The Review *of* Economics *and* Statistics

VOL. XC AUGUST 2008 NUMBER 3

NONPARAMETRIC TESTS FOR TREATMENT EFFECT HETEROGENEITY

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

Abstract—In this paper we develop two nonparametric tests of treatment effect heterogeneity. The first test is for the null hypothesis that the treatment has a zero average effect for all subpopulations defined by covariates. The second test is for the null hypothesis that the average effect conditional on the covariates is identical for all subpopulations, that is, that there is no heterogeneity in average treatment effects by covariates. We derive tests that are straightforward to implement and illustrate the use of these tests on data from two sets of experimental evaluations of the effects of welfare-to-work programs.

I. 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 policymakers 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. Firpo (2007), for example, developed an estimator for quantile treatment effect estimates under unconfoundedness.[3]

The hypothesis that the average effect of the treatment is zero for all subpopulations also is important for researchers interested in assessing assumptions concerning selection mechanisms and the appropriateness of particular estimators. For example, Heckman and Hotz (1989) introduced an important class of specification tests that test for a particular type of zero treatment effect in order to eliminate inappropriate estimators of average treatment effects. In particular, Heckman and Hotz proposed testing the null hypothesis of zero causal effects on lagged outcomes using alternative parametric estimators on panel data and discarding those estimators for which this null hypothesis is rejected. While their testing strategy presumed constant treatment effects, their approach clearly 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 effect after averaging over the subpopulations.

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 all subpopulations 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 (for example, Hahn, 1998; Imbens, Newey, & Ridder, 2006; Chen, Hong, & Tarozzi, 2008; Chen, 2007). Given the

Received for publication October 31, 2006. Revision accepted for publication June 12, 2007.

* Department of Economics, University of California, Berkeley; Department of Economics, Duke University, and NBER; Department of Economics, Harvard University, and NBER; and Department of Economics, University of Miami, and IZA, respectively.

We are grateful for helpful comments by Michael Lechner, the co-editor, two anonymous referees, and seminar participants at NYU, the University of Pennsylvania, Johns Hopkins University, the Harvard-MIT econometrics seminar, U.C. Berkeley, UCLA. U.C. Riverside, the World Bank, and for financial support from the National Science Foundation through grants SES 0452590 and SES 0339051.

[1] See Angrist and Krueger (1999), Heckman and Robb (1984), Heckman, LaLonde, and Smith (1999), Rosenbaum (2001), Wooldridge (2002), Imbens (2004), Lechner (2002), and Lee (2005) for surveys of this literature.

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

[3] Also see Lehmann (1974), Doksum (1974), Abadie, Angrist, and Imbens (2002), Chernozhukov and Hansen (2005), and Bitler, Gelbach, and Hoynes (2006) for quantile effects in other settings.

The Review of Economics and Statistics, August 2008, 90(3): 389–405

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 with 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 also is related to the strategy of testing conditional moment restrictions by using an expanding set of marginal moment conditions. See, for example, Bierens (1990) and 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 increase 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 (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 use empirical process methods to test equality of two regression functions, again in the context of a single regressor. Pinkse and Robinson focus on efficient estimation of the nonparametric functions and investigate the efficiency gains from pooling 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 data from two sets of experimental evaluations of the effects of welfare-to-work 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 with 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 not have 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 II, 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 III, 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 IV, 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.

II. Framework and Motivation

A. Setup

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 (see Crump, 2006). The setup we use is standard in the program evaluation literature and based on the potential outcome notation popularized by Rubin (1974). See Angrist and Krueger (1999), Heckman, LaLonde, and Smith (1999), 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 the control treatment and $Y_i(1)$ the outcome under the active 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 pretreatment variables, or covariates, denoted by X_i. The support of the covariates is $\mathbb{X} \subset \mathbb{R}^d$. Define the two conditional means, $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$, the two conditional variances, $\sigma_w^2(x) = \text{Var}\,(Y(w)|X = x)$, 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 $e(x) = \Pr(W = 1|X = x) = \mathbb{E}[W|X = x]$, the marginal treatment probability $\pi_1 = \Pr(W = 1) = \mathbb{E}[e(X)]$, and $\pi_0 = \Pr(W = 0)$.

Assumption 2.1 (Independent and Identically Distributed Random Sample): (Y_i, W_i, X_i), $i = 1, 2, \ldots, N$ *are independent and identically distributed random variables.*

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

Assumption 2.2 (Unconfoundedness):

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

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 impose additional conditions on the two regression functions, $\mu_w(x)$, and the two conditional variance functions, $\sigma_w^2(x)$.

Various estimators have been proposed for the average treatment effect in this setting, such as by Hahn (1998), Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003), Chen, Hong, and Tarozzi (2008), and Abadie and Imbens (2006).

B. 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 is

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

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 0.

The second pair of hypotheses is

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

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,

$$\begin{aligned} &H_0'': \mathbb{E}[\tau(X)] = 0, \\ &H_a'': \mathbb{E}[\tau(X)] \neq 0. \end{aligned} \tag{2.4}$$

Tests of the null hypothesis of a zero average effect are more commonly carried out, either explicitly or implicitly, using the estimated 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 (for example, Eubank & Spiegelman, 1990; Härdle & Marron, 1990; Hong & White, 1995; Horowitz & 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.

C. *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, one may be interested in whether there are subpopulations for which the effect is substantively and statistically significant. For example, one may be interested not only in whether a new drug has a nonzero average effect, but whether it has nonzero effect—positive or negative—for identifiable subgroups in the population in order to better target who should or should not use the drug. As a first step toward establishing such a conclusion, one would like 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 estimate directly 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 multiple-testing 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, such as indicators of racial, educational, or age groups. 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. Although the unconfoundedness assumption is not directly testable, there are types of tests that are suggestive of the plausibility of this assumption. We consider two such tests. The first set of tests was originally suggested by Heckman and Hotz (1989).[4] 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 be the case that the assumption of unconfoundedness conditional on Z is incorrect. Although this is not direct evidence against unconfoundedness conditional on the full set of covariates $X = (V, Z')'$, it is, at the very least, suggestive that unconfoundedness is a delicate assumption with the presence of V essential. Such tests can be particularly effective if the researcher has data on multiple lagged values of the outcome. In that case, one can choose V to be the one-period lagged value of the outcome. 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).[5]

A second strategy is Rosenbaum's (1997) consideration of the use of two or more control groups. Rosenbaum suggests that if potential biases would likely be different for both groups, then evidence that all control groups led to similar estimates is suggestive that unconfoundedness may be appropriate. One can implement this idea by comparing the two control groups. Let T_i be an indicator for the two control groups, $T_i = 0$ for the first and $T_i = 1$ for the second. One can test whether $Y_i(0) \perp\!\!\!\perp T_i|X_i$ in the two control groups. If we find evidence that this pseudo treatment has a systematic effect on the outcome, it must be that unconfoundedness is violated for at least one of the two control groups. 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 III B, we then provide

[4] See also the discussion in Imbens (2004).

[5] An even stronger version of this test would be to test for conditional independence of the treatment indicator and the lagged outcome conditional on the remaining covariates. Such a test could be implemented using methods in Gozalo and Linton (2003). In practice most deviations from the null hypothesis of conditional independence would involve differences in conditional means, so we focus on these differences in the current paper.

conditions under which these tests can be interpreted as nonparametric tests.

III. Testing

A. *Tests in Parametric Models*

We briefly discuss the parametric versions of the tests in (2.2) and (2.3). Such tests are standard in this parametric setting, but they help fix the ideas for how such testing procedures can be extended to nonparametric settings. Suppose the regression functions are specified as

$$\mu_w(x) = \alpha_w + \beta_w' x.$$

We can estimate α_w and $\beta_w \in \mathbb{R}^{K-1}$ using least squares:

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

Under general heteroskedasticity, with the conditional variance of $Y(w)$ equal to $\mathbb{V}(Y(w)|X) = \sigma_w^2(X)$, the normalized asymptotic covariance matrix of $(\hat\alpha_w, \hat\beta_w')'$ is

$$\Omega_w = N_w \cdot \left(\sum_{i|W_i=w} X_i X_i'\right)^{-1} \sum_{i|W_i=w} \sigma_w^2(X_i) X_i X_i' \left(\sum_{i|W_i=w} X_i X_i'\right)^{-1}. \tag{3.6}$$

In large samples,

$$\begin{pmatrix} \sqrt{N_0}\begin{pmatrix}\hat\alpha_0 - \alpha_0 \\ \hat\beta_0 - \beta_0\end{pmatrix} \\ \sqrt{N_1}\begin{pmatrix}\hat\alpha_1 - \alpha_1 \\ \hat\beta_1 - \beta_1\end{pmatrix}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix}\Omega_0 & 0 \\ 0 & \Omega_1\end{pmatrix}\right), \tag{3.7}$$

where N_0, N_1, and N are the sample size for the controls, treated, and the overall sample size, respectively. 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

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

This can be tested using the quadratic form

$$T = \begin{pmatrix}\hat\alpha_0 - \hat\alpha_1 \\ \hat\beta_0 - \hat\beta_1\end{pmatrix}' (\hat\Omega_0/N_0 + \hat\Omega_1/N_1)^{-1} \begin{pmatrix}\hat\alpha_0 - \hat\alpha_1 \\ \hat\beta_0 - \hat\beta_1\end{pmatrix}. \tag{3.8}$$

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

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

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

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

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

$$T' = (\hat\beta_0 - \hat\beta_1)'(\hat\Omega_{0,11}/N_0 + \hat\Omega_{1,11}/N_1)^{-1}(\hat\beta_0 - \hat\beta_1). \tag{3.10}$$

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

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

We now turn to the formulation of nonparametric versions of these tests. The first step toward that goal is the nonparametric estimation of the regression functions $\mu_w(x)$, for $w = 0, 1$.

B. *Nonparametric Estimation of Regression Functions*

In order to develop nonparametric extensions of the tests discussed in section III A, 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 (2008). See Chen (2007) 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 $\lambda = (\lambda_1, \ldots, \lambda_d)$ be a multi-index of dimension d, that is, a d-dimensional vector of nonnegative integers, with norm $|\lambda| = \Sigma_{k=1}^d \lambda_k$, 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 $p_r(x) = x^{\lambda(r)}$, $P_r(x) = (p_1(x), \ldots, p_r(x))'$. The nonparametric series estimator of the regression function $\mu_w(x)$, given K terms in the series, is given by

$$\hat\mu_{w,K}(x) = P_K(x)'\left(\sum_{i|W_i=w} P_K(X_i)P_K(X_i)'\right)^{-} \sum_{i|W_i=w} P_K(X_i)Y_i$$
$$= P_K(x)'\hat\xi_{w,K},$$

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

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

Define the $N_w \times K$ matrix $P_{w,K}$ with rows equal to $P_K(X_i)'$ for units with $W_i = w$, and Y_w to be the $N_w \times 1$ vector with elements equal to Y_i for the same units, so that $\hat\xi_{w,K} = (P_{w,K}'P_{w,K})^-(P_{w,K}'Y_w)$.

In addition, let us define (see assumption 3.2 below)

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

$$\Sigma_{P,w,K} \equiv \mathbb{E}[\sigma^2_w(X)P_K(X)P_K(X)'|W = w].$$

For fixed K, the limiting variance of $\sqrt{N_w}\hat{\xi}_{w,K}$ is $\Omega^{-1}_{P,w,K}\Sigma_{P,w,K}\Omega^{-1}_{P,w,K}$. We estimate this variance by $\hat{\Omega}^{-1}_{P,w,K}\hat{\Sigma}_{P,w,K}\hat{\Omega}^{-1}_{P,w,K}$, where

$$\hat{\Omega}_{P,w,K} = \frac{P'_{w,K}P_{w,K}}{N_w}, \quad \hat{\Sigma}_{P,w,K} = \frac{P'_{w,K}\hat{D}_{w,K}P_{w,K}}{N_w},$$

and

$$\hat{D}_{w,K} = \text{diag}\,\{\mathbf{1}_w(W_i)(Y_i - \hat{\mu}_{w,K}(X_i))^2; i = 1, \ldots, N\},$$

where $\mathbf{1}_w$ (W_i) is an indicator function that takes on the value 1 when $W_i = w$. 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 $s/d > 27/4$.*

(*ii*) *for $\varepsilon_{w,i} \equiv Y_i(w) - \mu_w(x_i)$*

(*a*) $\mathbb{E}[\varepsilon_{w,i}|X = x] = 0$,

(*b*) $\mathbb{E}[\varepsilon^2_{w,i}|X = x] = \sigma^2_w(x)$ *where* $0 < \underline{\sigma}^2 \leq \sigma^2_w(x) \leq \bar{\sigma}^2 < \infty$ *for all* $x \in \mathbb{X}$,

(*c*) $\mathbb{E}[\varepsilon^4_{w,i}] < \infty$.

Assumption 3.3 (Rates for Series Estimators): *The number of terms in the series, K, increases with the sample size N as $K = C \cdot N^\nu$, for an arbitrary positive constant C and some ν such that $d/(2s + 2d) < \nu < 2/13$.*

C. *Nonparametric Tests: Zero Conditional Average Treatment Effect*

In this section, we show how the tests based on parametric regression functions discussed in section III A can be used to test the null hypothesis against the alternative hypothesis given in (2.2) without the parametric model. In essence, we are going to provide conditions under which we can apply a sequence of parametric tests identical to those discussed in section III A 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, that is, for the hypotheses in (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, $\hat{\xi}_{0,K}$ and $\hat{\xi}_{1,K}$. Specifically, we use the normalized quadratic form,

$$T = ((\hat{\xi}_{1,K} - \hat{\xi}_{0,K})'\hat{V}^{-1}_{P,K}(\hat{\xi}_{1,K} - \hat{\xi}_{0,K}) - K)/\sqrt{2K} \qquad (3.12)$$

as the test statistic for the test of the null hypothesis, H_0, where

$$\hat{V}_{P,K} = \hat{\Omega}^{-1}_{P,0,K}\hat{\Sigma}_{P,0,K}\hat{\Omega}^{-1}_{P,0,K}/N_0 + \hat{\Omega}^{-1}_{P,1,K}\hat{\Sigma}_{P,1,K}\hat{\Omega}^{-1}_{P,1,K}/N_1. \qquad (3.13)$$

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, $V^{-1/2}_{P,K} \cdot (\hat{\xi}_{1,K} - \hat{\xi}_{0,K})$, into three parts. Define the pseudo-true values, $\xi^*_{w,K}$, for $w = 0, 1$, $K = 1, 2, \ldots, F$ as

$$\xi^*_{w,K} = \arg\min_{\xi} \mathbb{E}[(\mu(X) - P_K(X)'\xi)^2|W = w]$$
$$= (\mathbb{E}[P_K(X)P_K(X)'|W = w])^{-1}\mathbb{E}[P_K(X)Y|W = w], \qquad (3.14)$$

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

$$V^{-1/2}_{P,K} \cdot (\hat{\xi}_{1,K} - \hat{\xi}_{0,K}) = V^{-1/2}_{P,K} \cdot (\xi^*_{1,K} - \xi^*_{0,K})$$
$$+ V^{-1/2}_{P,K} \cdot (\hat{\xi}_{1,K} - \xi^*_{1,K}) - V^{-1/2}_{P,K} \cdot (\hat{\xi}_{0,K} - \xi^*_{0,K}).$$

For fixed K and in large samples, the last two terms are normally distributed and centered around 0. The asymptotic distribution of T is based on this approximate normality. This approximation ignores the first term, that is, the difference $V^{-1/2}_{P,K} \cdot (\xi^*_{1,K} - \xi^*_{0,K})$. For fixed K, this difference is not equal to 0, even if $\mu_0(x) = \mu_1(x)$ because the covariate distributions differ in the two treatment groups. In large samples, however, we can ignore this difference with large K. 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 $P_K(x)'\xi^*_{w,K}$ for all x. Hence, it follows that, for large enough K, $P_K(x)'V^{-1/2}_{P,K} \cdot (\xi^*_{1,K} - \xi^*_{0,K})$ is close to 0 for all x, implying $V^{-1/2}_{P,K}\xi^*_{0,K}$ and $V^{-1/2}_{P,K}\xi^*_{1,K}$ 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 $V^{-1/2}_{P,K} \cdot (\hat{\xi}_{w,K} - \xi^*_{w,K})$ 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,

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

and compare this with 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. Thus, the modification is a small-sample correction that does not affect the large-sample properties of the test. Interestingly, the modification would make the implementation of the nonparametric test identical to the implementation of the parametric test discussed in section III A, if the parametric model is

$$\mu_w(x) = P_K(x)'\xi_{w,K}. \tag{3.15}$$

This makes the nonparametric tests particularly simple to apply. However, in large samples the nonparametric test does not rely on the correct specification, but instead relies on the increasingly flexible specification as K increases with the sample size. The price one pays for this robustness of the nonparametric tests is in the decreased power of the test. If, in fact, the specification (3.15) is correct for $K = K_0$, the nonparametric version of the test with larger values of K than necessary, that is, $K > K_0$, combines the test of ξ_{K_0} being zero with tests of additional parameters—that is, the elements of ξ_K not in ξ_{K_0}—being equal to zero. The latter are, in fact, zero under the parametric model, and, therefore, this addition lowers the power of finding deviations of ξ_{K_0} from zero.

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

$$\mu_1(x) - \rho_0(x) = \mu_N \cdot \Delta(x),$$

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 *and* 3.1–3.3 *hold. Also suppose 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* $\rho_N^{-1} = O(N^{1/2-2\nu-\epsilon})$ *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).

D. Nonparametric Tests: Constant Conditional Average Treatment Effect

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

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

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

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

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

$$T' = ((\hat{\xi}_{11,K} - \hat{\xi}_{01,K})'\hat{V}_{P,11}^{-1}(\hat{\xi}_{11,K} - \hat{\xi}_{01,K}) - (K-1))/\sqrt{2(K-1)}. \tag{3.16}$$

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 http://www.econ.duke.edu/~vjh3/proofs.pdf.

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

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

and compare this with 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 *and* 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* $\rho_N^{-1} = O(N^{1/2-2\nu-\epsilon})$ *for some* $\epsilon > 0$. *Then* $\Pr(T' \geq M) \to 1$ *for all* M.

Proof: See http://www.econ.duke.edu/~vjh3/proofs.pdf.

Table 1.—Summary Statistics, Experimental GAIN Data

	Los Angeles (LA)		Riverside (RI)		Alameda (AL)		San Diego (SD)	
	$N_1 = 2{,}995$, $N_0 = 1{,}400$		$N_1 = 4{,}405$, $N_0 = 1{,}040$		$N_1 = 597$, $N_0 = 601$		$N_1 = 6{,}978$, $N_0 = 1{,}154$	
	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 year 1/1,000	1.44	(4.08)	2.37	(4.94)	1.44	(4.15)	2.55	(5.31)

IV. 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 reanalyze 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 1990s in several different counties in California.[6] The second set consists of four experimental Work INcentive (WIN) demonstration programs implemented in the mid-1980s in different locations of the United States. The WIN programs also were welfare-to-work programs that examined different strategies for improving the employment and earnings of welfare recipients.[7] 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.

A. *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.

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

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

Table 2.—Tests for Zero and Constant Average Treatment Effects for GAIN Data

	Zero Cond. Ave. TE					Constant Cond. Ave. TE					Zero Ave. TE				
						All Covariates									
County	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val
LA	26.53	(29)	0.60	−0.32	0.63	22.47	(28)	0.76	−0.74	0.23	0.37	(1)	0.54	−0.61	0.54
RI	178.75	(29)	0.00	19.66	0.00	65.92	(28)	0.00	5.07	0.00	72.46	(1)	0.00	8.51	0.00
AL	35.07	(29)	0.20	0.80	0.21	34.01	(28)	0.20	0.80	0.21	0.04	(1)	0.83	0.21	0.83
SD	82.60	(29)	0.00	7.04	0.00	65.78	(28)	0.00	5.05	0.00	3.64	(1)	0.06	1.91	0.06
						Top Down Selection of Covariates									
County	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val
LA	3.71	(6)	0.72	−0.66	0.75	3.68	(5)	0.60	−0.42	0.34	0.37	(1)	0.54	−0.61	0.54
RI	101.56	(5)	0.00	30.54	0.00	8.21	(4)	0.08	1.49	0.07	72.46	(1)	0.00	8.51	0.00
AL	0.30	(2)	0.86	−0.85	0.80	0.15	(1)	0.70	−0.60	0.27	0.04	(1)	0.83	0.21	0.83
SD	29.41	(8)	0.00	5.35	0.00	18.31	(7)	0.01	3.02	0.00	3.64	(1)	0.06	1.91	0.06
						Bottom Up Selection of Covariates									
County	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val
LA	4.28	(7)	0.75	−0.73	0.77	4.21	(6)	0.65	−0.52	0.30	0.37	(1)	0.54	−0.61	0.54
RI	112.47	(7)	0.00	28.19	0.00	13.25	(6)	0.04	2.09	0.02	72.46	(1)	0.00	8.51	0.00
AL	0.96	(4)	0.92	−1.08	0.86	0.74	(3)	0.86	−0.92	0.18	0.04	(1)	0.83	0.21	0.83
SD	42.42	(10)	0.00	7.25	0.00	32.49	(9)	0.00	5.54	0.00	3.64	(1)	0.06	1.91	0.06

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*-sq column is equal to the square of the normal column.

We conducted three specifications of the tests. In the first one, 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 other two versions considered alternative methods of selecting which of the 28 covariates to include in the regressions, using only the data for the control group and then applying the same specification to the treatment group. In the first method, we selected the covariates "top down," that is, starting with the full set of covariates and sequentially (one by one) dropping the covariate with the smallest t-statistic until all remaining covariates had a t-statistic larger than or equal to 2 (in absolute value). In the second method, we selected covariates "bottom up," that is, running K regressions with just an intercept and a covariate and selecting from them the covariate with the highest t-statistic. We then run $K - 1$ regressions with an intercept, the selected covariate, and one of all remaining covariates, selecting again from the potential covariates the one with the highest t-statistic. This process continued until no potential covariate had a t-statistic equal to or above 2 (in absolute value).

The results for the various tests we consider are reported in the three panels of table 2. The top panel corresponds to tests performed using all the covariates described above, and the middle and bottom panels present the results for the tests performed using the covariates selected by the "top down" and "bottom up" methods, respectively. (The degrees of freedom for the *chi*-squared version of the tests are recorded in this table under the "dof" heading and their p-value under the "p-val" 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 1% level for two out of the four of the GAIN counties, Riverside and San Diego. (For all of the tests, we also include the normal distribution–based version of the tests and their p-value.) Comparing the results across the three panels of this table, it is clear that the results are robust to the variables included in the regressions.

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. We reject this null hypothesis at the 5% conventional level for the same two counties when using all the covariates and the "bottom up" selection method, but only for San Diego when using the "top down" selection method. 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 zero 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

Table 3.—Summary Statistics, Experimental WIN Data

	Maryland (MD)		Arkansas (AK)		San Diego (SD)		Virginia (VA)	
	$N_1 = 524$, $N_0 = 547$		$N_1 = 115$, $N_0 = 128$		$N_1 = 658$, $N_0 = 646$		$N_1 = 939$, $N_0 = 428$	
	Mean	(s.d.)	Mean	(s.d.)	Mean	(s.d.)	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)
Nonwhite	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)

other outcomes) in the initial periods after random assignment—is what was reported in the MDRC analysis on this evaluation.[8] It has been widely cited as evidence that the program strategies used in Riverside County's 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.[9] 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 San Diego also benefited from the GAIN treatment. Moreover, there is clear evidence of treatment-effect heterogeneity across subgroups both in San Diego and (at least for two of the specifications) in Riverside.

B. *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 nonwhite. In addition, we have four quarters of earnings data. Table 3 presents summary statistics for the twelve covariates and the outcome variable, annual earnings in the first year after random assignment, for the four locations. Again, we consider three specifications, using all the covariates and applying "top down" and "bottom up" methods for selecting the covariates.

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 two out of the four locations of the WIN experiments (Arkansas and San Diego) at the 5% level. For Arkansas, we also reject the hypothesis of constant treatment-effects level. (The tests are not rejected for Arkansas when using the "bottom up" selection method, though.) In contrast, testing the null hypothesis of a zero average treatment effect results in the rejection of the null hypothesis only for San Diego. 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 at least one of the four locations analyzed in this set of evaluations.

V. 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 nonzero treatment effects and the presence 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 nonzero treatment effects that were missed by testing strategies that focused solely on inferences concerning average treatment effects.

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

[9] 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.

TABLE 4.—TESTS FOR ZERO AND CONSTANT AVERAGE TREATMENT EFFECTS FOR WIN DATA

	Zero Cond. Ave. TE					Constant Cond. Ave. TE					Zero Ave. TE				
	All Covariates														
County	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val
MD	13.76	(13)	0.39	0.15	0.44	12.19	(12)	0.43	0.04	0.48	0.03	(1)	0.85	−0.18	0.85
AK	23.83	(13)	0.03	2.12	0.02	21.37	(12)	0.05	1.91	0.03	0.48	(1)	0.49	0.69	0.49
SD	31.70	(13)	0.00	3.67	0.00	15.86	(12)	0.20	0.79	0.22	4.37	(1)	0.04	2.09	0.04
VA	10.96	(13)	0.61	−0.40	0.66	10.77	(12)	0.55	−0.25	0.40	0.01	(1)	0.94	−0.08	0.94
	Top Down Selection of Covariates														
County	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val
MD	4.41	(4)	0.35	0.15	0.44	1.80	(3)	0.61	−0.49	0.31	0.03	(1)	0.85	−0.18	0.85
AK	12.51	(4)	0.01	3.01	0.00	11.95	(3)	0.01	3.65	0.00	0.48	(1)	0.49	0.69	0.49
SD	21.10	(4)	0.00	6.04	0.00	2.75	(3)	0.43	−0.10	0.46	4.37	(1)	0.04	2.09	0.04
VA	0.80	(3)	0.85	−0.90	0.81	0.47	(2)	0.79	−0.77	0.22	0.01	(1)	0.94	−0.08	0.94
	Bottom Up Selection of Covariates														
County	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val	*chi*-sq	(dof)	*p*-val	normal	*p*-val
MD	5.04	(5)	0.41	0.01	0.50	1.94	(4)	0.75	−0.73	0.23	0.03	(1)	0.85	−0.18	0.85
AK	2.62	(2)	0.27	0.31	0.38	2.02	(1)	0.16	0.72	0.24	0.48	(1)	0.49	0.69	0.49
SD	21.35	(5)	0.00	5.17	0.00	2.99	(4)	0.56	−0.36	0.36	4.37	(1)	0.04	2.09	0.04
VA	0.02	(3)	1.00	−1.22	0.89	0.01	(2)	0.99	−0.99	0.16	0.01	(1)	0.94	−0.08	0.94

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*-sq column is equal to the square of the normal column.

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 et al. (forthcoming) 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.

REFERENCES

Abadie, Alberto, "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association* 97 (March 2002), 284–292.

Abadie, Alberto, Joshua Angrist, and Guido Imbens, "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica* 70 (January 2002), 91–117.

Abadie, Alberto, and Guido Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74 (January 2006), 235–267.

Angrist, Joshua, and Alan Krueger, "Empirical Strategies in Labor Economics" (pp. 1278–1366), in Orley Ashenfelter and David Card (Eds.), *Handbook of Labor Economics,* Vol. 3 (New York: Elsevier Science, 1999).

Bentkus, Vidmantas, "A Lyapunov-type Bound in R^d," *Theory of Probability and Applications* 49 (2005), 311–322.

Bierens, Herman, "Consistent Model Specification Tests," *Journal of Econometrics* 20 (October 1982), 105–134.

——— "A Consistent Conditional Moment Test of Functional Form," *Econometrica* 58 (November 1990), 1443–1458.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes, "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review* 96 (September 2006), 988–1012.

Blundell, Richard, and Monica Costa-Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Economic Journal* 1 (August 2002), 91–115.

Chen, Xiaohong, "Large Sample Sieve Estimation of Semi-Nonparametric Models" (pp. 5549–5632), in James Heckman and Edward Leamer (Eds.), *Handbook of Econometrics,* Vol. 6, Part 2 (New York: Elsevier Science, 2007).

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *The Annals of Statistics* 36 (April 2008), 808–843.

Chernozhukov, Victor, and Christian Hansen, "An IV Model of Quantile Treatment Effects," *Econometrica* 73 (January 2005), 245–261.

Crump, Richard, "Testing Parametric Relationships Between Nonparametric Curves Using Series Estimation," Department of Economics, U.C. Berkeley, unpublished manuscript (2006).

Crump, Richard, V. Joseph Hotz, Guido Imbens, and Oscar Mitnik, "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika* (forthcoming).

De Jong, Robert, and Herman Bierens, "On the Limit of a Chi-Square Type Test if the Number of Conditional Moments Tested Approaches Infinity," *Econometric Theory* 9 (March 1994), 70–90.

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

Eubank, Randall, and Clifford Spiegelman, "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *Journal of the American Statistical Association* 85 (June 1990), 387–392.

Firpo, Sergio, "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica* 75 (January 2007), 259–276.

Gozalo, Pedro, and Oliver Linton, "Conditional Independence Restrictions: Testing and Estimation," London School of Economics, unpublished manuscript (2003).

Gueron, Judith, and Edward Pauly, *From Welfare to Work* (New York: Russell Sage Foundation Press, 1991).

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (March 1998), 315–331.

Härdle, Wolfgang, and Enno Mammen, "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics* 21 (December 1993), 1926–1947.

Härdle, Wolfgang, and James Marron, "Semiparametric Comparison of Regression Curves," *The Annals of Statistics* 18 (March 1990), 63–89.

Heckman, James, and V. Joseph Hotz, "Alternative Methods for Evaluating the Impact of Training Programs" (with discussion), *Journal of the American Statistical Association* 84 (December 1989), 862–874.

Heckman, James, Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65 (April 1998), 261–294.

Heckman, James, Robert LaLonde, and Jeffrey Smith, "The Economics and Econometrics of Active Labor Market Programs" (pp. 1866–2097), in Orley Ashenfelter and David Card (Eds.), *Handbook of Labor Economics,* Vol. 3 (New York: Elsevier Science, 1999).

Heckman, James, and Richard Robb, "Alternative Methods for Evaluating the Impact of Interventions" (pp. 156–245), in James Heckman and Burton Singer (Eds.), *Longitudinal Analysis of Labor Market Data* (Cambridge: Cambridge University Press, 1984).

Hirano, Keisuke, Guido Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71 (July 2003), 1161–1189.

Hong, Yongmiao, and Halbert White, "Consistent Specification Testing Via Nonparametric Series Regression," *Econometrica* 63 (September 1995), 1133–1159.

Horowitz, Joel, and Vladimir Spokoiny, "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica* 69 (May 2001), 599–631.

Hotz, V. Joseph, Guido Imbens, and Jacob Klerman, "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program," *Journal of Labor Economics* 24 (July 2006), 521–566.

Hotz, V. Joseph, Guido Imbens, and Julie Mortimer, "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics* 125 (March–April 2005), 241–270.

Imbens, Guido, "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," this REVIEW 86 (February 2004), 1–29.

Imbens, Guido, Whitney Newey, and Geert Ridder, "Mean-squared-error Calculations for Average Treatment Effects," Department of Economics, Harvard University, unpublished manuscript (2006).

Lechner, Michael, "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society,* Series A 165 (February 2002), 659–682.

Lee, Myoung-jae, *Micro-Econometrics for Policy, Program, and Treatment Effects* (Oxford: Oxford University Press, 2005).

Lehmann, Erich, *Nonparametrics: Statistical Methods Based on Ranks* (San Francisco, CA: Holden-Day, 1974).

Neumeyer, Natalie, and Holger Dette, "Nonparametric Comparison of Regression Curves: An Empirical Process Approach," *The Annals of Statistics* (2003), 880–920.

Newey, Whitney, "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79 (July 1997), 147–168.

Pinkse, Joris, and Peter Robinson, "Pooling Nonparametric Estimates of Regression Functions with a Similar Shape" (pp. 172–197), in G. S. Maddala, Peter C. B. Phillips, and T. N. Srinivisan (Eds.), *Advances in Econometrics and Quantitative Economics: Essays in Honor of C. R. Rao* (New York: Wiley Blackwell, 1995).

Riccio, James, Daniel Friedlander, and Steven Freedman, *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program* (New York: MDRC, 1994).

Rosenbaum, Paul, "The Role of a Second Control Group in an Observational Study" (with discussion), *Statistical Science* 2 (August 1997), 292–316.

——— *Observational Studies,* 2nd ed. (New York: Springer Verlag, 2001).

Rosenbaum, Paul, and Donald Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (April 1983), 41–55.

Rubin, Donald, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66 (October 1974), 688–701.

Shorack, Galen, *Probability for Statisticians* (New York: Springer, 2000).

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

APPENDIX

Instead of working directly with $\xi_{w,K}$ and $\hat{\xi}_{w,K}$ it is useful to work with a normalized version of the parameters and the least squares estimates. Given assumption 3.1, the expectation $\Omega_{P,1,K} = \mathbb{E}[P_K(X)P_K(x)'|W = 1]$ is nonsingular for all K. Hence we may construct a sequence $R_K(x) = \Omega_{P,1,K}^{-1/2}P_K(x)$ 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 functions, $R_K(x)$. Now, the nonparametric series estimator of the regression function $\mu_w(x)$, given K terms in the series, is given by

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

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 $N_w \times 1$ vector with elements equal to Y_i for the same units, so that $\hat{\gamma}_{w,K} = (R'_{w,K}R_{w,K})^{-}(R'_{w,K}Y_w)$. In addition, we also define the following counterparts to the definitions introduced in section III B. Under this normalization, the second-moment matrices are

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

$$\Sigma_{w,K} = \mathbb{E}[\sigma_w^2(X)R_K(X)R_K(X)'|W = w],$$

which we may estimate via

$$\hat{\Omega}_{w,K} = \frac{R'_{w,K}R_{w,K}}{N_w}, \quad \hat{\Sigma}_{w,K} = \frac{R'_{w,K}\hat{D}_{w,K}R_{w,K}}{N_w}.$$

Finally, we have

$$N \cdot V = \frac{1}{\pi_0} \cdot \Omega_{0,K}^{-1}\Sigma_{0,K}\Omega_{0,K}^{-1} + \frac{1}{\pi_1} \cdot \Omega_{1,K}^{-1}\Sigma_{1,K}\Omega_{1,K}^{-1}$$

and

$$\hat{V} = \hat{\Omega}_{0,K}^{-1}\hat{\Sigma}_{0,K}\hat{\Omega}_{0,K}^{-1}/N_0 + \hat{\Omega}_{1,K}^{-1}\hat{\Sigma}_{1,K}\hat{\Omega}_{1,K}^{-1}/N_1.$$

Recall that $\pi_w = \Pr(W = w)$ and observe that under this normalization $\Omega_{1,K} = I_K$. It is crucial to emphasize that this change of basis does not affect the value of either test statistic. Thus, results proved in the appendix under this nonsingular transformation are directly applicable. Finally, we need to define

$$\zeta(K) = \sup_x \|R_K(x)\|,$$

where here and in the following, $\|\cdot\|$ denotes the Euclidean matrix norm, that is, for a matrix A, $\|A\|^2 = \text{tr}\,(A'A)$. Also, define C as a generic positive constant unless specifically stated otherwise.

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*)

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

(*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* $O_p(\zeta(K)K^{1/2}N^{-1/2}) = o_p(1)$. *In addition, suppose assumption 3.2 holds. Then* (*iv*)

$$\|\hat{\Sigma}_{w,K} - \Sigma_{w,K}\| = O_p(\zeta(K)^2K^{3/2}N^{-1}) + O_p(\zeta(K)K^{1/2}K^{-s/d}),$$

(*v*) *the eigenvalues of* $\Sigma_{w,K}$ *are bounded and bounded away from zero, and* (*vi*) *the eigenvalues of* $\hat{\Sigma}_{w,K}$ *are bounded and bounded away from zero in probability if* $O_p(\zeta(K)^2K^{3/2}N^{-1}) + O_p(\zeta(K)K^{1/2}K^{-s/d}) = o_p(1)$.

Proof: See http://www.econ.duke.edu/~vjh3/proofs.pdf.

Lemma A.2: *Suppose assumptions* 2.1–2.3 *and* 3.1–3.2 *hold. Then* (*i*) *the eigenvalues of* $N \cdot V$ *are bounded and bounded away from zero and* (*ii*) *the eigenvalues of* $N \cdot \hat{V}$ *are bounded and bounded away from zero in probability if* $O_p(\zeta(K)^2K^{3/2}N^{-1}) + O_p(\zeta(K)K^{1/2}K^{-s/d}) = o_p(1)$.

Proof: See http://www.econ.duke.edu/~vjh3/proofs.pdf.

In our case $\zeta(K)$ is $O(K)$ (see for example, Newey, 1997) so lemma A.1 implies that under assumption 3.3, $\|\hat{\Omega}_{w,K} - \Omega_{w,K}\| = o_p(1)$ and $\|\hat{\Sigma}_{w,K} - \Sigma_{w,K}\| = o_p(1)$. After applying the nonsingular transformation, the analagous formula to equation (3.14) is

$$\begin{aligned}\gamma^*_{w,K} &\equiv (\mathbb{E}[R_K(X)R_K(X)'|W = w])^{-1}\mathbb{E}[R_K(X)Y|W = w]\\ &= \Omega^{-1}_{w,K}\mathbb{E}[R_K(X)Y|W = w].\end{aligned}$$

Now define

$$\tilde{\gamma}_{w,K} \equiv \gamma^*_{w,K} + \frac{1}{N_w}\,\Omega^{-1}_{w,K}R'_{w,K}\varepsilon_w, \tag{A1}$$

where

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

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

$$\begin{aligned}\sqrt{N_w}\,(\tilde{\gamma}_{w,K} - \gamma^*_{w,K}) &= \frac{\sqrt{N_w}}{N_w}\,\Omega^{-1}_{w,K}R'_{w,K}\varepsilon_w\\ &= \frac{1}{\sqrt{N_w}}\sum_{i=1}^{N}\Omega^{-1}_{w,K}\mathbf{1}_w(W_i)R_K(X_i)\varepsilon_{w,i},\end{aligned} \tag{A2}$$

with

$$\begin{aligned}\mathbb{E}[\Omega^{-1}_{w,K}\mathbf{1}_w(W_i)R_K(X_i)\varepsilon_{w,i}] &= \Omega^{-1}_{w,K}\mathbb{E}[\mathbf{1}_w(W_i)R_K(X_i)\mathbb{E}[\varepsilon_{w,i}|X_i, \mathbf{1}_w(W_i)]]\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\mathbf{1}_w(W_i)R_K(X_i)\mathbb{E}[\varepsilon_{w,i}|X_i]] = 0,\end{aligned}$$

and,

$$\begin{aligned}\mathbb{V}[\Omega^{-1}_{w,K}\mathbf{1}_w(W_i)R_K(X_i)\varepsilon_{w,i}] &= \Omega^{-1}_{w,K}\mathbb{E}[\mathbf{1}_w(W_i)\varepsilon^2_{w,i}R_K(X_i)R_K(X_i)']\Omega^{-1}_{w,K}\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\mathbf{1}_w(W_i)R_K(X_i)R_K(X_i)'\mathbb{E}[\varepsilon^2_{w,i}|X_i, \mathbf{1}_w(W_i)]]\Omega^{-1}_{w,K}\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\mathbf{1}_w(W_i)R_K(X_i)R_K(X_i)'\mathbb{E}[\varepsilon^2_{w,i}|X_i]]\Omega^{-1}_{w,K}\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\mathbf{1}_w(W_i)\sigma^2_w(X_i)R_K(X_i)R_K(X_i)']\Omega^{-1}_{w,K}\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\sigma^2_w(X_i)R_K(X_i)R_K(X_i)'|\mathbf{1}_w(W_i) = 1]\Omega^{-1}_{w,K}\cdot\Pr(\mathbf{1}_w(W_i) = 1)\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\sigma^2_w(X_i)R_K(X_i)R_K(X_i)'|W_i = w]\Omega^{-1}_{w,K}\cdot\Pr(W_i = w))\\ &= \Omega^{-1}_{w,K}\mathbb{E}[\sigma^2_w(X)R_K(X)R_K(X)'|W = w]\Omega^{-1}_{w,K}\cdot\pi_w\\ &= \Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}\cdot\pi_w.\end{aligned}$$

Therefore,

$$\mathbb{V}[\sqrt{N_w}\,(\tilde{\gamma}_{w,K} - \gamma^*_{w,K})] = \frac{1}{N_w}\,N\cdot\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}\cdot\pi_w \;\to\; \Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}, \tag{A3}$$

and

$$\begin{aligned}&\mathbb{V}\left[\frac{\sqrt{N}}{\sqrt{N_w}}\,\sqrt{N_w}\,(\tilde{\gamma}_{w,K} - \gamma^*_{w,K})\right]\\ &\quad= \frac{N}{N_w}\frac{1}{N_w}\,N\cdot\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}\cdot\pi_w \;\to\; \frac{1}{\pi_w}\,\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}.\end{aligned}$$

Moreover,

$$\begin{aligned}S_{w,K} &= [\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}\cdot\pi_w]^{-1/2}\,\frac{1}{\sqrt{N_w}}\sum_{i=1}^{N}\Omega^{-1}_{w,K}\mathbf{1}_w(W_i)R_K(X_i)\varepsilon_{w,i}\\ &= \frac{1}{\sqrt{N_w}}\sum_{i=1}^{N}[\Omega^{-1}_{w,K}\,\Sigma_{w,K}\Omega^{-1}_{w,K}\cdot\pi_w]^{-1/2}\Omega^{-1}_{w,K}\mathbf{1}_w(W_i)R_K(X_i)\varepsilon_{w,i}\\ &\equiv \frac{1}{\sqrt{N_w}}\sum_{i=1}^{N}Z_i\end{aligned}$$

is a normalized summation of N_w independent random vectors distributed with expectation 0 and variance-covariance matrix I_K. Denote the distribution of $S_{w,K}$ by Q_{N_w} and define $\beta_3 \equiv N_w^{-3/2}\sum_{i=1}^{N}E\|Z_i\|^3$. Then, by theorem 1.1, Bentkus (2005),

$$\sup\nolimits_{\mathscr{A}\in A_K}|Q_{N_w}(\mathscr{A}) - \Phi(\mathscr{A})| \le C\beta_3K^{1/4},$$

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* $\nu < 2/13$. *Then,*

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

Proof: First we will show that β_3 is $O(\zeta(K)^3N^{-1/2})$. Consider,

$$\begin{aligned}\beta_3 &= N_w^{-3/2}\sum_{i=1}^{N}\mathbb{E}\|Z_i\|^3\\ &= N_w^{-3/2}\sum_{i=1}^{N}\mathbb{E}\|[\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}\cdot\pi_w]^{-1/2}\times\\ &\qquad\mathbf{1}_w(W_i)\cdot\Omega^{-1}_{w,K}R_K(X_i)\varepsilon_{w,i}\|^3\\ &\le \underline{\sigma}^2\cdot N_w^{-3/2}\sum_{i=1}^{N}\mathbb{E}\|[\Omega^{-1}_{w,K}\cdot\pi_w]^{-1/2}\times\\ &\qquad\mathbf{1}_w(W_i)\cdot\Omega^{-1}_{w,K}R_K(X_i)\varepsilon_{w,i}\|^3\\ &\le C\cdot N^{-3/2}\sum_{i=1}^{N}\mathbb{E}\|\Omega^{-1/2}_{w,K}R_K(X_i)\varepsilon_{w,i}\|^3\\ &\le C\cdot\zeta(K)^3N^{-1/2},\end{aligned}$$

where the last line follows by,

$$\|\Omega_{w,K}^{-1/2}R_K(X_i)\varepsilon_{w,i}\| \leq \lambda_{max}(\Omega_{w,K}^{-1/2})\|R_K(X_i)\varepsilon_{w,i}\| \leq \lambda_{max}(\Omega_{w,K}^{-1/2})\zeta(K)\cdot|\varepsilon_{w,i}|,$$

and so,

$$\mathbb{E}\|\Omega_{w,K}^{-1/2}R_K(X_i)\varepsilon_{w,i}\|^3 \leq \lambda_{max}(\Omega_{w,K}^{-1/2})^3\zeta(K)^3\cdot\mathbb{E}|\varepsilon_{w,i}|^3 \leq C\cdot\zeta(K)^3$$

by lemma A.1 and assumption 3.2. Thus,

$$C\cdot\beta_3 K^{1/4} \leq C\cdot N^{-1/2}\zeta(K)^3K^{1/4} \leq C\cdot N^{-1/2}K^{13/4},$$

which is $o(1)$ for $\nu < 2/13$. ■

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

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

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.

Lemma A.4: *Under assumptions* 2.1–2.3 *and* 3.1–3.3,

$$\sup_c |\Pr(S'_{w,K}S_{w,K} \leq c) - \chi_K^2(c)| \rightarrow 0.$$

Proof: Define the set $A(c) \equiv \{S \in \mathbb{R}^K | S'S \leq c\}$. 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 $\chi_K^2(c) = \Pr(Z'_K Z_K \leq c)$. Then,

$$\begin{aligned}
&\sup_c |\Pr[S'_{w,K}S_{w,K} \leq c] - \chi_K^2(c)| \\
&\quad = \sup_c |\Pr(S'_{w,K}S_{w,K} \leq c) - \Pr(Z'_K Z_K \leq c)| \\
&\quad = \sup_c |\Pr(S_{w,K} \in A(c)) - \Pr(Z_K \in A(c))| \\
&\quad \leq \sup_{\mathcal{A}\in A_K} |Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A})| \leq C\beta_3 K^{1/4} \\
&\quad = O(N^{-1/2}K^{13/4}),
\end{aligned}$$

which is $o(1)$ for $\nu < 2/13$ 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,

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

Proof:

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

The first term goes to zero by lemma A.4. The second term may be rewritten as

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

where Z_K is a $K \times 1$ vector of independent standard normal random variables, with the result following from the Berry-Esséen Theorem.[10] Thus for $\nu > 0$ the right-hand side converges to zero as well and the result is established. ■

The next lemma establishes rates of convergence for the estimators of the regression functions.

Lemma A.6: *Suppose assumptions* 3.1–3.3 *hold. Then,*

(*i*) *there is a sequence of vectors* $\gamma_{w,K}^0$ *such that*

$$\sup_x |\mu_w(x) - R_K(x)'\gamma_{w,K}^0| \equiv \sup_x |\mu_w(x) - \mu_{w,K}^0(x)| = O(K^{-s/d}).$$

In addition, we have the following rates of convergence:

(*ii*)

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

(*iii*)

$$\|\hat{\gamma}_{w,K} - \gamma_{w,K}^0\| = O_p(\zeta(K)KN^{-1}) + O(K^{-s/d}),$$

(*iv*)

$$\begin{aligned}
&\sup_x |R_K(x)'\gamma_{w,K}^* - R_K(x)'\gamma_{w,K}^0| \equiv \sup_x |\mu_{w,K}^*(x) - \mu_{w,K}^0(x)| \\
&\quad = O(\zeta(K)K^{1/2}K^{-s/d}),
\end{aligned}$$

(*v*)

$$\begin{aligned}
&\sup_x |\mu_w(x) - R_K(x)'\hat{\gamma}_{w,K}| \equiv \sup_x |\mu_w(x) - \hat{\mu}_{w,K}(x)| \\
&\quad = O_p(\zeta(K)^2KN^{-1}) + O(\zeta(K)K^{-s/d}).
\end{aligned}$$

Proof: See http://www.econ.duke.edu/~vjh3/proofs.pdf.

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,

$$\begin{aligned}
&(N_w\cdot(\hat{\gamma}_{w,K} - \gamma_{w,K}^*)'[\hat{\Omega}_{w,K}^{-1}\hat{\Sigma}_{w,K}\hat{\Omega}_{w,K}^{-1}]^{-1}(\hat{\gamma}_{w,K} - \gamma_{w,K}^*) - K)/\sqrt{2K} \\
&\quad \xrightarrow{d} \mathcal{N}(0, 1).
\end{aligned}$$

Proof: We need only show that

$$\|[\hat{\Omega}_{w,K}^{-1}\hat{\Sigma}_{w,K}\hat{\Omega}_{w,K}^{-1}]^{-1/2}\sqrt{N_w}\cdot(\hat{\gamma}_{w,K} - \gamma_{w,K}^*) - S_{w,K}\| = o_p(1),$$

then the result follows by lemmas A.3, A.4, and A.5. First, notice that we may write $\hat{\gamma}_{w,K}$ as

$$\hat{\gamma}_{w,K} = \gamma_{w,K}^* + \frac{1}{N_w}\hat{\Omega}_{w,K}^{-1}R'_{w,K}\varepsilon_{w,K}^*,$$

where

[10] See Theorem 1.1 in Shorack (2000).

$$\varepsilon^*_{w,K} \equiv Y_w - R_{w,K}\cdot\gamma^*_{w,K}.$$

Then,

$$\begin{aligned}
&\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\sqrt{N_w}\cdot(\hat\gamma_{w,K}-\gamma^*_{w,K}) - S_{w,K}\| \\
&= N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\varepsilon^*_{w,K} \\
&\quad - [\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \qquad \text{(A4)} \\
&\le N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\varepsilon^*_{w,K} \\
&\quad - [\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \\
&\quad + N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w \\
&\quad - [\hat\Omega^{-1}_{w,K}\ \Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \qquad \text{(A5)} \\
&\quad + N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w \\
&\quad - [\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\|. \qquad \text{(A6)}
\end{aligned}$$

Additional details for the following results may be found at http://www.econ.duke.edu/~vjh3/proofs.pdf. First equation (A4) is

$$\begin{aligned}
&N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\varepsilon^*_{w,K} \\
&\quad - [\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \qquad \text{(A7)} \\
&\le \|[\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\|\,\|\hat\Omega^{-1/2}_{w,K}R'_{w,K}(\varepsilon^*_{w,K}-\varepsilon_w)/N_w^{1/2}\| \\
&= O_p(\zeta(K)^2K^{3/2}K^{-2s/d}N^{1/2}).
\end{aligned}$$

Next, equation (A5) is

$$\begin{aligned}
&N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w \\
&\quad - [\hat\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \qquad \text{(A8)} \\
&\le \|[\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2} - [\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\|\,\|N_w^{-1/2}2\hat\Omega^{-1/2}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \\
&= O_p(\zeta(K)K^{3/2}N^{-1/2}) + O_p(\zeta(K)K^{3/2}K^{-s/d}).
\end{aligned}$$

Finally, equation (A6) is

$$\begin{aligned}
&N_w^{-1/2}\|[\hat\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\hat\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w \\
&\quad -[\Omega^{-1}_{w,K}\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\Omega^{-1}_{w,K}R'_{w,K}\cdot\varepsilon_w\| \qquad \text{(A9)} \\
&\le \|[\Sigma_{w,K}\Omega^{-1}_{w,K}]^{-1/2}\|\,\|\hat\Omega^{-1/2}_{w,K}-\Omega^{-1/2}_{w,K}\|\,\|N_w^{-1/2}R'_{w,K}\cdot\varepsilon_w\| \\
&= O_p(\zeta(K)K^{3/2}N^{-1/2}).
\end{aligned}$$

We may combine equations (A7), (A8), and (A9),

$$\begin{aligned}
&\|[\hat\Omega^{-1}_{w,K}\hat\Sigma_{w,K}\hat\Omega^{-1}_{w,K}]^{-1/2}\sqrt{N_w}\cdot(\hat\gamma_{w,K}-\gamma^*_{w,K}) - S_{w,K}\| \\
&\quad = O_p(\zeta(K)K^{3/2}K^{-s/d}) + O_p(\zeta(K)K^{3/2}N^{-1/2}),
\end{aligned}$$

which is $o_p(1)$ by assumptions 3.2 and 3.3. ■

Proof of theorem 3.1: To simplify notation let us define

$$\hat\delta_K = \hat\gamma_{1,K} - \hat\gamma_{0,K}, \quad \delta^*_K = \gamma^*_{1,K} - \gamma^*_{0,K}.$$

We may follow the logic of lemmas A.3, A.4, and A.5 to conclude that

$$T^* = ((\hat\delta_K - \delta^*_K)'\hat V^{-1}(\hat\delta_K - \delta^*_K) - K)/\sqrt{2K} \xrightarrow{d} \mathcal{N}(0,1).$$

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 $\gamma^0_{1,K} = \gamma^0_{0,K}$ for $\mu^0_{1,K}(x) = \mu^0_{0,K}(x)$. Then,

$$\begin{aligned}
\|\gamma^*_{1,K} - \gamma^*_{0,K}\| &= \|\gamma^*_{1,K} - \gamma^0_{1,K} + \gamma^0_{0,K} - \gamma^*_{0,K}\| \\
&\le \|\gamma^*_{1,K} - \gamma^0_{1,K}\| + \|\gamma^0_{0,K} - \gamma^*_{0,K}\| \qquad \text{(A10)} \\
&= O(K^{1/2}K^{-s/d}),
\end{aligned}$$

by lemma A.6 (*ii*), and

$$\begin{aligned}
\|\hat\gamma_{1,K} - \hat\gamma_{0,K}\| &= \|\hat\gamma_{1,K} - \gamma^0_{1,K} + \gamma^0_{0,K} - \hat\gamma_{0,K}\| \\
&\le \|\hat\gamma_{1,K} - \gamma^0_{1,K}\| + \|\gamma^0_{0,K} - \hat\gamma_{0,K}\| \\
&= O_p(\zeta(K)KN^{-1}) + O(K^{-s/d}), \qquad \text{(A11)}
\end{aligned}$$

by lemma A.6 (*iii*). Thus,

$$\begin{aligned}
|T^* - T| &= |((\hat\delta_K - \delta^*_K)'\hat V^{-1}(\hat\delta_K - \delta^*_K) - K)/\sqrt{2K} \\
&\quad - (\hat\delta_K'\hat V^{-1}\hat\delta_K - K)/\sqrt{2K}| \\
&\le \frac{2}{\sqrt{2K}}\cdot|(\hat\gamma_{1,K} - \hat\gamma_{0,K})'\hat V^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K})| \qquad \text{(A12)} \\
&\quad + \frac{1}{\sqrt{2K}}\cdot|(\gamma^*_{1,K} - \gamma^*_{0,K})'\hat V^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K})|. \qquad \text{(A13)}
\end{aligned}$$

The second factor of equation (A12) is

$$\begin{aligned}
&|(\hat\gamma_{1,K} - \hat\gamma_{0,K})'\hat V^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K})| \\
&\quad = N\cdot|\text{tr}\,((\hat\gamma_{1,K} - \hat\gamma_{0,K})'[N\cdot\hat V]^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K}))| \\
&\quad \le N\cdot\lambda_{max}([N\cdot\hat V]^{-1})\|\hat\gamma_{1,K} - \hat\gamma_{0,K}\|\,\|\gamma^*_{1,K} - \gamma^*_{0,K}\| \\
&\quad = N\cdot[O(1) + O_p(\zeta(K)K^{3/2}N^{-1/2}) + O_p(\zeta(K)K^{3/2}K^{-s/d})] \\
&\qquad \times [O_p(\zeta(K)KN^{-1}) + O(K^{-s/d})][O(K^{1/2}K^{-s/d})] \\
&\quad = O_p(\zeta(K)K^2K^{-3s/d}N).
\end{aligned}$$

Thus, equation (A12) is

$$\frac{2}{\sqrt{2K}}\cdot|(\hat\gamma_{1,K} - \hat\gamma_{0,K})'\hat V^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K})| = O_p(\zeta(K)K^{3/2}K^{-3s/d}N).$$

The second factor of equation (A13) is

$$\begin{aligned}
&|(\gamma^*_{1,K} - \gamma^*_{0,K})'\hat V^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K})| \\
&\quad = N\cdot|\text{tr}\,((\gamma^*_{1,K} - \gamma^*_{0,K})'[N\cdot\hat V]^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K}))| \\
&\quad \le N\cdot\lambda_{max}([N\cdot\hat V]^{-1})\|\gamma^*_{1,K} - \gamma^*_{0,K}\|^2 \\
&\quad = N\cdot[O(1) + O_p(\zeta(K)K^{3/2}N^{-1/2}) + O_p(\zeta(K)K^{3/2}K^{-s/d})] \\
&\qquad \times [O(K^{1/2}K^{-s/d})]^2 \\
&\quad = O_p(\zeta(K)K^{5/2}K^{-2s/d}N^{1/2}).
\end{aligned}$$

Thus, equation (A13) is

$$\frac{1}{\sqrt{2K}}\cdot|(\gamma^*_{1,K} - \gamma^*_{0,K})'\hat V^{-1}(\gamma^*_{1,K} - \gamma^*_{0,K})| = O_p(\zeta(K)K^2K^{-2s/d}N^{1/2}).$$

Thus,

$$|T^* - T| = O_p(\zeta(K)K^{3/2}K^{-3s/d}N) + O_p(\zeta(K)K^2K^{-2s/d}N^{1/2}),$$

which is $o_p(1)$ under assumptions 3.2 and 3.3 and the result follows. ■

Proof of theorem 3.2: Recall that $\zeta(K)$ satisfies $\underline{C} \cdot K < \zeta(K) < \bar{C} \cdot K$ for some $0 < \underline{C}$, $\bar{C} < \infty$. Next, consider,

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

Thus,

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

Because $s/d > 27/4$ by assumption 3.2 and $1/(2s/d + 2) < \nu < 2/13$ by assumption 3.3, and since we may rewrite

$$\|\hat{\gamma}_{w,K} - \gamma^0_{w,K}\| = O_p(\zeta(K)^{1/2}K^{1/2}N^{-1/2}) + O(K^{-s/d})$$

since $K = N^\nu$, then,

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

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

and

$$\begin{aligned}
&\zeta(K)\frac{\|\hat{\gamma}_{0,K} - \gamma^0_{0,K}\|}{\rho_N \cdot C_0} \\
&= O(K) \cdot [O_p(\zeta(K)^{1/2}K^{1/2}N^{-1/2}) + O(K^{-s/d})] \cdot O(N^{1/2-2\nu-\varepsilon}) \\
&= o_p(1), \zeta(K)\frac{\|\hat{\gamma}_{1,K} - \gamma^0_{1,K}\|}{\rho_N \cdot C_0} \\
&= O(K) \cdot [O_p(\zeta(K)^{1/2}K^{1/2}N^{-1/2}) + O(K^{-s/d})] \cdot O(N^{1/2-2\nu-\varepsilon}) \\
&= o_p(1).
\end{aligned}$$

It follows that

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

with probability going to 1 as $N \to \infty$. Thus,

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

with probability going to 1 as $N \to \infty$. Since

$$\begin{aligned}
&N^{1/2}\zeta(K)^{-1/2}K^{-1/2}\zeta(K)^{-1} \cdot \rho_N \cdot C_0 \\
&\quad \ge CN^{1/2}\zeta(K)^{-1/2}K^{-1/2}\zeta(K)^{-1}N^{-1/2+2\nu+\varepsilon} \\
&\quad \ge CN^{\varepsilon},
\end{aligned}$$

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

$$\Pr(N^{1/2}\zeta(K)^{-1/2}K^{-1/2}\|\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\| > M') \to 1. \qquad \text{(A14)}$$

Next, we show that this implies that

$$\Pr\left(\frac{\tilde{C} \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'V^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}} > M\right) \to 1, \qquad \text{(A15)}$$

for an arbitrary constant $\tilde{C} \in \mathbb{R}_{++}$. Denote the minimum and maximum eigenvalues of $[N \cdot V]^{-1}$, $\lambda_{min}([N \cdot V]^{-1})$, and $\lambda_{\max}([N \cdot V]^{-1})$, by $\underline{\lambda}$ and $\bar{\lambda}$, respectively, and note that by lemma A.2 $\underline{\lambda}$ is bounded away from 0 and $\bar{\lambda}$ is bounded.

$$\begin{aligned}
&\Pr\left(\frac{\tilde{C} \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'V^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}} > M\right) \\
&= \Pr\left(\frac{\tilde{C} \cdot N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'[N \cdot V]^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}} > M\right) \\
&= \Pr(\tilde{C} \cdot N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'[N \cdot V]^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) > \sqrt{2}\,MK^{1/2} + K) \\
&\ge \Pr(\underline{\lambda}\tilde{C} \cdot N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) > \sqrt{2}\,MK^{1/2} + K) \\
&= \Pr(N\zeta(K)^{-1}K^{-1} \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) \\
&> (\underline{\lambda}\tilde{C})^{-1}\zeta(K)^{-1}(\sqrt{2}\,MK^{-1/2} + 1)) \\
&= \Pr(N^{1/2}\zeta(K)^{-1/2}K^{-1/2}\|\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\| \\
&> (\underline{\lambda}\tilde{C})^{-1/2}\zeta(K)^{-1/2}(\sqrt{2}\,MK^{-1/2} + 1)^{1/2}).
\end{aligned}$$

Since for any M, for large enough N, we have

$$(\underline{\lambda}\tilde{C})^{-1/2}\zeta(K)^{-1/2}(\sqrt{2}\,MK^{-1/2} + 1)^{1/2} < 2(\underline{\lambda}\tilde{C})^{-1/2},$$

it follows that this probability is for large N bounded from below by the probability

$$\Pr(N^{1/2}\zeta(K)^{-1/2}K^{-1/2}\|\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}\| > 2(\underline{\lambda}\tilde{C})^{-1/2}),$$

which goes to 1 by equation (A14). To conclude we must show that this implies that

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

Let $\hat{\underline{\lambda}} = \lambda_{\min}([N \cdot \hat{V}]^{-1})$ for simplicity of notation. Let A_1 denote the event that $\hat{\underline{\lambda}} > \underline{\lambda}/2$, which satisfies $\Pr(A_1) \to 1$ as $N \to \infty$ by lemmas A.1 and A.2 along with assumptions 3.2 and 3.3. Also define the event A_2,

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

Note that

$$\begin{aligned}
&\Pr\left(\frac{\tilde{C} \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'V^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}} > M\right)\\
&\quad = \Pr\left(\frac{\tilde{C} \cdot N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'[N \cdot V]^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}} > M\right)\\
&\quad \leq \Pr\left(\frac{\bar{\lambda}\tilde{C} \cdot N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}} > M\right),
\end{aligned}$$

which goes to 1 as $N \to \infty$ by equation (A15). Since $\tilde{C}$ was arbitrary we may choose $\tilde{C} = (\underline{\lambda}/2) \cdot \bar{\lambda}^{-1}$ and so $\Pr(A_2) \to 1$ as $N \to \infty$. Thus, $\Pr(A_1 \cap A_2) \to 1$ as $N \to \infty$. Finally, note that the event $A_1 \cap A_2$ implies that

$$\begin{aligned}
T &= \frac{(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'\hat{V}^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}}\\
&= \frac{N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'[N \cdot \hat{V}]^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}}\\
&\geq \frac{\hat{\underline{\lambda}} N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}}\\
&> \frac{(\underline{\lambda}/2)N \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K}{\sqrt{2K}}\\
&> M.
\end{aligned}$$

Hence $\Pr(T > M) \to 1$. ■