 Nonparametric Tests: Constant Conditional Average Treatment Effect

Next, we consider tests of the null hypothesis against the alternative hypothesis given in (2.3). Suppose without loss of generality that $R_{1K}(x) = 1$ for all K. For this test we partition $\hat{\gamma}_{w,K}$ as

$$\hat{\gamma}_{w,K} = \begin{pmatrix} \hat{\gamma}_{w0,K} \\ \hat{\gamma}_{w1,K} \end{pmatrix},$$

with $\hat{\gamma}_{w0,K}$ a scalar and $\hat{\gamma}_{w1,K}$ a $K-1$-dimensional vector, and the matrix $\hat{V}$ as

$$\hat{V} = \begin{pmatrix} \hat{V}_{00} & \hat{V}_{01} \\ \hat{V}_{10} & \hat{V}_{11} \end{pmatrix}$$

where $\hat{V}_{00}$ is scalar, $\hat{V}_{01}$ is a $1 \times (K-1)$ vector, $\hat{V}_{10}$ is a $(K-1) \times 1$ vector and $\hat{V}_{11}$ is a $(K-1) \times (K-1)$ matrix. The test statistic is then:

$$T' = \left(N \cdot (\hat{\gamma}_{11,K} - \hat{\gamma}_{01,K})' \hat{V}_{11}^{-1} (\hat{\gamma}_{11,K} - \hat{\gamma}_{01,K}) - (K-1) \right) / \sqrt{2(K-1)}. \tag{3.16}$$

Theorem 3.3 *Suppose Assumptions 2.1-2.3 and 3.1-3.3 hold. Then if $\tau(x) = \tau$ for some τ and for all $x \in \mathbb{X}$,*

$$T' \xrightarrow{d} \mathcal{N}(0,1).$$

Proof: See supplementary materials on website.

In practice we may again wish to use the chi-squared approximation. Now we calculate the quadratic form

$$Q = N \cdot (\hat{\gamma}_{11,K} - \hat{\gamma}_{01,K})' \hat{V}_{11}^{-1} (\hat{\gamma}_{11,K} - \hat{\gamma}_{01,K}) = \sqrt{2(K-1)} \cdot T' + K - 1,$$

and compare this to the critical values of a chi-squared distribution with $K-1$ degrees of freedom.

Again, we analyze the properties of the test when the null hypothesis is false.

Theorem 3.4 (CONSISTENCY OF TEST)

Suppose Assumptions 2.1-2.3, 3.1–3.3 hold. Suppose also that under the alternative hypothesis $\mu_1(x) - \mu_0(x) = \tau + \rho_N \cdot \Delta(x)$ with $\Delta(x)$ s times continuously differentiable, and $|\Delta(x_0)| = C_0 > 0$ for some x_0, and $\rho_N^{-1} = O(N^{1/2-3\nu/2-\epsilon})$ for some $\epsilon > 0$. Then $\Pr(T' \geq M) \to 1$ for all M.

Proof: See supplementary materials on website.

4 Application

In this section we apply the tests developed in this paper to data from two sets of experimental evaluations of welfare-to-work training programs. We first re-analyze data from the MDRC evaluations of California's Greater Avenues to INdependence (GAIN) programs. These experimental evaluations of job training and job search assistance programs took place in the 1990's in several different counties in California.[4] The second set consists of four experimental Work INcentive (WIN) demonstration programs implemented in the mid-eighties in different locations of the U.S. The WIN programs also were welfare-to-work programs that examined different strategies for improving the employment and earnings of welfare recipients.[5] The design of both evaluations entailed random assignment of welfare recipients to a treatment group that received training and job assistance services and a control group that did not. Thus, estimating the average effect from these data is straightforward. While the effects of treatments were analyzed for a number of different outcomes, we focus here on the labor market earnings of participants in the first year after random assignment for both sets of evaluations.

4.1 Treatment Effect Tests for the GAIN Data

In this section, we present the results of tests concerning the effects of the GAIN programs in four of California's counties, namely Los Angeles (LA), Riverside (RI), Alameda (AL) and San Diego (SD) counties, on participants' labor market earnings in the first year after random assignment. The sample sizes for the treatment and control groups in each of these counties are provided at the top of Table 1. For each county, we conducted tests for zero and constant conditional average treatment effects, where we condition on measures of participants' background characteristics – including gender, age, ethnicity (Hispanic, black, or other), an indicator for high school graduation, an indicator for the presence of exactly one child (all individuals have at least one child), and an indicator for the presence of children under the age of 6 – as well as on the quarterly earnings of participants in the ten quarters prior to random assignment. Descriptive statistics (means and standard deviations) for these conditioning covariates, as well as for the earnings outcome variable, are found in Table 1, separately by county. All the conditioning data on earnings are in thousands of dollars per quarter. For all of the tests, we controlled for all seven individual characteristics linearly, plus a quadratic term for age, plus all ten quarterly earnings variables and ten indicators for zero earnings in each quarter. This leads to a total of 28 covariates (listed in Table 1) in the regressions, plus an intercept.

The results for the various tests we consider are reported in Table 2. (The degrees of freedom for the chi-squared version of the tests are recorded in this table under the "dof" heading.) We first consider the test of the null hypothesis that $\tau(x) = 0$ against the alternative that $\tau(x) \neq 0$ for some x ("Zero Cond. Ave. TE"). For this test, we get a clear rejection of the zero conditional average treatment effect at the 5% level for three out of the four of the GAIN

---

[4]For a description of these evaluations and their 3-year findings, see Riccio, Friedlander and Freedman (1994). Also see Hotz, Imbens and Klerman (2006) for a re-analysis of the longer term effects of the GAIN programs.

[5]For a description of these evaluations, see Gueron and Pauly (1991). Also see Hotz, Imbens and Mortimer (2005) for a re-analysis of these data.

counties, with the test statistic for only Los Angeles County being smaller than conventional critical values. (For all of the tests, we also include the normal distribution based version of the tests.) Results for the second test of the null hypothesis that $\tau(x) = \tau$ against the alternative that $\tau(x) \neq \tau$ for some x ("Constant Cond. Ave. TE") also are presented in Table 2. Again, we reject this null hypothesis at conventional levels for three out of the four counties. Finally, for comparison purposes, we include the simple test for the null hypothesis that the average effect of the treatment is equal to zero ("Zero Ave. TE"). This is the traditional test that is typically reported when testing treatment effects in the program evaluation literature. It is based on the statistic calculated as the difference in average outcomes for the treatment and control groups divided by the standard error of this difference. Based on this traditional test, we cannot reject the null hypothesis of no treatment effect in three out of the four counties. In particular, only for the Riverside data is there a clear rejection of a zero average treatment effect on earnings.

This latter finding, namely that only Riverside County's GAIN program showed significant effects on earnings (and other outcomes) in the initial periods after random assignment, is what was reported in the MDRC analysis on this evaluation[6]. It has been widely cited as evidence that the program strategies used in Riverside county GAIN program, namely emphasis on job search assistance and little or no basic skills training used by the other GAIN county programs, was the preferred strategy for moving welfare recipients from welfare to work.[7] However, as the results for the other two tests presented in Table 2 make clear, these conclusions are not robust. The findings from the two tests developed in this paper applied to these data clearly suggest that some subgroups in counties other than Riverside benefited from the GAIN treatments in those counties. Moreover, there is clear evidence of treatment effect heterogeneity across subgroups in all but Los Angeles County.

4.2   Treatment Effect Tests for the WIN Data

In this section, we present results for the same set of tests using data from the Work INcentive (WIN) experiments in Baltimore, Maryland (MD), Arkansas (AK), San Diego County (SD) and Virginia (VA). Here we have data on four binary indicators for individual characteristics, an indicator for one child, an indicator for a high school diploma, for never being married, and for being non white. In addition, we have four quarters of earnings data. Table 3 presents summary statistics for the 12 covariates and the outcome variable, annual earnings in the first year after random assignment, for the four locations.

Results of the tests for the four WIN evaluation locations are presented in Table 4, which has the same format as Table 2. With respect to the test of zero conditional average treatment effects, we find that we can reject this null hypothesis in three out of the four locations of the WIN experiments at the 5% level. For two out of those three locations, we also reject the hypothesis of constant treatment effects. In contrast, testing the null hypothesis of a zero average treatment effect results in the rejection of the null hypothesis for only one out of the four

---

[6]See Riccio, Friedlander and Freedman (1994).

[7]Also see Hotz, Imbens and Klerman (2006) for an explicit analysis of the relative effectiveness of alternative treatment strategies based on this same GAIN data.

locations. Overall, the conclusion is again that a researcher who relied only on the traditional tests of a zero average effect would have missed the presence of treatment effects for two out of the four locations analyzed in this set of evaluations.

5   Conclusion

In this paper, we develop and apply tools for testing the presence of and heterogeneity in treatment effects in settings with selection on observables (unconfoundedness). In these settings, researchers have largely focused on inference for the average effect or the average effect for the treated. Although researchers have typically allowed for general treatment effect heterogeneity, there has been little formal investigation of the presence of such heterogeneity and the presence of more complex patterns of treatment effects that could not be detected with traditional tests concerning average treatment effects. At best, researchers have estimated average effects for subpopulations defined by categorical individual characteristics. Here, we develop simple-to-apply tools for testing both the presence of non-zero treatment effects and of treatment effect heterogeneity. Analyzing data from eight experimental evaluations of welfare-to-work training programs, we find considerable evidence of treatment effect heterogeneity and of non-zero treatment effects that were missed by testing strategies that focused solely on inferences concerning average treatment effects.

We note that there is a related issue with respect to the presence of heterogeneity when estimating average treatment effects. In particular, allowing for general forms of heterogeneity can lead to imprecise estimates of such effects. To address this issue, Crump, Hotz, Imbens and Mitnik (2006) explore the potential gains of focusing on the estimation of average effects for subpopulations which have more overlap in the covariate distributions. They provide a systematic treatment of the choice of these subpopulations and develop estimators of treatment effects that have optimal asymptotic properties with respect to their precision.

6 Appendix

Before proving Theorem 3.1 we present a couple of preliminary results.

Lemma A.1 *Suppose Assumptions 2.1-2.3 and 3.1 hold. Then* (i)

$$\left\|\hat{\Omega}_{w,K} - \Omega_{w,K}\right\| = O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right),$$

and (ii) *The eigenvalues of* $\Omega_{w,K}$ *are bounded and bounded away from zero and* (iii) *The eigenvalues of* $\hat{\Omega}_{w,K}$ *are bounded and bounded away from zero in probability if* $O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right) = o_p(1)$.

Proof: See supplementary materials on website.

Lemma A.2 *Suppose Assumptions 2.1-2.3 and 3.1 hold. Then* (i) *The eigenvalues of* V *are bounded and bounded away from zero and* (ii) *The eigenvalues of* $\hat{V}$ *are bounded and bounded away from zero in probability if* $O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right) = o_p(1)$.

Proof: See supplementary materials on website.
Newey (1994) showed that $\zeta(K)$ is $O(K)$, so Lemma $A.1$ implies that if $K^3/N \to 0$ (as implied by Assumption 3.3), $\|\hat{\Omega}_{w,K} - \Omega_{w,K}\| = o_p(1)$.

Next, recall from equation (3.15) the pseudo true value $\gamma_{w,K}^*$ is

$$\gamma_{w,K}^* \equiv \left(\mathbb{E}[R_K(X)R_K(X)'|W = w]\right)^{-1}\mathbb{E}[R_K(X)Y|W = w] = \Omega_{w,K}^{-1}\mathbb{E}[R_K(X)Y|W = w].$$

and define

$$\tilde{\gamma}_{w,K} \equiv \gamma_{w,K}^* + \Omega_{w,K}^{-1}R'_{w,K}\varepsilon_w/N_w \tag{A.1}$$

where

$$\varepsilon_w \equiv Y_w - \mu_w(\mathbf{X}).$$

Then we can write $\sqrt{N_w}(\tilde{\gamma}_{w,K} - \gamma_{w,K}^*)$ as

$$\Omega_{w,K}^{-1}\frac{1}{\sqrt{N_w}}R'_{w,K}\varepsilon_w = \frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^{N}\Omega_{w,K}^{-1}R_K(X_i)\varepsilon_{w,i} \tag{A.2}$$

with

$$\mathbb{E}[\Omega_{w,K}^{-1}R_K(X_i)\varepsilon_{w,i}] = \Omega_{w,K}^{-1}\mathbb{E}\left[R_K(X_i)\mathbb{E}\left[\varepsilon_{w,i}|X_i\right]\right] = 0$$

and

$$\mathbb{V}\left[\Omega_{w,K}^{-1}R_K(X_i)\varepsilon_{w,i}\right] = \sigma_w^2 \cdot \Omega_{w,K}^{-1}$$

Therefore,

$$S_{w,K} \equiv \frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^{N}\left[\sigma_w^2 \cdot \Omega_{w,K}\right]^{-\frac{1}{2}}R_K(X_i)\varepsilon_{w,i} \equiv \frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^{N}Z_i \tag{A.3}$$

is a normalized summation of N_w independent random vectors distributed with expectation $\mathbf{0}$ and variance-covariance matrix I_K.

[16]

Denote the distribution of $S_{w,K}$ by Q_{N_w} and define $\beta_3 \equiv \sum_{i|W_i=w} \mathbb{E}\left\|\frac{Z_i}{\sqrt{N_w}}\right\|^3$. Then, by Theorem 1.1, Bentkus (2005),

$$sup_{\mathcal{A}\in A_K} |Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A})| \leq C\beta_3 K^{1/4}$$

where A_K is the class of all measurable convex sets in K-dimensional Euclidean space, C is an absolute constant, and Φ is a multivariate standard Gaussian distribution.

Lemma A.3 *Suppose Assumptions 2.1-2.3 and 3.1$-$3.3. In particular let $K(N) = N^\nu$ where $\nu < \frac{2}{19}$. Then,*

$$\sup_{\mathcal{A}\in A_K} |Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A})| \to 0$$

Proof First we will show that β_3 is $O(K^{\frac{9}{2}}N^{-\frac{1}{2}})$

$$
\begin{aligned}
\beta_3 \equiv \sum_{i|W_i=w} \mathbb{E}\left\|\frac{Z_i}{\sqrt{N_w}}\right\|^3 &= N_w^{-\frac{3}{2}} \sum_{i|W_i=w} \mathbb{E}\left\|\left[\sigma_w^2 \cdot \Omega_{w,K}\right]^{-\frac{1}{2}} R_K(X_i)\varepsilon_{w,i}\right\|^3 \\
&= \left(N_w \cdot \sigma_w^2\right)^{-\frac{3}{2}} \sum_{i|W_i=w}^N \mathbb{E}\left\|\Omega_{w,K}^{-\frac{1}{2}} R_K(X_i)\varepsilon_{w,i}\right\|^3 \\
&\leq \left(N_w \cdot \sigma_w^2\right)^{-\frac{3}{2}} \sum_{i|W_i=w}^N \mathbb{E}\left[\|\Omega_{w,K}^{-\frac{1}{2}}\|^3 \|R_K(X_i)\varepsilon_{w,i}\|^3\right]
\end{aligned}
$$

First, consider

$$\|\Omega_{w,K}^{-\frac{1}{2}}\|^3 = \left[\text{tr}(\Omega_{w,K}^{-1})\right]^{\frac{3}{2}} \leq \left[K \cdot \lambda_{max}(\Omega_{w,K}^{-1})\right]^{\frac{3}{2}} \leq C \cdot K^{\frac{3}{2}}$$

which is $O(K^{\frac{3}{2}})$ because $\lambda_{min}(\Omega_{w,K})$ is bounded away from zero by Lemma $A.1$. Next, consider

$$\mathbb{E}\|R_K(X_i)\varepsilon_{w,i}\|^3 \leq \sup_x \|R_K(x)\|^3 \cdot \mathbb{E}|\varepsilon_{w,i}|^3 \leq C \cdot K^3$$

where the third moment of $\varepsilon_{w,i}$ is bounded by Assumption 3.2 and so the factor is $O(K^3)$. Since σ_w^2 is also bounded by Assumption 3.2, β_3 is $O(K^{\frac{9}{2}}N^{-\frac{1}{2}})$. Thus,

$$C\beta_3 K^{1/4} = C \cdot \sum_{i|W_i=w}^N \mathbb{E}\left\|\frac{Z_i}{\sqrt{N_w}}\right\|^3 K^{1/4} \leq C_1 \cdot K^{\frac{9}{2}} N_w^{-\frac{1}{2}} \cdot K^{1/4} = C_1 \cdot K^{\frac{19}{4}} N_w^{-1/2}$$

and the result follows. $\blacksquare$

We may proceed further to detail conditions under which the quadratic form, $S'_{w,K}S_{w,K}$, properly normalized, converges to a univariate standard Gaussian distribution. The quadratic form $S'_{w,K}S_{w,K}$ can be written as

$$S'_{w,K}S_{w,K} = \sum_{j=1}^K \left(\frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^N Z_{ij}\right)^2$$

where Z_{ij} is the jth element of the vector Z_i. Thus, $S'_{w,K}S_{w,K}$ is a sum of K uncorrelated, squared random variables with each random variable converging to a standard Gaussian distribution by the previous result. Intuitively, this sum should converge to a chi-squared random variable with K degrees of freedom.

[17]

Lemma A.4 *Under Assumptions 2.1-2.3 and 3.1-3.3,*

$$\sup_c \left| \Pr(S'_{w,K} S_{w,K} \le c) - \chi^2_K(c) \right| \to 0.$$

Proof Define the set $A(c) \equiv \left\{ S \in \mathbb{R}^K \,|\, S'S \le c \right\}$. Note that $A(c)$ is a measurable convex set in $\mathbb{R}^K$. Also note that for $Z_K \sim \mathcal{N}(0, I_K)$, we have that $\chi^2_K(c) = \mathrm{P}(Z'_K Z_K \le c)$. Then,

$$
\begin{aligned}
\sup_c \left| \Pr[S'_{w,K} S_{w,K} \le c] - \chi^2_K(c) \right| &= \sup_c \left| \Pr(S'_{w,K} S_{w,K} \le c) - \Pr(Z'_K Z_K \le c) \right| \\
&= \sup_c \left| \Pr(S_{w,K} \in A(c)) - \Pr(Z_K \in A(c)) \right| \\
&\le \sup_{\mathcal{A} \in A_K} \left| Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A}) \right| \\
&\le C \beta_3 K^{1/4} \\
&= O(K^{\frac{19}{4}} N_w^{-1/2})
\end{aligned}
$$

which is o(1) for $\nu < \frac{2}{19}$ by Lemma $A.3$. ∎

The proper normalization of the quadratic form yields the studentized version, $(S'_{w,K} S_{w,K} - K)/\sqrt{2K}$. This converges to a standard Gaussian distribution by the following lemma.

Lemma A.5 *Under Assumptions 2.1-2.3 and 3.1-3.3,*

$$\sup_c \left| \Pr \left(\frac{S'_{w,K} S_{w,K} - K}{\sqrt{2K}} \le c \right) - \Phi(c) \right| \to 0.$$

Proof

$$
\begin{aligned}
&\sup_c \left| \Pr \left(\frac{S'_{w,K} S_{w,K} - K}{\sqrt{2K}} \le c \right) - \Phi(c) \right| \\
=\;& \sup_c \left| \Pr \left(S'_{w,K} S_{w,K} \le K + c\sqrt{2K} \right) - \Phi(c) \right| \\
\le\;& \sup_c \left| \Pr \left(S'_{w,K} S_{w,K} \le K + c\sqrt{2K} \right) - \chi^2(K + c\sqrt{2K}) \right| + \sup_c \left| \chi^2(K + c\sqrt{2K}) - \Phi(c) \right|
\end{aligned}
$$

The first term goes to zero by Lemma $A.4$. For the second term we may apply the Berry-Esséen Theorem which yields,

$$\sup_c \left| \Pr \left(\frac{Z'_K Z_K - K}{\sqrt{2K}} \le c \right) - \Phi(c) \right| \le C \cdot K^{-\frac{1}{2}}.$$

Thus for $\nu > 0$ the right-hand side converges to zero as well and the result is established. ∎

In order to proceed we need the following selected results from Imbens, Newey and Ridder (2006). These results establish convergence rates for the estimators of the regression function.

Lemma A.6 (IMBENS, NEWEY AND RIDDER): *Suppose Assumptions 3.1 - 3.3 hold. Then,*

(i) there is a sequence $\gamma^0_{w,K}$ such that

$$\sup_x \left| \mu_w(x) - R_K(x)' \gamma^0_{w,K} \right| \equiv \sup_x \left| \mu_w(x) - \mu^0_{w,K} \right| = O \left(K^{-\frac{s}{d}} \right)$$

(ii)

$$\sup_x \left| R_K(x)' \gamma^*_{w,K} - R_K(x)' \gamma^0_{w,K} \right| \equiv \sup_x \left| \mu^*_{w,K} - \mu^0_{w,K} \right| = O \left(\zeta(K) K^{\frac{1}{2}} K^{-\frac{s}{d}} \right)$$

[18]

(iii)

$$\left\| \gamma_{w,K}^* - \gamma_{w,k}^0 \right\| = O\left(K^{\frac{1}{2}} K^{-\frac{s}{d}} \right)$$

(iv)

$$\left\| \hat\gamma_{w,K} - \gamma_{w,k}^0 \right\| = O_p\left(K^{\frac{1}{2}} N^{-\frac{1}{2}} + K^{-\frac{s}{d}} \right)$$

The following lemma describes the limiting distribution of the infeasible test statistic.

Lemma A.7 *Under Assumptions 2.1-2.3 and 3.1−3.3,*

$$\left(N_w \cdot \left(\hat\gamma_{w,K} - \gamma_{w,K}^* \right)' \left(\hat\sigma_{w,K}^2 \cdot \hat\Omega_{w,K}^{-1} \right)^{-1} \left(\hat\gamma_{w,K} - \gamma_{w,K}^* \right) - K \right) / \sqrt{2K} \xrightarrow{d} \mathcal{N}(0,1)$$

Proof We need only show that,

$$\left\| \left[\hat\sigma_{w,K}^2 \cdot \hat\Omega_{w,K}^{-1} \right]^{-\frac{1}{2}} \sqrt{N_w} \left(\hat\gamma_{w,K} - \gamma_{w,K}^* \right) - S_{w,K} \right\| = o_p(1).$$

then the result follows by Lemmas $(A.3)$, $(A.4)$, and $(A.5)$.

First, notice that we can rewrite $\hat\gamma_{w,K}$ as

$$\hat\gamma_{w,K} = \gamma_{w,K}^* + \hat\Omega_{w,K}^{-1} R_{w,K}' \varepsilon_{w,K}^* / N_w$$

where

$$\varepsilon_{w,K}^* \equiv Y_w - R_{w,K} \gamma_{w,K}^*.$$

Then,

$$
\begin{aligned}
& \left\| \left[\hat\sigma_{w,K}^2 \cdot \hat\Omega_{w,K}^{-1} \right]^{-\frac{1}{2}} \sqrt{N_w} \left(\hat\gamma_{w,K} - \gamma_{w,K}^* \right) - S_{w,K} \right\| \\
=\; & \left\| \left[\hat\sigma_{w,K}^2 \cdot \hat\Omega_{w,K}^{-1} \right]^{-\frac{1}{2}} \sqrt{N_w} \cdot \hat\Omega_{w,K}^{-1} \cdot R_{w,K}' \varepsilon_{w,K}^* / N_w - \left[\sigma_w^2 \cdot \Omega_{w,K} \right]^{-\frac{1}{2}} \sqrt{N_w} \cdot R_{w,K}' \varepsilon_w / N_w \right\| \\
=\; & \left\| \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_{w,K}^* / \sqrt{N_w} - \sigma_w^{-1} \Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| \\
=\; & \left\| \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_{w,K}^* / \sqrt{N_w} - \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right. \\
& \qquad + \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} - \sigma_w^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \\
& \qquad \left. + \sigma_w^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} - \sigma_w^{-1} \Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| \\
\leq\; & \left\| \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_{w,K}^* / \sqrt{N_w} - \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| \\
& + \left\| \hat\sigma_{w,K}^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} - \sigma_w^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| \\
& + \left\| \sigma_w^{-1} \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} - \sigma_w^{-1} \Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| \\
=\; & \left| \hat\sigma_{w,K}^{-1} \right| \left\| \hat\Omega_{w,K}^{-\frac{1}{2}} R_{w,K}' \left(\varepsilon_{w,K}^* - \varepsilon_w \right) / \sqrt{N_w} \right\| & \text{(A.4)} \\
& + \left| \hat\sigma_{w,K}^{-1} - \sigma_w^{-1} \right| \left\| \hat\Omega_{w,K}^{-\frac{1}{2}} \cdot R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| & \text{(A.5)} \\
& + \left| \sigma_w^{-1} \right| \left\| \left(\hat\Omega_{w,K}^{-\frac{1}{2}} - \Omega_{w,K}^{-\frac{1}{2}} \right) R_{w,K}' \varepsilon_w / \sqrt{N_w} \right\| & \text{(A.6)}
\end{aligned}
$$

[19]

First, consider equation $(A.4)$,

$$\left|\hat{\sigma}_{w,K}^{-1}\right|\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}R'_{w,K}\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)/\sqrt{N_{w}}\right\|$$

$$=\quad\left(\sigma_{w}^{-1}+o_{p}\left(1\right)\right)\cdot\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}R'_{w,K}\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)/\sqrt{N_{w}}\right\|$$

$$=\quad\left(O\left(1\right)+o_{p}\left(1\right)\right)\cdot\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}R'_{w,K}\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)/\sqrt{N_{w}}\right\|$$

where the consistency of the sample variance follows by Lemma B.2 in the supplementary materials on the website.

$$\mathbb{E}\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}R'_{w,K}\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)/\sqrt{N_{w}}\right\|^{2}$$

$$=\quad\mathbb{E}\left[\frac{1}{N_{w}}\mathrm{tr}\left(\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)'R_{w,K}\hat{\Omega}_{w,K}^{-1}R'_{w,K}\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)\right)\right]$$

$$=\quad\mathbb{E}\left[\left(\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)'R_{w,K}\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)\right)\right]$$

$$\leq\quad\mathbb{E}\left[\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)'\left(\varepsilon_{w,K}^{*}-\varepsilon_{w}\right)\right] \tag{A.7}$$

$$=\quad\mathbb{E}\left[\left(\mu_{w}(\mathbf{X})-R_{w,K}\gamma_{w,K}^{*}\right)'\left(\mu_{w}(\mathbf{X})-R_{w,K}\gamma_{w,K}^{*}\right)\right]$$

$$\leq\quad N_{w}\cdot\sup_{x}\left|\mu_{w}(x)-R_{K}(x)'\gamma_{w,K}^{*}\right|^{2}$$

$$\leq\quad N_{w}\cdot\sup_{x}\left(\left|\mu_{w}(x)-R_{K}(x)'\gamma_{w,K}^{0}\right|+\left|R_{K}(x)'\gamma_{w,K}^{0}-R_{K}(x)'\gamma_{w,K}^{*}\right|\right)^{2}$$

$$=\quad N_{w}\left(O\left(K^{-\frac{s}{d}}\right)+O\left(\zeta(K)K^{\frac{1}{2}}K^{-\frac{s}{d}}\right)\right)^{2} \tag{A.8}$$

$$=\quad O\left(N\right)\cdot\left(O\left(\zeta(K)K^{\frac{1}{2}}K^{-\frac{s}{d}}\right)\right)^{2}$$

so that equation $(A.4)$ is $O_{p}\left(\zeta(K)K^{\frac{1}{2}}K^{-\frac{s}{d}}N^{\frac{1}{2}}\right)$ by Markov's inequality. $(A.7)$ follows by the fact that $(I_{N_{w}}-R_{w,K}(R'_{w,K}R_{w,K})^{-1}R'_{w,K})$ is a projection matrix and is thus positive semi-definite. $(A.8)$ follows from Lemma $A.6$ (i) and (ii).

Next consider equation $(A.5)$. We will work first with the second factor,

$$\mathbb{E}\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}\cdot R'_{w,K}\varepsilon_{w}/\sqrt{N_{w}}\right\|^{2}$$

$$=\quad\mathbb{E}\left[\frac{1}{N_{w}}\mathrm{tr}\left(\varepsilon_{w}'R_{w,K}\hat{\Omega}_{w,K}^{-1}R'_{w,K}\varepsilon_{w}\right)\right]$$

$$=\quad\mathbb{E}\left[\mathrm{tr}\left(\varepsilon_{w}'R_{w,K}\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}\varepsilon_{w}\right)\right]$$

$$=\quad\mathbb{E}\left[\mathrm{tr}\left(R_{w,K}\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}\varepsilon_{w}\varepsilon_{w}'\right)\right]$$

$$=\quad\mathrm{tr}\left(\mathbb{E}\left[R_{w,K}\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}\mathbb{E}\left[\varepsilon_{w}\varepsilon_{w}'|\mathbf{X}\right]\right]\right)$$

$$=\quad\sigma_{w}^{2}\cdot\mathrm{tr}\left(\mathbb{E}\left[R_{w,K}\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}\right]\right)$$

$$=\quad\sigma_{w}^{2}\cdot\mathbb{E}\left[\mathrm{tr}\left(R_{w,K}\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}\right)\right]$$

$$=\quad\sigma_{w}^{2}\cdot\mathbb{E}\left[\mathrm{tr}\left(\left(R'_{w,K}R_{w,K}\right)^{-1}R'_{w,K}R_{w,K}\right)\right]$$

$$=\quad\sigma_{w}^{2}\cdot\mathrm{tr}\left(I_{K}\right)$$

$$=\quad\sigma_{w}^{2}\cdot K$$

[20]

so that the second factor is $O\left(K^{\frac{1}{2}}\right)$ by Markov's inequality. Then by Lemma B.2, equation $(A.5)$ is $o_p(1)$.

Finally, consider equation $(A.6)$,

$$\left|\sigma_w^{-1}\right|\left\|\left(\hat{\Omega}_{w,K}^{-\frac{1}{2}}-\Omega_{w,K}^{-\frac{1}{2}}\right)R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|$$
$$\leq\quad C\cdot\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}-\Omega_{w,K}^{-\frac{1}{2}}\right\|\left\|R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|$$

The first factor is $O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right)$ by Lemma $A.1$ and the continuous mapping theorem, and

$$\mathbb{E}\left\|R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|^2$$
$$=\quad\mathbb{E}\left[\frac{1}{N_w}\text{tr}\left(\varepsilon'_w R_{w,K}R'_{w,K}\varepsilon_w\right)\right]$$
$$=\quad\mathbb{E}\left[\frac{1}{N_w}\text{tr}\left(R'_{w,K}\varepsilon_w\varepsilon'_w R_{w,K}\right)\right]$$
$$=\quad\text{tr}\left(\frac{1}{N_w}\mathbb{E}\left[R'_{w,K}\mathbb{E}\left[\varepsilon_w\varepsilon'_w|\mathbf{X}\right]R_{w,K}\right]\right)$$
$$=\quad\sigma_w^2\cdot\text{tr}\left(\mathbb{E}\left[R'_{w,K}R_{w,K}/N_w\right]\right)$$
$$=\quad\sigma_w^2\cdot\text{tr}\left(\Omega_{w,K}\right)$$
$$\leq\quad\sigma_w^2\cdot K\cdot\lambda_{max}\left(\Omega_{w,K}\right)$$
$$\leq\quad C\cdot K$$

so that the second factor is $O\left(K^{\frac{1}{2}}\right)$ by Assumption 3.2, Lemma $A.1$ (ii) and Markov's inequality. Thus, equation $(A.6)$ is $O_p\left(\zeta(K)KN^{-\frac{1}{2}}\right)$.

Combining these results yields:

$$\left\|\left[\hat{\sigma}_{w,K}^2\cdot\hat{\Omega}_{w,K}^{-1}\right]^{-\frac{1}{2}}\sqrt{N_w}\left(\hat{\gamma}_{w,K}-\gamma^*_{w,K}\right)-S_{w,K}\right\|$$
$$=\quad O_p\left(\zeta(K)K^{\frac{1}{2}}K^{-\frac{s}{d}}N^{\frac{1}{2}}\right)+o_p\left(1\right)+O_p\left(\zeta(K)KN^{-\frac{1}{2}}\right)$$

which is $o_p(1)$ under Assumptions 3.2 and 3.3. ∎

Proof of Theorem 3.1: First, note that by Lemma $A.7$ we have that

$$\sqrt{N_w}\left(\hat{\gamma}_{w,K}-\gamma^*_{w,K}\right)\overset{d}{\to}\mathcal{N}\left(0,\sigma_{w,K}^2\cdot\Omega_{w,K}^{-1}\right). \tag{A.9}$$

We can rewrite this result as $\frac{\sqrt{N_w}}{\sqrt{N}}\cdot\sqrt{N}\left(\hat{\gamma}_{w,K}-\gamma^*_{w,K}\right)$. Then we have that,

$$\sqrt{N}\left(\hat{\gamma}_{0,K}-\gamma^*_{0,K}\right)\overset{d}{\to}\mathcal{N}\left(0,\frac{\sigma_{0,K}^2}{1-c}\cdot\Omega_{0,K}^{-1}\right) \tag{A.10}$$

and

$$\sqrt{N}\left(\hat{\gamma}_{1,K}-\gamma^*_{1,K}\right)\overset{d}{\to}\mathcal{N}\left(0,\frac{\sigma_{1,K}^2}{c}\cdot\Omega_{1,K}^{-1}\right). \tag{A.11}$$

We may follow the logic of Lemmas $(A.3)$, $(A.4)$, and $(A.5)$ to conclude that

$$T^* \equiv \left(N \cdot \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - (\gamma_{1,K}^* - \gamma_{0,K}^*) \right)' \cdot \hat{V}^{-1} \cdot \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - (\gamma_{1,K}^* - \gamma_{0,K}^*) \right) - K \right) / \sqrt{2K}$$

converges in distribution to a $\mathcal{N}(0,1)$ random variable. To complete the proof we must show that $|T^* - T| = o_p(1)$.

Note that under the null hypothesis $\mu_1(x) = \mu_0(x)$ so we may choose the same approximating sequence $\gamma_{1,K}^0 = \gamma_{0,K}^0$ for $\mu_{1,K}^0(x) = \mu_{0,K}^0(x)$. Then,

$$
\begin{aligned}
\left\| \gamma_{1,K}^* - \gamma_{0,K}^* \right\| &= \left\| \gamma_{1,K}^* - \gamma_{1,K}^0 + \gamma_{0,K}^0 - \gamma_{0,K}^* \right\| \\
&\le \left\| \gamma_{1,K}^* - \gamma_{1,K}^0 \right\| + \left\| \gamma_{0,K}^0 - \gamma_{0,K}^* \right\| \\
&= O(K^{\frac{1}{2}} K^{-\frac{s}{d}})
\end{aligned}
\tag{A.12}
$$

by Lemma $A.6$ (iii), and

$$
\begin{aligned}
\left\| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right\| &= \left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 + \gamma_{0,K}^0 - \hat{\gamma}_{0,K} \right\| \\
&\le \left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 \right\| + \left\| \gamma_{0,K}^0 - \hat{\gamma}_{0,K} \right\| \\
&= O_p(K^{\frac{1}{2}} N^{-\frac{1}{2}} + K^{-\frac{s}{d}})
\end{aligned}
\tag{A.13}
$$

by Lemma $A.6$ (iv). So then,

$$
\begin{aligned}
|T^* - T| &= \Bigg| \left(N \cdot \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - (\gamma_{1,K}^* - \gamma_{0,K}^*) \right)' \hat{V}^{-1} \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - (\gamma_{1,K}^* - \gamma_{0,K}^*) \right) - K \right) / \sqrt{2K} \\
&\quad - \left(N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K \right) / \sqrt{2K} \Bigg| \tag{A.14} \\
&= \frac{N}{\sqrt{2K}} \cdot \Bigg| \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - (\gamma_{1,K}^* - \gamma_{0,K}^*) \right)' \hat{V}^{-1} \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - (\gamma_{1,K}^* - \gamma_{0,K}^*) \right) \\
&\quad - (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) \Bigg| \tag{A.15} \\
&= \frac{N}{\sqrt{2K}} \cdot \left| -2 \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) + (\gamma_{1,K}^* - \gamma_{0,K}^*)' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right| \tag{A.16} \\
&\le \frac{N}{\sqrt{2K}} \cdot 2 \cdot \left| (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right| \tag{A.17} \\
&\quad + \frac{N}{\sqrt{2K}} \cdot \left| (\gamma_{1,K}^* - \gamma_{0,K}^*)' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right| \tag{A.18}
\end{aligned}
$$

Consider $(A.17)$,

$$
\begin{aligned}
2 \cdot \left| (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right| &= 2 \cdot \left| \text{tr} \left((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right) \right| \tag{A.19} \\
&\le 2 \cdot \left\| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right\| \cdot \left\| \gamma_{1,K}^* - \gamma_{0,K}^* \right\| \cdot \lambda_{max}(\hat{V}^{-1}) \tag{A.20} \\
&= \left(O_p(K^{\frac{1}{2}} N^{-\frac{1}{2}} + K^{-\frac{s}{d}}) \cdot O(K^{\frac{1}{2}} K^{-\frac{s}{d}}) \right) \tag{A.21}
\end{aligned}
$$

Where $(A.21)$ follows from Lemma $A.2$, $(A.12)$, $(A.13)$ and Assumption 3.3.

Now, consider $(A.18)$,

$$
\begin{aligned}
\left| (\gamma_{1,K}^* - \gamma_{0,K}^*)' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right| &= \left| \text{tr} \left((\gamma_{1,K}^* - \gamma_{0,K}^*)' \hat{V}^{-1} (\gamma_{1,K}^* - \gamma_{0,K}^*) \right) \right| \tag{A.22} \\
&\le \left\| \gamma_{1,K}^* - \gamma_{0,K}^* \right\|^2 \cdot \lambda_{max}(\hat{V}^{-1}) \tag{A.23} \\
&= O(K K^{-\frac{2s}{d}}) \tag{A.24}
\end{aligned}
$$

Where $(A.24)$ follows from Lemma $A.2$, $(A.12)$ and Assumption 3.3.

So then,

$$\begin{aligned}
|T^* - T| &= \frac{N}{\sqrt{2K}} \cdot \left(O_p(K^{\frac{1}{2}} N^{-\frac{1}{2}} + K^{-\frac{s}{d}}) \cdot O(K^{\frac{1}{2}} K^{-\frac{s}{d}}) + O(K K^{-\frac{2s}{d}}) \right) \\
&= O_p\left(N^{\frac{1}{2}} K^{\frac{1}{2}} K^{-\frac{s}{d}} \right) + O_p\left(N K^{-\frac{2s}{d}} \right) + O\left(N K^{\frac{1}{2}} K^{-\frac{2s}{d}} \right)
\end{aligned}$$

All three terms are $o_p(1)$ under Assumptions 3.2 and 3.3 and the result follows. ∎

Proof of Theorem 3.2 First, note that $\zeta(K) = \sup_x \|R_K(x)\|$ satisfies $\underline{C} \cdot K < \zeta(K) < \overline{C} \cdot K$ for some $0 < \underline{C}, \overline{C} < \infty$. Second,

$$\begin{aligned}
\rho_N \cdot \sup_{x \in \mathbb{X}} |\Delta(x)| &= \sup_x |\mu_1(x) - \mu_0(x)| \\
&\leq \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right| + \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right| \\
&\quad + \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - R_K(x)' \gamma_{0,K}^0 \right| \\
&\leq \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right| + \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right| \\
&\quad + \sup_{x \in \mathbb{X}} \left| R_K(x)' \hat{\gamma}_{0,K} - R_K(x)' \gamma_{0,K}^0 \right| + \sup_{x \in \mathbb{X}} \left| R_K(x)' \hat{\gamma}_{1,K} - R_K(x)' \gamma_{1,K}^0 \right| \\
&\quad + \sup_{x \in \mathbb{X}} \left| R_K(x)' \hat{\gamma}_{1,K} - R_K(x)' \hat{\gamma}_{0,K} \right| \\
&\leq \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right| + \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right| \\
&\quad + \sup_{x \in \mathbb{X}} \|R_K(x)\| \cdot \left\| \hat{\gamma}_{0,K} - \gamma_{0,K}^0 \right\| + \sup_{x \in \mathbb{X}} \|R_K(x)\| \cdot \left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 \right\| \\
&\quad + \sup_{x \in \mathbb{X}} \|R_K(x)\| \cdot \left\| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right\| \\
&= \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right| + \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right| \\
&\quad + \zeta(K) \cdot \left\| \hat{\gamma}_{0,K} - \gamma_{0,K}^0 \right\| + \zeta(K) \cdot \left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 \right\| + \zeta(K) \cdot \left\| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right\|
\end{aligned}$$

Thus

$$\begin{aligned}
\left\| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right\| &\geq \zeta^{-1}(K) \cdot \rho_N \cdot \sup_{x \in \mathbb{X}} |\Delta(x)| - \zeta^{-1}(K) \cdot \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right| \\
&\quad - \zeta^{-1}(K) \cdot \sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right| - \left\| \hat{\gamma}_{0,K} - \gamma_{0,K}^0 \right\| - \left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 \right\| \\
&\geq \zeta^{-1}(K) \cdot \rho_N \cdot C_0 \cdot \left(1 - \frac{\sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right|}{\rho_N \cdot C_0} \right. \\
&\quad \left. - \frac{\sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right|}{\rho_N \cdot C_0} - \zeta(K) \frac{\left\| \hat{\gamma}_{0,K} - \gamma_{0,K}^0 \right\|}{\rho_N \cdot C_0} - \zeta(K) \frac{\left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 \right\|}{\rho_N \cdot C_0} \right).
\end{aligned}$$

Because $s/d > 25/4$ by Assumption 3.2 and $1/(2s/d + 3) < \nu < 2/19$ by Assumption 3.3,

$$\frac{\sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{0,K}^0 - \mu_0(x) \right|}{\rho_N \cdot C_0} = O\left(K^{-s/d} \right) \cdot O\left(N^{1/2 - 3\nu/2 - \varepsilon} \right) = o(1),$$

[23]

$$\frac{\sup_{x \in \mathbb{X}} \left| R_K(x)' \gamma_{1,K}^0 - \mu_1(x) \right|}{\rho_N \cdot C_0} = O\left(K^{-s/d}\right) \cdot O\left(N^{1/2 - 3\nu/2 - \varepsilon}\right) = o(1),$$

$$\zeta(K) \frac{\left\| \hat{\gamma}_{0,K} - \gamma_{0,K}^0 \right\|}{\rho_N \cdot C_0} = O\left(K\right) \cdot O_p\left(K^{1/2} N^{-1/2}\right) \cdot O\left(N^{1/2 - 3\nu/2 - \varepsilon}\right) = o_p(1),$$

and

$$\zeta(K) \frac{\left\| \hat{\gamma}_{1,K} - \gamma_{1,K}^0 \right\|}{\rho_N \cdot C_0} = O\left(K\right) \cdot O_p\left(K^{1/2} N^{-1/2}\right) \cdot O\left(N^{1/2 - 3\nu/2 - \varepsilon}\right) = o_p(1),$$

it follows that

$$\| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \| \geq \zeta^{-1}(K) \cdot \rho_N \cdot C_0$$

with probability going to one as $N \to \infty$. Thus

$$N^{1/2} K^{-1/2} \| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \| \geq N^{1/2} K^{-1/2} \zeta^{-1}(K) \cdot \rho_N \cdot C_0$$

with probability going to one as $N \to \infty$. Since

$$N^{1/2} K^{-1/2} \zeta^{-1}(K) \cdot \rho_N \cdot C_0 \geq C N^{1/2} K^{-1/2} \zeta^{-1}(K) N^{-1/2 + 3\nu/2 + \varepsilon} \geq C N^\varepsilon$$

which goes to infinity with the sample size, it follows that for any M',

$$\Pr\left(N^{1/2} K^{-1/2} \| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \| > M' \right) \longrightarrow 1. \tag{A.25}$$

Next, we show that this implies that

$$\Pr\left(\frac{N \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right)' V^{-1} \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right) - K}{\sqrt{2K}} > 2M \right) \longrightarrow 1.$$

Let $\lambda_{\min}(A)$ be the minimum eigenvalue of a matrix A. Denote $\lambda_{\min}(V^{-1})$ by $\underline{\lambda}$ and note that by Lemma A.2 it follows that $\underline{\lambda}$ is bounded away from zero.

$$\Pr\left(\frac{N \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right)' V^{-1} \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right) - K}{\sqrt{2K}} > 2M \right)$$

$$= \Pr\left(N \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right)' V^{-1} \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right) > M\sqrt{8} K^{1/2} + K \right)$$

$$\geq \Pr\left(N\underline{\lambda} \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right)' \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right) > M\sqrt{8} K^{1/2} + K \right)$$

$$= \Pr\left(N K^{-1} \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right)' \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right) > \underline{\lambda}^{-1} \left(1 + M\sqrt{8} K^{-1/2} \right) \right)$$

$$= \Pr\left(N^{1/2} K^{-1/2} \| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \| > \underline{\lambda}^{-1/2} \left(1 + M\sqrt{8} K^{-1/2} \right)^{1/2} \right).$$

Since for any M, for large enough N, we have $\underline{\lambda}^{-1/2} \left(1 + M\sqrt{8} K^{-1/2} \right)^{1/2} < 2\underline{\lambda}^{-1/2}$, it follows that this probability is for large N bounded from below by the probability

$$= \Pr\left(N^{1/2} K^{-1/2} \| \hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \| > 2\underline{\lambda}^{-1/2} \right),$$

which goes to one by (A.25).

Finally, we show that this implies that

$$\Pr(T > M) = \Pr\left(\frac{N \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right)' \hat{V}^{-1} \left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K} \right) - K}{\sqrt{2K}} > M \right) \longrightarrow 1.$$

[24]

Let $\hat{\underline{\lambda}} = \lambda_{\min}(\hat{V}^{-1})$ be the minimum eigenvalue of the matrix $\hat{V}^{-1}$. Lemmas A.1 and A.2, Assumption 3.3 and the consistency of $\hat{\sigma}_{0,K}^2$, $\hat{\sigma}_{1,K}^2$ and $\hat{c}$ imply that $\hat{\underline{\lambda}} - \underline{\lambda} = o_p(1)$. Since $\underline{\lambda}$ is bounded away from zero, it follows that $\hat{\underline{\lambda}}$ is bounded away from zero with probability going to one. Let A denote the event that $\lambda_{\min}(\hat{V}^{-1}) > \underline{\lambda}/2$ and $\frac{N(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-K}{\sqrt{2K}} > 2M/\underline{\lambda}$. The probability of the event A converges to one since,

$$\Pr\left(\frac{N\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right)' V^{-1}\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right) - K}{\sqrt{2K}} > 2M\right) \geq \Pr\left(\frac{N\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right)\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right) - K}{\sqrt{2K}} > 2M/\underline{\lambda}\right)$$

Also, A implies that

$$\frac{N\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right)' \hat{V}^{-1}\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right) - K}{\sqrt{2K}}$$

$$\geq \frac{N\hat{\underline{\lambda}}\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right)'\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right) - K}{\sqrt{2K}}$$

$$> \frac{N\underline{\lambda}/2\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right)'\left(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\right) - K}{\sqrt{2K}} > \underline{\lambda}/2 \cdot 2M/\underline{\lambda} = M.$$

Hence $\Pr(T > M)) \to 1$. ∎

References

ABADIE, A., (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, Vol 97, 284-292.

ABADIE, A., J. ANGRIST, AND G. IMBENS, (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica.* Vol. 70, No. 1, 91-117.

ABADIE, A., AND G. IMBENS, (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica.* Vol. 74, No. 1, 235-267

ANGRIST, J. D. AND A. B. KRUEGER (2000), "Empirical Strategies in Labor Economics," in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

BENTKUS, V., (2005), "A Lyapunov-type Bound in R^d," *Theory of Probability and Applications*, Vol 49(2), 311-322.

BIERENS, H., (1982), "Consistent Model Specification Tests," *Journal of Econometrics*, Vol 20, 105-134.

BIERENS, H., (1990), "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, Vol 58, 1443-1458.

BITLER, M., J. GELBACH, AND H. HOYNES (2002) "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," unpublished paper, Department of Economics, University of Maryland.

BLUNDELL, R. AND M. COSTA-DIAS (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02.

CHEN, X., (2005), "Large Sample Sieve Estimation of Semi-Nonparametric Models," forthcoming, *H*andbook of Econometrics, Vol VI, Heckman and Leamer (eds), North-Holland Publishers, Amsterdam.

CHEN, X., HONG, H., AND TAROZZI, A., (2004), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Error, Missing Data and Treatment Effects." Working Paper.

CHERNOZHUKOV, V., AND C. HANSEN., (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, Vol. 73, No 1., 245-261.

CRUMP, R., V. J. HOTZ, V. J., G. IMBENS, AND O. MITNIK, (2006), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," unpublished manuscript, Department of Economics, UC Berkeley.

DE JONG, R., AND H. BIERENS, (1994), "On the Limit of a Chi-Square Type Test if the Number of Conditional Moments Tested Approaches Infinity," *Econometric Theory*, Vol 9, 70-90.

DOKSUM, K., (1974), "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *The Annals of Statistics*, Vol 2, 267-277.

EUBANK, R., AND C. SPIEGELMAN, (1990), "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *Journal of the American Statistical Association*, Vol 85, 387-392.

FIRPO, S., (2004), "Efficient Semiparametric Estimation of Quantile Treatment Effects" Working Paper.

GUERON, J., AND E. PAULY, (1991), *From Welfare to Work*, Russell Sage Foundation, New York.

HAHN, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,"*Econometrica* 66 (2), 315-331.

HÄRDLE, W., AND E. MAMMEN, (1993), "Comparing Nonparametric Versus Parametric Regression Fits,"*The Annals of Statistics*, Vol 21(4), 1926-1947.

HÄRDLE, W., AND J. MARRON, (1990), "Semiparametric Comparison of Regression Curves,"*The Annals of Statistics*, Vol 18(1), 63-89.

HECKMAN, J., AND V. J. HOTZ, (1989), "Alternative Methods for Evaluating the Impact of Training Programs," (with discussion), *Journal of the American Statistical Association.*, 84(804): 862-874.

HECKMAN, J., H. ICHIMURA, AND P. TODD, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261–294.

HECKMAN, J., R. LALONDE, AND J. SMITH (2000), "The Economics and Econometrics of Active Labor Markets Programs," in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

HECKMAN, J., AND R. ROBB, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

HIRANO, K., G. IMBENS, AND G. RIDDER, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,"*Econometrica*, 71(4): 1161-1189.

HONG, Y., AND H. WHITE, (1995), "Consistent Specification Testing Via Nonparametric Series Regression,"*Econometrica*, Vol 63(5), 1133-1159.

HOROWITZ, J., AND V. SPOKOINY, (2001), "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica*, 69(3): 599-631.

HOTZ, V.J., G. IMBENS AND J. KLERMAN, (2006), "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program," forthcoming in *Journal of Labor Economics*.

HOTZ, V.J., G. IMBENS AND J. MORTIMER, (2005), "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics*, Vol. 125, 241-270.

IMBENS, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86(1): 1-29.

IMBENS, G., W. NEWEY AND G. RIDDER, (2006), "Mean-squared-error Calculations for Average Treatment Effects," unpublished manuscript, Department of Economics, UC Berkeley.

LECHNER, M, (2002), "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society*, Series A, 165: 659–82.

LEE, M.-J., (2005), *Micro-Econometrics for Policy, Program, and Treatment Effects* Oxford University Press, Oxford.

LEHMANN, E., (1974), *Nonparametrics: Statistical Methods Based on Ranks* Francisco, CA: Holden-Day.

LI, C.-K., AND R. MATHIAS, (2002), "Interlacing Inequalities for Totally Nonnegative Matrices,"*Linear Algebra and its Applications*, Vol 341, 35-44.

Neumeyer, N., and H. Dette, (2003), "Nonparametric Comparison of Regression Curves: An Empirical Process Approach," *The Annals of Statistics*, Vol 31, 880-920.

Pinkse, J., and P. Robinson, (1995), "Pooling Nonparametric Estimates of Regression Functions with a Similar Shape," in *Statistical Methods of Econometrics and Quantitative Economics: A Volume in Honour of C.R. Rao*, G.S. Maddala, P.C.B. Phillips and T.N. Srinivisan, eds., 172-197.

Riccio, J., D. Friedlander and S. Freedman, (1994) *GAIN: Benefits, costs, and three-year impacts of a welfare-to-work program.* Manpower Demonstration Research Corporation, New York.

Rosenbaum, P., (1997), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292–316.

Rosenbaum, P., (2001), *Observational Studies*, second edition, Springer Verlag, New York.

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

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66: 688-701.

Wooldridge, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Table 1: SUMMARY STATISTICS EXPERIMENTAL GAIN DATA

	Los Angeles (LA) $N_1 = 2995,$ $N_0 = 1400$		Riverside (RI) $N_1 = 4405,$ $N_0 = 1040$		Alameda (AL) $N_1 = 597,$ $N_0 = 601$		San Diego (SD) $N_1 = 6978,$ $N_0 = 1154$	
	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
Female	0.94	(0.24)	0.88	(0.33)	0.95	(0.22)	0.84	(0.37)
Age	38.52	(8.43)	33.64	(8.20)	34.72	(8.62)	33.80	(8.59)
Age-squared/100	15.55	(6.83)	11.99	(5.96)	12.79	(6.41)	12.16	(6.24)
Hispanic	0.32	(0.47)	0.27	(0.45)	0.08	(0.26)	0.25	(0.44)
Black	0.45	(0.50)	0.16	(0.36)	0.70	(0.46)	0.23	(0.42)
HS Diploma	0.35	(0.48)	0.52	(0.50)	0.59	(0.49)	0.57	(0.50)
1 Child	0.33	(0.47)	0.39	(0.49)	0.42	(0.49)	0.43	(0.50)
Children under 6	0.10	(0.30)	0.16	(0.37)	0.31	(0.46)	0.13	(0.34)
Earnings Q-1/1,000	0.22	(0.87)	0.45	(1.41)	0.21	(0.85)	0.59	(1.48)
Earnings Q-2/1,000	0.22	(0.88)	0.57	(1.55)	0.21	(0.87)	0.71	(1.68)
Earnings Q-3/1,000	0.23	(0.86)	0.60	(1.60)	0.20	(0.87)	0.76	(1.77)
Earnings Q-4/1,000	0.22	(0.87)	0.61	(1.60)	0.26	(1.02)	0.81	(1.88)
Earnings Q-5/1,000	0.20	(0.88)	0.67	(1.70)	0.25	(1.11)	0.83	(1.92)
Earnings Q-6/1,000	0.19	(0.81)	0.70	(1.76)	0.23	(0.89)	0.84	(1.90)
Earnings Q-7/1,000	0.19	(0.81)	0.71	(1.79)	0.26	(1.05)	0.84	(1.95)
Earnings Q-8/1,000	0.18	(0.80)	0.73	(1.84)	0.22	(1.01)	0.83	(1.96)
Earnings Q-9/1,000	0.18	(0.80)	0.72	(1.83)	0.23	(1.00)	0.83	(1.99)
Earnings Q-10/1,000	0.17	(0.74)	0.73	(1.82)	0.24	(1.09)	0.84	(2.01)
Zero Earn Q-1	0.88	(0.33)	0.78	(0.41)	0.86	(0.34)	0.73	(0.44)
Zero Earn Q-2	0.88	(0.33)	0.76	(0.42)	0.86	(0.34)	0.72	(0.45)
Zero Earn Q-3	0.87	(0.33)	0.76	(0.43)	0.86	(0.34)	0.71	(0.45)
Zero Earn Q-4	0.87	(0.33)	0.75	(0.43)	0.86	(0.34)	0.71	(0.45)
Zero Earn Q-5	0.88	(0.32)	0.74	(0.44)	0.86	(0.35)	0.71	(0.46)
Zero Earn Q-6	0.89	(0.31)	0.74	(0.44)	0.86	(0.35)	0.70	(0.46)
Zero Earn Q-7	0.88	(0.33)	0.74	(0.44)	0.87	(0.34)	0.71	(0.45)
Zero Earn Q-8	0.89	(0.32)	0.73	(0.44)	0.87	(0.33)	0.72	(0.45)
Zero Earn Q-9	0.89	(0.31)	0.74	(0.44)	0.87	(0.33)	0.73	(0.45)
Zero Earn Q-10	0.89	(0.31)	0.74	(0.44)	0.87	(0.34)	0.73	(0.44)
Earnings Yr 1/1,000	1.44	(4.08)	2.37	(4.94)	1.44	(4.15)	2.55	(5.31)

Table 2: TESTS FOR ZERO AND CONSTANT AVERAGE TREATMENT EFFECTS FOR GAIN DATA

County	Zero Cond. Ave TE chi-sq	(dof)	normal	Constant Cond. Ave. TE chi-sq	(dof)	normal	Zero Ave. TE chi-sq	(dof)	normal
LA	34.58	(29)	0.73	34.56	(28)	0.88	0.37	(1)	-0.61
RI	248.09	(29)	28.77	171.22	(28)	19.14	72.46	(1)	8.51
AL	46.68	(29)	2.32	46.52	(28)	2.48	0.04	(1)	0.21
SD	97.51	(29)	9.00	88.14	(28)	8.04	3.64	(1)	1.91

For the zero and constant conditional average treatment effect test the chi-sq column is equal to $\sqrt{2K}$ times the normal column plus K, where K is the degrees of freedom. For the column with the zero average treatment effect results the chi-squared colum is equal to the square of the normal column.
Critical values for Chi-squared distribution: $\mathcal{X}^2_{0.95}(1) = 3.84$ $\mathcal{X}^2_{0.99}(1) = 6.63$ $\mathcal{X}^2_{0.95}(28) = 41.34$ $\mathcal{X}^2_{0.99}(28) = 48.28$ $\mathcal{X}^2_{0.95}(29) = 42.56$ $\mathcal{X}^2_{0.99}(29) = 49.59$.

Table 3: SUMMARY STATISTICS EXPERIMENTAL WIN DATA

	Maryland (MD) $N_1 = 524,$ $N_0 = 547$ mean	(s.d.)	Arkansas (AK) $N_1 = 115,$ $N_0 = 128$ mean	(s.d.)	San Diego (SD) $N_1 = 658,$ $N_0 = 646$ mean	(s.d.)	Virginia (VA) $N_1 = 939,$ $N_0 = 428$ mean	(s.d.)
One Child	0.48	(0.50)	0.44	(0.50)	0.47	(0.50)	0.47	(0.50)
High School Dipl	0.40	(0.49)	0.48	(0.50)	0.54	(0.50)	0.44	(0.50)
Never Married	0.36	(0.48)	0.35	(0.48)	0.26	(0.44)	0.29	(0.45)
Non White	0.69	(0.46)	0.85	(0.36)	0.68	(0.47)	0.65	(0.48)
Earnings Q-1/1,000	0.43	(0.89)	0.19	(0.53)	0.41	(1.08)	0.28	(0.75)
Earnings Q-2/1,000	0.44	(0.97)	0.21	(0.58)	0.41	(1.03)	0.29	(0.76)
Earnings Q-3/1,000	0.43	(0.93)	0.18	(0.48)	0.43	(1.08)	0.32	(0.78)
Earnings Q-4/1,000	0.44	(0.98)	0.18	(0.45)	0.41	(1.01)	0.31	(0.75)
Zero Earn Q-1	0.69	(0.46)	0.82	(0.38)	0.75	(0.43)	0.80	(0.40)
Zero Earn Q-2	0.70	(0.46)	0.83	(0.38)	0.74	(0.44)	0.78	(0.42)
Zero Earn Q-3	0.71	(0.45)	0.80	(0.40)	0.73	(0.44)	0.76	(0.43)
Zero Earn Q-4	0.70	(0.46)	0.81	(0.39)	0.73	(0.44)	0.75	(0.43)
Earnings Year 1/1,000	1.65	(3.18)	0.89	(1.93)	2.06	(4.16)	1.50	(2.81)

Table 4: TESTS FOR ZERO AND CONSTANT AVERAGE TREATMENT EFFECTS FOR WIN DATA

County	Zero Cond. Ave TE			Constant Cond. Ave. TE			Zero Ave. TE		
	chi-sq	(dof)	normal	chi-sq	(dof)	normal	chi-sq	(dof)	normal
MD	18.71	(13)	1.12	17.46	(12)	1.11	0.03	(1)	-0.18
AK	27.69	(13)	2.88	27.29	(12)	3.12	0.48	(1)	0.69
SD	25.47	(13)	2.44	19.98	(12)	1.63	4.37	(1)	2.09
VA	27.56	(13)	2.86	27.53	(12)	3.17	0.01	(1)	-0.08

For the zero and constant conditional average treatment effect test the chi-sq column is equal to $sqrt2K$ times the normal column plus K, where K is the degrees of freedom. For the column with the zero average treatment effect results the chi-squared colum is equal to the square of the normal column.
Critical values for Chi-squared distribution: $\mathcal{X}^2_{0.95}(1) = 3.84$ $\mathcal{X}^2_{0.99}(1) = 6.64$ $\mathcal{X}^2_{0.95}(12) = 21.03$ $\mathcal{X}^2_{0.99}(12) = 26.22$ $\mathcal{X}^2_{0.95}(13) = 22.36$ $\mathcal{X}^2_{0.99}(13) = 27.69$.