# A General Nonparametric Test for Misspecification

by
Ralph Bradley
Robert McClelland*
US Bureau of Labor Statistics

JEL Code: C12, C14, Specification testing, degenerate moment functions

We establish a new consistent, uniformly most powerful test for misspecification. Unlike most previous work, we use a constrained cross-validation scheme for smoothing parameter selection, so that the test does not require arbitrary parameter selections. We address the degeneracy issue through a bootstrap procedure that also allows us to establish the asymptotic distribution of our test statistic. By allowing the use of higher-order kernels without the calculation of higher order derivatives, our test should be simpler to calculate and more powerful than similar tests. Finally, we show that our test can be applied to discrete as well as continuous regressors.

* Please address correspondence to: Robert McClelland, Room 3105, 2 Massachusetts Ave. NE, Washington, DC 20212; (202)-606-6579; x607 FAX (202) 606-7080; McClelland_R@BLS.GOV.
The authors are especially grateful to Anna Sanders for her extraordinarily patient editing. The views expressed in this paper are solely those of the author, and do not necessarily reflect the policy of the Bureau of Labor Statistics(BLS) or the views of other staff members of BLS.

## I. Introduction

Recent literature on misspecification tests has focused on tests that are consistent against all alternatives. For example, Wooldridge (1992), and Lee, Granger and White (1993) develop tests for neglected nonlinearity using the same theory as Bierens (1990). While these tests have a desirable robustness, there are serious implementation issues. Bradley and McClelland (1993) (1994), Hong and White (1993), Stinchcombe and White (1993), Bierens and Ploberger (1994), Whang and Andrews (1993), De Jong and Bierens (1994), Hong (1993), and Horowitz and Härdle (1994) all develop new tests based on the underlying theory of Bierens (1990), and that address these issues.[1]

To obtain consistency against all misspecifications, these tests necessarily use nonparametric estimators such as kernel estimators. As a consequence, a smoothing parameter must be selected for each of these tests. Unfortunately, selection of this parameter affects the finite sample size and power of the statistic and only Bradley and McClelland (1993) (1994) offer any guidance in its selection in finite samples.

Most of the above tests are not only consistent against all alternatives, they are also in the class of the uniformly most powerful tests (UPT). The UPT, established in Bradley and McClelland (1993) and Stinchcombe and White (1992), has as its first moment the expectation of the product of the regression residual and the conditional expectation of the residual given the regressors. Unfortunately, because the

---

[1] Horowitz and Härdle (1994) is mentioned for completeness. However, the test in this paper is not based on a underlying statistic that is largest under misspecification and consistent against all alternatives. To remove problems of dimensionality when there is more than one regressor, Härdle and Horowitz use the misspecified functional form as the argument to the kernel regression function. This feature does not allow the test to be consistent against all alternatives and therefore is not in the class of tests that we discuss in this paper.

conditional expectation of the residual equals zero everywhere in the domain under the null hypothesis, tests of its correlation with other variables degenerate to a constant.

Attempts to correct this problem fall into two dissimilar categories. Wooldridge (1992) is an example of the first category in which he places restrictions on the expansion of his series estimator to keep the statistic from degenerating. In the second category Hong and White prevent degeneration by standardizing the sample moment of their test at a rate greater than root-n.[2] It is now widely recognized that the second category of tests have higher power than the first. What has not been generally recognized is that this occurs partly because the first solution forces the noncentrality parameter of the tests to converge to a form that is not the UPT.

In a later paper, Hong (1993) designs a second category test that allows the smoothing parameter to converge at an "optimal" rate so that the nonparametric estimator achieves the optimal convergence in terms of the integrated mean square error (IMSE) criterion for second order kernel functions. Although further reduction of the IMSE is possible through the use of higher order kernels, the use of the second derivative of the nonparametric components in his test would require extremely messy and tedious estimation of higher order derivatives.

A final issue that has received little attention involves the type of data sets upon which these tests may be used. With the exception of De Jong (1991), who extends the Bierens (1990) test to time series data, these tests are restricted by the common assumptions that all variables are independently and identically distributed and continuous.

In this paper we establish a new consistent test for misspecification whose non centrality parameter converges to the UPT under misspecification. Unlike most other tests, it establishes an automatic mechanism to select the smoothing parameter. By using a cross-validation scheme, the test makes no

---

[2] This is the rate used for the sample moments for parametric statistical tests. See Davidson and McKinnon (1993) page 112-113.

demands of the researcher, such as arbitrarily selecting a smoothing parameter. Rather than use either of the two previous approaches to the degeneracy issue, we employ a bootstrap procedure that uses variation in the resampling to prevent degeneration of the test statistic. Finally, we show that our test can be applied to discrete as well as continuous regressors.

## II. Notation, Definitions and the Null Hypothesis

Suppose the random vector (y,x), $y \in Y \subset \Re$ and $x \in X \subset \Re^k$ on the probability space $(Y \times X, \mathcal{J}, F)$, has a joint probability distribution p(y,x) where the following holds:

$$E(y|x) = f(x) \tag{2.1}$$

$$y = f(x) + \varepsilon \tag{2.2}$$

where $\varepsilon \in \Re$ is the error of the model. It is clear that $E(\varepsilon|x) = 0$ for all x and $E(\varepsilon)=0$. In parametric regression estimation, it is assumed that f(x) falls in a family of known real parametric functions $f(x,\theta)$ on $\Re^k \times \Theta$ where $\Theta$ is the parameter space and is a compact subset of $\Re^q$. We wish to test that a specific parameterization of f(x), denoted $f(x,\theta)$, satisfies the null hypothesis:

$$H_o: \Pr(E(y|x)=f(x,\theta_0))=1 \text{ for some } \theta_0 \in \Theta. \tag{2.3}$$

The alternative hypothesis is that

$$H_A: \Pr(E(y|x)=f(x,\theta)) < 1 \text{ for all } \theta \in \Theta \tag{2.4}$$

In general, there are an infinite variety of ways that the alternative hypothesis in (2.4) can hold.

Given a consistent estimator for $\theta$, that varies with sample of size n, which we denote as $\hat{\theta}_n$, a moment-based test, or "m-test", of the null hypothesis in (2.3) can be constructed. These m-tests are based on a sample of observations from $Y \times X$. If the null in (2.3) is true at $\theta=\theta_o$ then using a sample of size n, there is a moment function:

$$m_n: Y \times X \times \Theta \times \Pi \rightarrow \Re^1 \tag{2.5}$$

such that:

$$E(m_n(y_i, x_i, \theta_o, \pi))=0 \tag{2.6}$$

for all observations i and some parameters $\theta_0$ in $\Theta$ and some infinite dimensional nuisance parameter $\pi$ in $\Pi$. A statistical test is *consistent* against all alternatives if (2.6) does not hold whenever (2.4) is true. Given this moment function (which we can also write as $m_{in}(\theta,\pi)$), one may use its sample analog to construct a test of the form $M_n = n\hat{m}_n \hat{V}_n \hat{m}_n$ ,where:

$$\hat{m}_n \equiv n^{-1}\sum_{i=1}^{n} m_{in}(\hat{\theta}_n, \hat{\pi}_n) \tag{2.7}$$

$\hat{\theta}_n$ is a consistent estimate of the true value of nuisance parameter $\pi_0$ and $\hat{V}_n$ is an estimator that asymptotically converges to the variance of $a_n\hat{m}_n$, for some sequence $a_n=o(n^{\delta})$, $.5 \leq \delta < 1$. Under certain regularity conditions (see Newey 1994), $M_n$ converges in distribution to a $\chi^2(1)$ that is invariant to the nuisance parameter $\pi$. Note, however, that these tests still generally require some nonparametric estimate of $\pi$.

The form of the most powerful moment tests is established in Bradley and McCelland (1993) and Stinchcombe and White (1992), where it is shown that the test must use a moment function that contains $[f(x_i) - f(x_i, \theta)][y_i - f(x_i, \theta)]$. The term in the first bracket is the conditional expectation of $y_i - f(x_i, \theta)$. To be consistent against all alternatives, the test must nonparametrically estimate this expectation. Tests of this type belong to the class of UPTs but vary in aspects of the nonparametric estimator . Letting $f(\cdot)$ be defined by (2.1) and (2.2) above and letting $f(\cdot,\theta)$ be a specific parameterized version of $f(\cdot)$, some examples of $a_n$ degenerate moment functions that are the components for most UPTs : are listed below.

$$m_{in}=[f(x_i)-f(x_i, \theta)][y_i - f(x_i, \theta)] \tag{2.8a}$$

$$m_{in}=[y_i - f(x_i)]^2 - [y_i - f(x_i, \theta)]^2 \quad \text{b)}$$

The tests in Wooldridge (1992), Hong and White (1993), Horowitz and Härdle (1994), Zheng (1990), and Hong (1993) are based on (2.8a). In constrast, Yatchew (1992) and Whang and Andrews (1991) develop tests based on (2.8b).

The difficulty behind designing a test of this form is that the moment function is necessarily a *degenerate moment function*, i.e. $m_{in}(\theta_0, \pi_0)=0$, under the null. This degeneracy implies that the distribution of the test statistic converges to a point, making the statistic useless. A more formal definition is as follows:

**Definition 2.1** Let $m_{in}$: $Y\times X\times\Theta\times\Pi\rightarrow\Re^1$ be generated on an i.i.d. sample from $(Y \times X)$. Suppose that plim $(\hat{\theta}-\theta)=0$ and that plim $\rho(\pi_n,\pi_0) = 0$ for some suitable metric $\rho(.,.)$ and for $\pi_0\in\Pi$. For each $(\theta,\pi) \in \Theta \times \Pi$ assume that $m_{in}(\theta,\pi)$ is measurable, and is twice Frechet differentiable with respect to $\pi$, and twice differentiable with respect to $\theta$. Let $a_n$ be a nonstochastic sequence $\{a_n \in \Re^+ : a_n n^{-1/2} \rightarrow \infty$ $a_n n^{-1} \rightarrow 0\}$. Denote $\bar{m}_n(\theta,\pi)=En^{-1}\sum_{i=1}^{n} m_{in}(\theta,\pi)$. The moment function $m_n$ is *$a_n$ degenerate* at $(\theta,\pi)$ if

(a) $a_n[m_n(\theta,\pi)-\bar{m}_n(\theta,\pi)]\xrightarrow{p}0$

(b) $a_n n^{-1/2}\nabla_\theta \bar{m}_n(\theta,\pi)\longrightarrow 0$

(c) $a_n\delta\bar{m}_n(\hat{\pi}-\pi_0;\theta,\pi_0)\longrightarrow 0$ for $\pi_0\in\Pi_0\subseteq\Pi$

where $\Pi_0$ is some neighborhood of $\pi_0$, $\nabla_\theta$ is the gradient with respect to $\theta$, and $\delta$ is the Frechet derivative operator.

Initially, the tests in the papers listed above attempted to impose restrictions that avoided the degeneracy. These test are either U-statistics or von Mise statistics, and the restrictions are usually placed on an expansion of these statistics. Earlier tests such as Wooldridge(1992) are typically *first order* tests in that they are based on the asymptotic theory of a first order expansion of $m_{in}$. When using first order testing, one places limits on the convergence of the bias of the nonparametric estimator so that its variance converges to zero faster than the bias. This avoids the degeneracy under the null because the smoothing parameter converges at an adequately slow rate so that the bias is always present and nonzero. In these tests, we say that the bias *dominates* the variance. The tests in both Wooldridge (1992) and Yatchew (1992) are examples of this type: Wooldridge (1992) uses a sequence of non nested alternatives with an appropriately slow asymptotic growth of the smoothing parameters, while Yatchew (1992) suggests a sample splitting procedure where one part of the sample is used to calculate the nonparametric estimator of the conditional expectation and then this estimator is multiplied by the estimated residuals of the other part of the sample. Although the tests in these papers do not degenerate, there are problems with these approaches. For example, non-nested testing requires slow convergence of the non parametric estimator,

sample splitting makes inefficient use of the data, and weighting requires the choice of new arbitrary parameters.

The major drawback of allowing the bias to dominate the variance is that if $m_{in}(\theta,\pi)$ is of the form in 2.8a) then the limit of the actual estimated statistic will not be in the class of UPTs. Although the limit on the convergence of the bias prevents degeneracy, by construction the presence of the bias of the nonparametric estimator does not allow the first moment of $\hat{m}_n$ in (2.7) to converge to the first moment of the UPT. (See Bierens 1987 for the convergence properties of the kernel regression.).

An alternative approach is that of Hong and White (1993), DeJong and Bierens (1994), and Horowitz and Härdle (1994), who use the central limit theorems (CLTs) of degenerate U-Statistics (in, for example, de Jong 1987 and Hall 1984) to exploit the degeneracy of the moment function. Most often, they use a standardization that is greater than root-n, and established regularity conditions where possible. Typically, these tests are *second order* tests in that the asymptotic theory is established on the second order expansion of $m_{in}$ that exploits the first order degeneracy of (b) and (c) in definition (2.1). Instead of controlling the convergence of the bias of the nonparametric estimator, they control the convergence of the variance. In these tests, we say that the variance *dominates* the bias. Unlike the first order tests, these second order tests do converge to a statistic in the UPT class.

In a variation of this second approach, Hong (1993) imposes the same rate of convergence on both the variance and the bias. This is a distinct advantage over previous tests because it allows him to use the "optimal" rate of convergence (in the IMSE sense) for the window width. Unfortunately, Hong's test is difficult to implement. Because the squared bias and the variance converge at the same rate, the denominator of his test must contain a variance component as well as a bias component. This bias component requires the estimation of second derivatives of the nonparametric component of the test so that if higher order kernels are to be used in order to improve the rate of convergence, messy higher order

derivatives must then be calculated. Finally, by not allowing the bias to disappear, Hong's test does not converge to a UPT.

## III. A Root-n Consistent Test

Our test is a second order test based on the first type of degenerate moment function in 2.8. Instead of using a greater than root-n standardization, we use a bootstrap procedure that forces the variance to dominate the bias by using the sample itself as an additional source of variation. This allows the test to achieve root-n consistency. Unlike Hong's Test, we can use higher order kernels without changing the structure of the denominator of the test. Therefore, we can readily use higher order kernels to achieve a convergence of the window width that produces a lower integrated mean squared error than the Hong test. Finally, by using a moment function of the form in (2.8a) and allowing the bias to disappear, our test does converge to an element of the UPT.

We begin with definitions and assumptions about the data generating process (DGP):

*Definition 3.1*

Let $\nu$ be an element the class of continuous and r differentiable functions For $s \leq r$ the supremum Sobolev Norm of $\nu$ is defined as:

$$\|\nu\|_{s,\infty} = \max_{|\lambda| \leq s} \sup_{z \in Z} \left| D^{\lambda} \nu(z) \right|,$$

where D is the differentiation operation with respect to the argument.

*Definition 3.2*

A Sobolev Space is defined as

$$W^{s}_{\infty,r}(Z) = \left\{ \nu \in C^{r}[Z] : \|\nu\|_{s,\infty} < \infty \right\}, \; s \leq r .$$

*Assumption 3.1*

$(Y\times X,\mathcal{J},F)$ is a complete probability space. (a) The stochastic process $\{(y_i, x_i): Y\times X \to \Re^{k+1}$ for i = 1,2,..,n; n=1,2...} is generated from this space, and for each n, $(y_i,x_i)$ is i.i.d. (b) The support of $x_i$, $X\subset\Re^k$, is compact. (c) $\sup_{x\in X} E|y_i|^{2+\delta} \ |x_i=x) < \infty$.

The compactness of X is used to avoid boundary issues. This assumption can be bypassed by including a trimming function on our test. We can also extend our results to martingales or to $\alpha$ or $\phi$ mixing variables, but this distracts from the main ideas of this paper.

Throughout the paper, we assume that when $H_o$ is true we can consistently estimate the true $\theta_o$. This standard assumption is the next assumption.

*Assumption 3.2*

(a) Under $H_o$, $n^{1/2}[\hat{\theta}-\theta_o] = O_p(1)$.

(b) The function $f(x,\theta) \in W_{\infty,2}{}^s$.

We also need to impose the following conditions on the kernel $K(\cdot)$.

*Assumption (3.3)*

$K:T\to \Re$ is a symmetric bounded kernel with compact support, where $T \equiv [-1,1]^k$ and K is differentiable of order s, with the s-order being Lipschitz. Finally, for $u \in \Re^k$

$$\int_T K(u)du = 1$$

and there exists some $t > 2k$ such that

$$\int_T u_1^{i_1} u_2^{i_2} \dots u_k^{i_k} K(u)du = 0 \ \text{ for } |i| = \sum_{i=1}^{k} i_j < t$$

$$\int_T u^{i_1} u^{i_2} \dots u^{i_k} K(u)du \neq 0 \ \text{ for } |i| = t .$$

Assumption 3.3 allows us to induce the bias to converge to zero more rapidly than the variance so that the variance of our estimator dominates the bias. We can therefore ignore the calculation of the bias in the

denominator of our test statistic. Note that we have not mentioned the window width, $\gamma_n$, that is the smoothing parameter in our test.

Below is an example of a kernel that implements assumption (3.3). We let $k : \Re \rightarrow \Re$ and the K(.) that satisfies (3.3) has the following form:

$$K(u) = \prod_{i=1}^{k} k(u_i) \tag{3.1}$$

$$k(v) = \sum_{j=1}^{1/2(t-2)} \alpha_j v^{2j} \left(\frac{3}{4}\right)(1 - v^2) I(|v| \leq 1) \tag{3.2}$$

The $\alpha$'s are derived as a solution to the following simultaneous system of (1/2)(t-2) equations:

$$\sum_{j=0}^{1/2(t-2)} \alpha_j E(v^{2(i+j)}) = \delta_{i0}, \quad 0 \leq i \leq (1/2)(t-2) \tag{3.3}$$

The kernel in this example is an Epanechnikov kernel, and is the most efficient kernel in the sense that it minimizes the integrated mean squared error for the kernel regression estimator.

*Assumption(3.4)*

Let $p_o(x)$ be the probability density function (pdf) for the random vector for the continuous variables in x. Then $p_o(x) \in W_{\infty,t}^{\infty}$ where t is defined in assumption(3.3).

By assuming that the probability density function (pdf) of x is differentiable t times, where t > 2k, Assumptions (3.3) and (3.4) imposes a convergence rate of the bias that goes to zero more rapidly than the variance. As in most of the nonparametric literature, the proofs for the reduction in bias involve taking higher order derivatives for the pdfs of the continuous variables.

As described in the introduction, we use a bootstrap procedure to prevent the degeneration of the test statistic. To do this we generate for each observation i in a given sample $y_i$ ,$x_i$ , i= 1,...,n, a new random vector of size n' by sampling with replacement from the set of integers {1,2,..,n}. We denote this random vector for the ith observation as $N_i$. We then define the cardinal variable S(A) as the number of occurrences of event A. Therefore, $S(j \in N_i)$ is the number of times j occurs in the random vector $N_i$. The

random vectors $N_1, N_2, \ldots N_n$ along with the operator $S(\cdot)$ are the components of the bootstrapping process. To prevent degeneration, we make the following assumption:

*Assumption (3.5)*

$n' = O(n^{3/2}\gamma_n^{k/2})$

This assumption establishes the relationship between the cardinality of $N_i$, n', and the size of the sample, n. The rate of increase of n' guarantees that our statistic is $O_p(1)$ by offsetting the rate the underlying moment of our test converges to zero when based on a kernel regression using the smoothing parameter, $\gamma_n$ . The assumptions of the convergence of $\gamma_n$ are as follows:

*Assumption (3.6)*

The window width $\gamma_n$ satisfies:

(a) $\gamma_n \rightarrow 0$

(b) $n\gamma_n^{3k} \rightarrow \infty$

(c) $n\gamma_n^{2t+k/2} \rightarrow 0$.

Assumption (3.6a) is a standard window width assumption for kernel regressions. Assumption (3.6b) ensures that the probability limit of the variance estimator of our underlying test is invariant to the underlying nuisance parameters. It is also important in terms establishing the essential condition for Hall's (1984) CLT for degenerate statistics. Finally, assumptions (3.6c) along with (3.5) allow us to construct a test such under $H_A$, the bias disappears in the asymptotic distribution. Thus assumptions (3.5) and (3.6) are essential ingredients for our test to converge to a UPT.

Given the above assumptions, we can now describe the major components of our test. The moment function of interest is

$$m_{in} = p(x_i)[f(x_i) - f(x_i, \theta)][y_i - f(x_i, \theta)] \qquad (3.4)$$

Our test is then based on estimating (3.4) with:

$$\hat{m}_n = n^{-1}\sum_{i=1}^{n}[\hat{r}_n(x_i) - \hat{p}_n(x_i)f(x_i,\hat{\theta}_n)][y_i - f(x_i,\hat{\theta}_n)] \tag{3.5}$$

where we use a "bootstrapped" Nadaraya-Watson kernel estimator for:

$$\hat{f}_n(x_i) = [n\hat{p}_n(x_i)]^{-1}\sum_{j=1}^{n} y_j S(j \in N_i)K_n(x_j - x_i) \text{ if } \hat{p}_n \neq 0$$
$$\hat{p}_n(x_i) = \frac{1}{n}\sum_{j=1}^{n} S(j \in N_i)K_n(x_j - x_i) \tag{3.6}$$

$$\hat{r}_n(x_i) = \hat{p}_n(x_i)\hat{f}_n(x_i)$$

The estimator $\hat{p}_n(x)$ is the kernel estimator for the product of the bootstrap indicator times the probability density of x. The function $K_n(x\text{-}x_j) = \gamma_n^{-k} K(\gamma_n^{-1}[x\text{-}x_j])$, where $\gamma_n$ is the window width of the kernel regression estimator. The parametric estimator of the conditional expectation is denoted as $f(x, \hat{\theta})$ since the only component that is estimated is the parameter vector $\theta$.

The only remaining detail is an assumption that establishes an automatic mechanism to select the window width.

*Assumption (3.7)*

Let $\hat{\sigma} = \{\hat{\sigma}_{x1},..,\hat{\sigma}_{xk}\}$ where $\hat{\sigma}_{xi}$ is the sample standard deviation of the ith element in the random vector. x Assume that the kernel function $K_n(\cdot)$ automatically standardizes each element in x by dividing by its counterpart in $\hat{\sigma}$.

Define R to be a set of int(log(n)) points equally spaced between .025 and 4. Then, $\gamma_n = c*n^{\delta}$ where $\delta$ satisfies the restrictions in assumption (3.6) and

$$c^* = \underset{c \in R}{\operatorname{argmin}} \left(\sum_{i=1}^{n}(y_i - \hat{f}_{-i}(x_i,c))^2\right),$$

where $\hat{f}_{-i}(x_i,c)$ is the kernel regression that omits the ith observation and uses the window width $\gamma_n = cn^{\delta}$. The residuals from the cross validation of assumption (3.7) are:

$$\hat{\varepsilon}_i^* = y_i - \hat{f}_{-i}(x_i, c^*) \tag{3.7}$$

$$\hat{f}_{-i}(x_i, c^*) = \frac{\sum_{j\neq 1}^{n} y_j K\left(\frac{x_i - x_j}{c^* n^{\delta}}\right)}{\sum_{j\neq 1}^{n} K\left(\frac{x_i - x_j}{c^* n^{\delta}}\right)}$$

The purpose of assumption (3.7) is to allow finite sample power improvements while still ensuring that $\hat{m}_n$ will have a zero expectation under the null. We cannot do an unconstrained cross validation because this would force the rate of convergence of $\gamma_n$ to violate assumption (3.6). The bounds of R must be O(1) so that we can maintain the necessary convergence of the kernel. When assumption (3.7) is combined with the bounds on T in assumption (3.3), the kernel is allowed to move from the point where positive weight is given to differences in $x_i$-$x_j$ that are $8(n^{\delta})$ sample standard deviations apart to the point where there are $.05(n^{\delta})$ sample standard deviations apart. Notice that $8\cdot n^{\delta} - 0.5\cdot n^{\delta} = O(1)\cdot n^{\delta}$. Any other points of $x_i$-$x_j$ that are greater than $8(n^{\delta})$ sample standard deviations are not weighted. This addresses the problems of higher bias at the boundary that is discussed in Härdle (1990) pages 130-132, since no outliers of $x_i$-$x_j$ will be used.

Our test statistic is :

$$M_n = \hat{V}_n^{-1/2}\sqrt{n}[\hat{m}_n - \hat{R}_n] \tag{3.8}$$

where

$$\hat{R}_n = \gamma_n^{-k} K(0)\hat{s}_n^2 \tag{3.9}$$

$$\hat{s}_n^2 = 1/n\sum_{i=1}^{n} S(i \in N_i)\hat{\varepsilon}_i^{*2} \tag{3.10}$$

$$\hat{V}_n = 4C(K)n^{-2}\sum\sum_{i<j}[\hat{\varepsilon}^{*2}_i\ \hat{\varepsilon}^{*2}_j\ K_n(x_i - x_j)] \tag{3.11}$$

$$C(K) = \int K^2(u)du \tag{3.12}$$

$$\hat{\varepsilon}_i = y_i - \hat{f}_n(x_i) \tag{3.13}$$

Nothing in this test is left for the researcher to decide. The simple cross validation mechanism is conducted over a compact finite set who boundary asymptotically approaches fixed values.

## IV. Asymptotic Properties of the Test

As we noted in section III, the test as delineated in (3.8) is based on a degenerate statistic. We use the CLTs from Hall (1984) and DeJong (1987) to prove the asymptotic normality of (3.8). Unlike, Hong and White (1993), we do not increase the variance by over standardizing, but instead use a bootstrap procedure. The random variable $m_{in}$ is the basis for our test and is linear in $\pi_i=\{p(x_i)f(x_i), p(x_i)\}$. This linearity simplifies the asymptotic proofs. The moment function in equation (3.1) is degenerate under the null of correct specification, i.e., when $\theta$ equals the true value $\theta_o$. The sample estimate for (3.1) is (3.3), where $\hat{r}_n(x_i)$ and $\hat{p}_n(x_i)$ are used to estimate $r_0(x_i)$ and $p_0(x_i)$.

We now establish a lemma that justifies the use of the bootstrapping, and that characterizes the limiting distribution of our statistic.

*Lemma 4.1*

Suppose Assumptions 3.1, 3.5, and 3.6 hold. Let $\varepsilon$ come from a zero mean distribution with finite variance. Let

$$W_n = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{nij}, \tag{4.1}$$

$$W_{nij} = \varepsilon_i\varepsilon_j S(j \in N_i)K_n(x_i - x_j),$$

$$\varepsilon_i = y_i - E(y_i|x_i)$$

Define

$$V_0 = 2C(K)E\left(\sigma^4(x)p(x)\right), \tag{4.2}$$

where

$$C(K) = \int K^2(u)du.$$

Then

$$V_0^{-1/2}\sqrt{n}(W_n - EW_n) \xrightarrow{d} N(0,1).$$

The bootstrap procedure adds enough additional variation so that even though $W_n$ is based on a degenerate statistic without the bootstrap, it is $O_p(1)$ with the bootstrap. We use the Hall (1984) CLT for degenerate U-Statistics in order to prove asymptotic normality.

We next develop a heteroskedastic consistent estimator for $V_0$.

*Lemma 4.2*
Given assumption 3.1 with $\delta$=2, while assumptions 3.2, 3.3, and 3.5 also hold.

Then, $\hat{V}_n - V_0 = o_p(1)$.

With Lemma 4.1 and 4.2, we can prove that the distribution of our statistic converges to a standard normal distribution under the null hypothesis..

*Theorem 4.1*
Suppose assumptions 3.1 through 3.7, then

$M_n \xrightarrow{d} N(0,1)$

under $H_0$ .

We now focus on the distribution of $M_n$ under global and local alternative hypotheses. We define a sequence $H_n$ under the local alternative $H_{an}$

$H_{an}$: $E(y|x) = H_n(x,\theta) = f(x,\theta_0) + n^{-1/2}\gamma_n^{-k/4}\Delta_n(x)$ (4.3)

where $\Delta_n(x)$ is a sequence of uniformly bounded functions that converges uniformly to a limit function $\Delta(x)$. This is different from $n^{-1/2}$ convergence between the true and alternative under the parametric alternatives of an m-test as outlined in Newey (1985).

We need to make additional assumptions in deriving the local asymptotic power of our test.

*Assumption (3.8)*
Let $\hat{\theta}$ be an estimator for $\theta_0$ in (4.3). There is a $\bar{\theta}$ such that the following holds

$$[\hat{\theta} - \bar{\theta}] = O(n^{-1/2})$$

and

$$[\theta_0 - \bar{\theta}] = O(n^{-1/2}\gamma_n^{-k/4})$$

*Assumption (3.9)*
The random variable, u, in (2.) has finite fourth moments.

*Theorem 4.2*
Let assumptions (3.1) to (3.9) hold. Define

$$\Delta^*(x) = \Delta(x) - E(\partial f(x,\theta)/\partial\theta)\beta$$

$$\beta = \text{plim } n^{1/2}\gamma_n^{k/4}(\bar{\theta}_n - \theta_0)$$

Under the sequence of local alternative models in (4.3) $M_n$ is asymptotically distributed as $N(\mu, V_0)$ where $\mu = E[\Delta^*(x)]$. Under $H_A$ for any nonstochastic sequence $\{C_n = o(n\gamma_n^{k/2})\}$

$$P[M_n > C_n] \longrightarrow 1.$$

From Theorem 4.2, we can conclude that $M_n$ has power against alternatives whose distance from $H_o$ is $O(n^{-1/2}\gamma_n^{-k/4})$. Notice that $\mu$ equals plim $E([y_i - f(x_i,\bar{\theta}_n)][f(x_i) - f(x_i,\bar{\theta}_n)])$ which is exactly the first moment of the UPT. However, there is a curse of dimensionality. As more regressors are added the order of local alternatives that can be detected become slower. This problem can be readily improved by using higher order kernels, and the rate of convergence can be made arbitrarily close to $n^{-1/2}$. Unlike Hong (1993), using higher-order kernels is straightforward.

In many applications there are discrete right hand side variables. We now relax our assumptions to include both discrete and continuous right hand side variables. Let x be partitioned into $[x_1, x_2]$ where $x_1$ are the $k_1$ continuous variables and $x_2$ are the $k_2$ discrete variables where $k = k_1 + k_2$. We now add the following assumptions.

*Assumption 3.9*

The kernel K(.,.): $\Re^{k1}\times\Re^{k2} \to \Re$ is chosen such that for $z_1,z_2$ element $\Re^{k1}\times\Re^{k2}$

(a) $K(z_1,0)$ satisfies all the assumptions of (3.3) in terms of $z_1$.

(b) $\sqrt{n} \sup_{|z_2|> \lambda/\gamma_n} \int K(z_1,z_2)dz_1 \to 0$ for all $\lambda > 0$

Assumption 3.9 places appropriate constraints on the tails of $K(z_1,.)$ so that $E(K(z_1,z_2)|z_2)$ converges to the indicator function $I(z=z_2)$.

*Assumption 3.10*

$x_2$ has support $X_2$ which is a subset of $\Re^{k2}$. $X_2$ has the following additional properties:

(a) $X_2$ has a finite number of elements where $x_2 \in X_2$ implies that $p(x2) > 0$.

(b) $\sum_{x_2\in X_2} p(x_2) = 1$

Let $p(x_1|x_2)$ be the density of $x_1$ given $x_2$. Then the following holds:

(c) $p(x_1|x_2) \in W_{t,\infty}^{s}$ ,

and $E(y|x_1,x_2)P(x_1|x_2) \in W_{t,\infty}^{s}$ with respect to the first argument $x_1$.

The assumption (3.1) holds for $X=X_1\times X_2$ where $X_1$ is the support for the continuous variables.

Assumptions (3.5) and (3.6) hold for $k=k_1$.

Perhaps the easiest way to describe the behavior of the test when discrete right hand side variables are used is to show some primitive results when there are no continuous variables. In this case, assumption (3.9) has a kernel such that $K(0) = 1$.

Letting $I_n(x_i=x_j)=I(x_i=x_j)/\gamma_n$ we define

$$\hat{\hat{r}}_n(x_i) = \frac{1}{n}\sum_{j=1}^{n} y_j I_n(x_j = x_i) S(j \in N_i) \qquad (4.4)$$

$$\hat{\hat{p}}_n(x_i) = \frac{1}{n}\sum_{j=1}^{n} I_n(x_j = x_i) S(j \in N_i) \tag{4.5}$$

$$\hat{\hat{m}}_n = \frac{1}{n}\sum_{i=1}^{n} (\hat{\hat{r}}_n(x_i - \hat{\hat{p}}_n(x_i) f(x_i, \hat{\theta}_n))(y_i - f(x_i, \hat{\theta}_n)) \tag{4.6}$$

Then the following holds:

$$\sqrt{n}\left(\hat{\hat{m}}_n - \hat{m}_n\right) = \frac{1}{\sqrt{n}}\frac{1}{n}\sum_{j=1}^{n} (f(x_j) - f(x_i) + u_j - [f(x_i) - f(x_i, \hat{\theta}_n)]) S(j \in N_i) \times$$

$$\left\{I_n(x_i = x_j) - K_n(x_i - x_j)\right\}\hat{u}_i$$

$$\leq \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \left|(f(x_j) - f(x_i) + u_j - [f(x_i, \hat{\theta}_n) - f(x_i)])\hat{u}_i K_n(x_i - x_j) S(j \in N_i)\right| I_n(x_i \neq x_j)$$

$$\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} Z_{ij} S(j \in N_i) \sup_{|z| > \mu(x_i)/\gamma_n} \sqrt{n} K(z)$$

where $\mu(x_i) = \inf_{x_j \in X_j / x_i} |x_i - x_j|$ and $Z_{ij} = \left|(f(x_j) - f(x_i) + u_j - [f(x_i, \hat{\theta}_n) - f(x_i)])\hat{u}_i\right|$.

By assumption (3.10)

$$\sup_{|z| > \mu(x_i)/\gamma_n} \sqrt{n} K(z) \longrightarrow 0 .$$

Then, $p\lim \sqrt{n}(\hat{\hat{m}}_n - \hat{m}_n) = 0.$

The statistic $\hat{m}_n$ is based on a continuous kernel that maintains all the smoothness properties in assumption (3.3) and therefore, we can easily have both discrete and continuous right hand side variables. Notice that because the support $X_2$ is finite, it will induce the test for pure discrete right hand to converge in probability to its underlying first moment rather than in distribution. The reason for this is that the kernel needs only to be estimated at a finite set of points in the sample space and these points do not grow with sample size.

*Theorem 4.3*

Under the assumptions 3.1 through 3.8 for $x_1$ and $k=k_1$, and the additional assumptions of 3.9 and 3.10. Theorem 4.2 carries over with k replaced by $k_1$ and $C(K) = \int K(u_1, u_2)^2 du_1$ where $u_1 \in \Re^{k_1}$ corresponds to the continuous variables.

We end this section by presenting an alternative estimator for $V_0$ in (4.2). When higher order kernels are used the computation of $C(K) = \int K^2(u)du$ is difficult when there are several right hand side regressors. We offer the following lemma to compute an consistent estimator for $V_0$ that requires no integration.

*Lemma 4.3*

The estimator:

$$\hat{J}_n = \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i<j} \hat{\varepsilon}_i^{*2} \hat{\varepsilon}_j^{*2} K_n^2(x_i - x_j)$$

where $\varepsilon_n^*$ is defined in (3.7) is a consistent estimator for $V_0$.

## V Conclusions

We have in this paper proposed a test for misspecification that is consistent against all functional forms and is of the form of the uniformly most powerful tests. By design the researcher does not arbitrarily choose any parameter of the test. The researcher need only to estimate his original parametric model, and then using the estimated residuals and the left and right hand side variables. The cross validation mechanism along with the kernel of compact support allows the data to determine an optimal window width while trimming out observations with density that are arbitrarily close to zero. The bounds of the cross validation are fixed so that the window width will still have the required rate of decrease in order to achieve standard asymptotic results.

It is important to note that our test prevents degeneracy by using a resampling technique. Although the asymptotic distribution of our bootstrap test is identical to the Hong and White (1993), this research opens a new avenue in which to promote additional finite sample efficiency. In order to maintain simplicity, our resampling was done so that each observation had a 1/n chance of being chosen. This automatically made the random variable $S(j \in N_i)$ independent from the sample, allowing us to concentrate on showing how resampling could prevent the degeneracy of our statistic. However, there are other resampling strategies. One is to let $p_i = 1/\sqrt{\hat{\sigma}_n^2(x_i)}$ where e $\hat{\sigma}_n^2(x_i)$ is an estimator for the residual variance. One

could make resampling section be proportional related to $p_i$ so that the sample points with smaller variance would have a higher chance of being chosen. This could increase the finite sample power of the test.

While this test is still subject to the curse of dimensionality it can be offset by the use of higher order kernels. Although the test is not optimal in the sense of Hong (1993) we can achieve a higher local power by using higher order kernels which at this point cannot be done in Hong without estimated third order and higher derivatives of the conditional expectation function. Finally, we show that our test can be applied to discrete variables, allowing the test to be used on a large number of data sets.

## References


Bierens, H.J., 1982, Consistent Model Specification Tests, Journal of Econometrics 20, 105-134.

————————, 1987, Kernel Estimators of Regression Functions, in: T. R. Bewley, ed., Advances in Econometrics / Fifth World Congress, Vol. 1 (Cambridge University Press: New York), 99-144.

————————, 1990, A Consistent Conditional Moment Test of Functional Form, Econometrica 58, 1443-1458.

Bierens, H.J., and W. Ploberger, 1994, Asymptotic Power of the Integrated Conditional Moment Test Against Large Local Alternatives.

Bradley, R. and R. McClelland, 1993, An Improved Nonparametric Test for Misspecification of Functional Form, Manuscript

Bradley, R. and R. McClelland, 1994, A Kernel Test for Neglected Nonlinearity, manuscript

de Jong, P. 1987, A Central Limit Theorem for Generalized Quadratic Forms, Probability Theory and Related Fields 75, 261-277.

————————— and H.J. Bierens, 1994, On the Limit Behavior of the Chi-Square Test if the Number of Conditional Moments Tested Approaches Infinity, Econometric Theory 9, 70-90.

de Jong, R.M. 1991, The Bierens Test Under Data Dependence, manuscript

Hall, P., 1984, Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators, Journal of Multivariate Analysis 14, 1-16.

Härdle, W., 1990, Applied Nonparametric Regression, (Cambridge University Press, New York).

Hong, Y., 1993, Consistent Specification Testing Using Optimal Nonparametric Kernel Estimation, CAE Working Paper #93-13, Center for Analytic Economics, Cornell University.

————————— and H. White, 1993, M-Testing Using Finite and Infinite Dimensional Parameters Estimators, manuscript, University of California San Diego.

Horowitz, J. and W. Härdle, 1994, Testing a Parametric Model Against a Semiparametric Alternative, Econometric Theory 10, 821-848.

Lee, T.H., White, H. and C.W.J. Granger, 1993, Testing for Neglected Nonlinearity in Time Series Models, Journal of Econometrics 56, 269-290.

Newey, W. K., 1985, Maximum Likelihood Specification Testing and Conditional Moment Tests, Econometrica 53, 1047-1070.

Newey, W, 1994, Kernel Estimatioin of Partial Means a a General Variance Estimator, Econometric Theory 10, 233-253

Stinchcombe, M. and H. White, 1993, An Approach to Consistent Specification Testing Using Duality and Banach Limit Theory, Discussion Paper, University of California, San Diego.

Whang, Y.J. and D. Andrews, 1993, Tests of Specification for Parametric and Semiparametric Models, Journal of Econometrics 57, 277-318.

Wooldridge, J., 1992, A Test for Functional Form Against Nonparametric Alternatives, Econometric Theory 8, 452-475.

Yatchew, A.J., 1992, Nonparametric Regression Tests Based on Least Squares, Econometric Theory 8, 435-451

Zheng, X., 1990, A Consistent Test of Functional Form Via Nonparametric Estimation Techniques, Princeton University Discussion Paper.

## Proofs

We start with Lemma A.1 which is an important ingredient for the lemmas and theorems in the paper.

***Lemma A.1***

Given assumptions 3.1 through 3.6, let N be continuously differentiable and $E||N^2(x)||<\infty$ and $\varepsilon_i$ is defined in (4.1). Then

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_i N(x_j)K_n(x_i - x_j)$$

is $Op(1/\sqrt{n})$.

Proof:

We rewrite the expression as a V-Statistic:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{i}\{\varepsilon_i N(x_j)+\varepsilon_j N(x_i)\}K_n(x_i - x_j).$$

We apply the theorem of Powell, Stock, and Stocker(1989), and we need to verify that

$$P_n = E\left\|\{\varepsilon_i N(x_j)+\varepsilon_j N(x_i)\}K_n(x_i - x_j)\right\|^2 = o(n).$$

$$P_n \le E\left\{[\varepsilon_i^2 N(x_j^2)+\varepsilon_j^2 N(x_i^2)]K_n^2(x_i - x_j)\right\}$$

$$= 2\int\int \frac{1}{\gamma_n^{2k}}K^2\left(\frac{x_i - x_j}{\gamma_n}\right)\sigma^2(x_i)N^2(x_j)p(x_i)p(x_j)dx_i dx_j$$

$$= 2\int\int \frac{1}{\gamma_n^{2k}}K(u)\sigma^2(x_i)N^2(x_i - \gamma_n u)p(x_i)p(x_i - \gamma_n u)dx_i\gamma_n^k du$$

$$= O(\gamma_n^{-k}) = O(n(n\gamma_n^k)^{-1}) = o(n)$$

QED

***Lemma 4.1***

Proof:

Under $H_o$, $E(W_{nij}|x_j) = E(W_{nij}|x_i)=0$. Given Assumption (3.1), the following holds:

$$E(W_n) = En^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{nij} = n^{-2}\sum_{i=1}^{n} E\varepsilon_i^2 S(i \in N_i)\gamma_n^{-k}K(0) = \sigma^2(n\gamma_n^k)^{-1}K(0)(n'/n) \quad .$$

Using A(3.5), this implies:

$$W_n - E(W_n) = n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} W_{nij} + (n\gamma_n^k)^{-1}n^{-1}K(0)\sum_{i=1}^{n}\{\varepsilon_i^2 S(i \in N_i) - (n'/\ n)\sigma^2\} = n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} W_{nij} + Op([n\gamma_n^{k/2}]^{-1}).$$

The last equality follows from Chebyshev's inequality, and $\delta$=2 from Assumption (3.1) and

$S(i\in N_i)=O_p(n'/n)=Op(n^{1/2}\gamma_n^{k/2})$ from A(3.5).

Then

$$\sqrt{n}(W_n - EW_n) = (n^{-3/2})\sum_{i=1}^{n}\sum_{j\neq i}^{n} W_{nij} + o_p(1)$$

given $n\gamma_n^k \to \infty$. To show that $\sqrt{n}(W_n - EW_n) \to N(0,1)$ in distribution, we show that:

$$V_n^{-1/2}U_n \xrightarrow{d} N(0,1), \text{where} \quad U_n = n^{-3/2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} W_{nij}, \text{ and } V_n = \text{var}(U_n).$$

Under $H_o$, because $E(U_n|x_i) = E(U_n|x_j) =0$, $U_n$ is a degenerate second order U-Statistic so that we require a central limit theorem (CLT). From de Jong's (1987) CLT for generalized forms we know that for $V_n^{-1/2}U_n$ to be asymptotically N(0,1), it suffices that $G_{ni}/V_n^2 = o(1)$ for i = 1,2,3, where :

$$\begin{aligned}
V_n = \text{Var}(U_n) &= \{n(n-1)2n^{-3}E(S^2(j \in N_i)\varepsilon_i^2\varepsilon_j^2(K_n^{ij})^2)\} \\
&= 2\frac{n(n-1)}{n^3}[n'(\frac{1}{n}\frac{n-1}{n}) + \frac{n'^2}{n^2}]\int_X\int_X \gamma_n^{-2k}K^2(\frac{x_1 - x_2}{\gamma_n})\sigma^2(x_1)\sigma^2(x_2)p(x_1)p(x_2)dx_1dx_2 \\
&= 2\gamma_n^k\int_X\int_X \gamma_n^{-2k}K^2(\frac{x_1 - x_2}{\gamma_n})\sigma^2(x_1)\sigma^2(x_2)p(x_1)p(x_2)dx_1dx_2 + o(1), \text{ by Assumption (3.5)} \\
&= \int_X\{\int_T K^2(u)\sigma^2(x_1 + \gamma_n u)p(x_1 + \gamma_n u)du\}\sigma^2(x_1)p(x_1)dx_1 + o(1), \text{ T is defined in Assumption (3.3).} \\
&= 2C(K)\int_X \sigma^4(x_1)p^2(x_1)dx_1 + o(1) = 2C(K)E\{\sigma^2(x)p(x)\} + o(1) = V_0 + o(1)
\end{aligned}$$

Therefore, $V_n - V_0 = o(1)$. Define $K_{nij}$ as $K((x_i - x_j)/\gamma_n)$

$$G_{n1} = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} E(W_{nij}/n^{3/2})^4 = \{(n-1)n8n^{-6}E(S^4(j \in N_i))E\{\varepsilon_1^4\varepsilon_2^4(K_{n12})^4\} \leq n^{-3}(n-1)\gamma_n^{2k}cE(K_{n12})^4$$

$$= n^{-3}(n-1)\gamma_n^{2k} c \int_X \int \{\gamma_n^{-k}(K(\frac{x_1 - x_2}{\gamma_n})\}^4 p(x_1)p(x_2)dx_1 dx_2$$

$$= n^{-3}(n-1)\gamma_n^{-k} c \int_T \int_X K(v)^4 p(x_1)p(x_1 + \gamma_n v)dx_1 dv$$

$$= n^{-3}(n-1)\gamma_n^{-k} c \int_V K(v)^4 dv \int_X p^2(x_1)dx_1\{1 + o(1)\} = O(n^{-2}\gamma_n^k)$$

$$G_{n2} = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{n} E\{W_{nij}^2 W_{nik}^2 / n^6\} \leq \{(n-1)(n-2)(n-3)16n^{-6}ES_{12}^2 ES_{13}^2 E\varepsilon_1^2\varepsilon_2^2(K_{n12})^2\varepsilon_1^2\varepsilon_3^2(K_{n13})^2\}$$

$$= O(n^{-1})\gamma_n^{2k} E\{(K_{n12})^2(K_{n13})^2\} = O(n^{-1})$$

$$G_{n3} = \sum\sum\sum\sum_{i\neq j\neq k\neq h}\{E(W_{nij}W_{nik}W_{nhk}W_{nkh}) + E(W_{nij}W_{nih}W_{njk}W_{nkh}) + E(W_{nik}W_{nih}W_{njk}W_{njh})\}n^{-6}$$

$$\leq 3CE\{S(1 \in N_2)S(1 \in N_3)S(4 \in N_2)S(4 \in N_3)\}\frac{n(n-1)(n-2)(n-3)}{n^6} E\left\{\varepsilon_1^2\varepsilon_2^2\varepsilon_3^2\varepsilon_4^2 K_{n12}K_{n13}K_{n42}K_{n43}\right\}$$

by A(3.5)

$$\leq 3O(1)\gamma_n^{2k}\int_X\int_X\int_X\int\{\gamma_n^{-k}K((x_3 - x_1)/\gamma_n)\}\{\gamma_n^{-k}K((x_3 - x_1)/\gamma_n)\}\{\gamma_n^{-k}K((x_2 - x_4/\gamma_n\}\{\gamma_n^{-k}K((x_3 - x_4/\gamma_n\}$$

$$= 3O(1)\gamma_n^k\int_X\int_T\int_T\int_T K(v)K(v+w)K(-r-w)K(-w)p(x_1)p(x_1+\gamma_n v)p(x_1+\gamma_n v+\gamma_n w)p(x_1+\gamma_n v+\gamma_n w+\gamma_n r)$$

$(x_2 = x_1 + \gamma_n v,\ x_3 = x_2 + \gamma_n w,\ x_4 = x_3 + \gamma_n r)$

$= O(\gamma_n{}^k)$

Since $V_n$ is O(1), we have shown that $G_{ni}/V_n{}^2 = o(1)$ for i=1,2,3 given $n\gamma_n{}^k \to \infty$ and $\gamma_n{}^k \to 0$.

QED

***Lemma 4.2***

To simplify, we set $\hat{\varepsilon}_i^* = \hat{\varepsilon}_i$ where $\hat{\varepsilon}_i^*$ is defined in (3.8).

Let $\hat{J}_n = 2n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} K_n\hat{\varepsilon}_i^2\hat{\varepsilon}_j^2$, then $\hat{V}_n = 2C(K)\hat{J}_n$.

Since $\hat{\varepsilon}_i = Y_i - \hat{f}(x_i, c^*)$ and $Y_i = f(x_i) + \varepsilon_i$ where $\hat{f}_n(x_i, c^*)$ is defined in assumption (3.7).

$\hat{J}_n = 2n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} K_n(x_i - x_j)\varepsilon_i^2\varepsilon_j^2$

$$+8n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K_n(x_i-x_j)\varepsilon_i\varepsilon_j^2\left[f(x_i)-\hat{f}(x_j,c^*)\right]$$

$$+4n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K_n(x_i-x_j)\varepsilon_j^2\left[f(x_i)-\hat{f}(x_i,c^*)\right]^2$$

$$+8n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K_n(x_i-x_j)\varepsilon_j\varepsilon_i\left[f(x_j)-\hat{f}(x_j,c^*)\right]\left[f(x_i)-\hat{f}(x_i,c^*)\right]$$

$$+8n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K_n(x_i-x_j)\varepsilon_j\left[f(x_j)-\hat{f}(x_j,c^*)\right]\left[f(x_i)-\hat{f}(x_i,c^*)\right]$$

$$+2n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K_n(x_i-x_j)\varepsilon_j\left[f(x_j)-\hat{f}(x_j,c^*)\right]^2\left[f(x_i)-\hat{f}(x_i,c^*)\right]^2$$

$$=2n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\varepsilon_i^2\varepsilon_j^2K_n(x_i-x_j)+8A_{1n}+4A_{2n}+8A_{3n}+8A_{4n}+2A_{5n}$$

We show that $A_{in}=o_p(1)$ $\quad i=1,\ldots,5$

Let $P=\sup_u K(u)$

$$A_{1n}\le P\gamma_n^k n^{-2}\left[\sum_{i=1}^{n}|\varepsilon_i|\right]\left[\sum_{j=1}^{n}\varepsilon_j^2\right]\left[\sum_{j=1}^{n}\left|f(x_j)-\hat{f}(x_j,c^*)\right|\right]$$

$$\le O_p(\gamma_n^k)O_p(1)O_p\left(n^{-1/2}\gamma^{-k/2}+\gamma_n^t\right)$$

$$\le o_p(1)\qquad \text{because } n\gamma_n^k\to\infty, \text{ and } \left[\sum_{j=1}^{n}\varepsilon_j^2\right]n^{-1/2}=\text{Op}(1)$$

$$A_{2n}\le P\gamma_n^k n^{-2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\varepsilon_j^2\left[f(x_i)-\hat{f}(x_i,c^*)\right]^2$$

$$\le P\gamma_n^k\left[n^{-1}\sum_{j=1}^{n}\varepsilon_j^2\right]\left[n^{-1}\sum_{i=1}^{n}\left[f(x_i)-\hat{f}(x_i,c^*)\right]^2\right]$$

$$\le O_p(\gamma_n^k)\left(O_p(n^{-1/2})\left[O_p(n\gamma_n^k)^{-1}+O_p(\gamma_n^{2t})\right]\right)=o_p(1).$$

The third inequality follows from results in Bierens (1987) that $\text{var}\left(f(x_i)-\hat{f}(x_i,c^*)\right)=O(n\gamma_n^k)^{-1}$ and the bias $f(x_i)-\hat{f}(x_i,c^*)=O(\gamma_n^t)$.

$$A_{3n} \le P\gamma_n^k n^{-1}\sum_{i=1}^{n}\left|\varepsilon_i\left[f(x_i)-\hat{f}(x_i,c^*)\right]\right| n^{-1}\sum_{j=1}^{n}\left|\varepsilon_j\left[f(x_i)-\hat{f}(x_j,c^*)\right]\right|$$

$$\le P\gamma_n^k O_p(n^{-1/2}\gamma_n^{-k/2}) \; + \; O_p\,(n^{-1/2}\gamma_n^t)$$

$$= op(1)$$

$$A_{4n} \le P\gamma_n^k n^{-2}\sum_{i=1}^{n}\left|\varepsilon_i\left[f(x_i)-\hat{f}(x_i,c^*)\right]\right|\sum_{j=1}^{n}\left|f(x_j)-\hat{f}(x_j,c^*)\right|$$

$$\le P\gamma_n^k\left[n^{-1}\sum_{i=1}^{n}\varepsilon_i^2\right]^{1/2} n^{-1}\sum_{i=1}\left[\left|f(x_i)-\hat{f}(x_i,c^*)\right|\right]^2$$

$$\le O_p(\gamma_n^k)O_p\left(\left[n\gamma_n^k\right]^{-3/2}\right) = o_p(1)$$

and

$$A_{5n} \le P\gamma_n^k\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\varepsilon_j(f(x_j)-f(x_j,c^*)^2\right\}\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{i-1}(f(x_i)-f(x_i,c^*))^2\right\}$$

$$= O(\gamma_n^k)O_p(n^{-1}\gamma_n^{-2k}) = o_p(1) \text{ by Assumption(3.6b)}$$

We then show that

$$\text{p lim } 2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_i^2\varepsilon_j^2 K_n(x_i-x_j) = 2E\left(\sigma^4(x)p(x)\right)$$

Let $\mu_4 = E(\varepsilon_j^4)$. Note that

$$E\left(\varepsilon_i^4\varepsilon_j^4 K_n^2(x_i-x_j)\right) = \iint \mu_4(x_i)\mu_4(x_j)K_n^2(x_i-x_j)dx_i, dx_j$$

$$= O\left(\gamma_n^{-k}\right) = o(n), \text{ since } n\gamma_n^k \to \infty$$

Therefore the first term in the expansion of $\hat{J}_n$ satisfies the conditions of the U Statistics Projection

Theorem in Powell, Stock, and Stocker(1989), and therefore.

$$\text{p lim } 2n^{-2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\varepsilon_i^2\varepsilon_j^2 K_n(x_i-x_j) \to 2C(K)E\left(p(x)\sigma^4 p(x)\right) \text{ since}$$

$$E\left(\varepsilon_i^2\varepsilon_j^2 K_n(x_i-x_j)\right) = \int_x\int_x \sigma^2(x_i)\sigma^2(x_2)K_n(x_1-x_2)p(x_1)p(x_2)dx_1, dx_2$$

$$= \int_x\int_T \sigma^2(x_1)\sigma^2(x_1-\gamma_n u)K(u)p(x_1)p(x_1+\mu)dx_1, d\mu$$

$$= \int_T\int_x \sigma^2(x_1)\sigma^2(x_1)p(x_1)p(x_1)dxK(u)du + o(1)$$

$$= \left[C(K)E\left(\sigma^4(x)p(x)\right) + o(1)\right]$$

QED

***Theorem 4.1***

We prove this theorem by Lemmas 4.1, 4.2

.

We must show that $\sqrt{n}\left(\hat{m}_n - W_n\right) = Op(1)$ under $H_0$. Also, under $H_0$ (f(x) in 2.1 is equal to $F(x, \theta_0)$).

$$\hat{m}_n = n^{-1}\sum_{i=1}^{n}\left[\hat{r}_n(x_i) - \hat{p}_n(x_i)f(x_i,\hat{\theta}_n)\right]\left[y_i - f(x_i,\hat{\theta}_n)\right]$$

$$= n^{-2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\left[y_j - f(x_i,\hat{\theta}_n)\right]K_n(x_i - x_j)S_{ij}\right]\hat{u}_i$$

$$= n^{-2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\left[f(x_j - \theta_0) + \varepsilon_j - f(x_i,\hat{\theta}_n)\right]K_n(x_i - x_j)S_{ij}\right]\left[\varepsilon_j - \left[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)\right]\right]$$

$$= n^{-2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\left[\varepsilon_j + f(x_j,\theta_0) - f(x_i,\theta_o) - \left[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)\right]\right]K_n(x_i - x_j)S_{ij}\right]\left[\varepsilon_i - \left[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)\right]\right]$$

$$= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_j S_{ij}K_n(x_i - x_j)\varepsilon_i$$

$$+n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_i K_n(x_i - x_j)S_{ij}\left[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)\right]$$

$$+n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[f(x_j,\theta_0) - f(x_i,\theta_0) - \left[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)\right]\right]K_n(x_i - x_j)S_{ij} \bullet$$

$$\left[\varepsilon_i - \left[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)\right]\right]$$

$$= U_{1n} + U_{2n} + U_{3n}$$

Because

$$\sqrt{n}U_{1n} = \sqrt{n}W_n$$

we only need to show that

$$\sqrt{n}\left(U_{2n} + U_{3n}\right) = o_p(1)$$

To show this we first show that

$$\sqrt{n}U_{2n} = o_p p(1)$$

$$\sqrt{n}U_{2n} = n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_j K_n\left(x_i - x_j\right)S_{ij}\left[f\left(x_i,\hat{\theta}_n\right) - f\left(x_i,\theta_0\right)\right]$$

$$= n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_j K_n\left(x_i - x_j\right)S_{ij}\frac{\partial f\left(x_i,\bar{\theta}_n\right)'}{\partial q}\left[\hat{\theta}_n - \theta_n\right]$$

for some

$$\bar{\theta}_n \in \left\{\theta : \left\|\theta - \theta_0\right\| \le \left\|\hat{\theta}_n - \theta\right\|\right\}.$$

Since $S_{ij} = O(n^{1/2}\gamma_n^{k/2})$, and using lemma A.1, and assumption 3.2

$$\sqrt{n}U_{2n} = O(\frac{\gamma_n^{k/2}}{n})o(n)O(n^{-1/2}) = o(1)$$

$$\sqrt{n}U_{3n} = n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n} f(x_j,\theta_0)K_n(x_i - x_j)S_{ij}\varepsilon_i$$

$$-f(x_j,\theta_0)K_n(x_i - x_j)S_{ij}[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)]$$

$$-[f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)]K_n(x_i - x_j)S_{ij}[\varepsilon_i + f(x_i,\hat{\theta}_n) - f(x_i,\theta_0)]$$

$$= A_{1n} + A_{2n} + A_{3n}$$

By Lemma A.1, and $S_{jj} = O(n^{1/2}\gamma_n^{k/2})$, $A_{1n} = O(\gamma_n^{k/2})$.

By assumption (3.2) and Lemma A.1 $A_{2n} = O(n^{-1/2}\gamma_n^{k/2})$,

and $A_{3n} = O(n^{-1/2}\gamma_n^{k/2})$.

QED

***Theorem 4.2***

Proof

$$\hat{m}_n - \hat{R}_n = n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[f\left(x_j,\theta_0\right)+n^{-1/2}\gamma_n^{-k/4}\Delta\left(x_j\right)+\varepsilon_j-f\left(x_j,\hat{\theta}_n\right)\right].$$
$$K_n\left(x_i-x_j\right)S_{ij}\left[f\left(x_i,\theta_0\right)+n^{-1/2}\gamma^{-k/4}\Delta\left(x_i\right)+\varepsilon_i-f\left(x_i,\hat{\theta}_n\right)\right]$$
$$= n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[f\left(x_j,\theta_0\right)-f\left(x_i,\theta_0\right)+\varepsilon_j\right]S_{ij}K_n\left(x_i-x_j\right)\varepsilon_i$$
$$+n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[n^{-1/2}\gamma_n^{-k/4}\Delta\left(x_j\right)-\left[f\left(x_i,\hat{\theta}_n\right)-f\left(x_i,\theta_0\right)\right]\right].$$
$$S_{ij}K_n\left(x_i-x_j\right)\left[n^{-1/2}\gamma_n^{-k/4}\Delta\left(x_i\right)+\left[f\left(x_i,\theta_0\right)-f\left(x_i,\hat{\theta}_n\right)\right]\right]$$
$$+n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[f\left(x_j,\theta_0\right)-f\left(x_i,\theta_0\right)+\varepsilon_j\right]S_{ij}K_n\left(x_i-x_j\right).$$
$$\left[f\left(x_i,\theta_0\right)-f\left(x_i,\hat{\theta}_0\right)+n^{-1/2}\gamma_n^{-1/4}\Delta\left(x_i\right)\right]$$
$$+n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[n^{-1/2}\gamma_n^{-1/4}\Delta\left(x_j\right)-f\left(x_i,\hat{\theta}_n\right)-f\left(x_i,\theta_0\right)\right].$$
$$S_{ij}K_n\left(x_i-x_j\right)\varepsilon_i$$
$$= V_{1n}+V_{2n}+V_{3n}+V_{4n}.$$

In Lemma 4.1, we have shown

$$\sqrt{n}V_{1n}\rightarrow N\left(O,V_0\right).$$

We then will show the following:

(i) plim $\sqrt{n}V_{2n}=\mu$, where $\mu$ is defined in the theorem

(ii) plim $\sqrt{n}V_{3n}=0$

(iii) plim $\sqrt{n}V_{4n}=0$.

Starting with (i)

$$\sqrt{n}E\left(V_{2n}\right)=n^{-2}\Sigma_{t=1}^{n}\Sigma_{j\neq i}^{n}E\left(n^{-1/2}\gamma_n^{-k/4}\Delta\left(x_j\right)-\frac{df}{d\Theta}\left(x_i,\tilde{\theta}_n\right)'\left[\hat{\theta}_n-\Theta_0\right]\right).$$

(i)
$$S_{ij}K_n\left(x_i-x_j\right)\left[n^{-1/2}\gamma_n^{-k/4}\Delta\left(x_i\right)-\frac{df}{d\Theta}\left(x_i,\tilde{\theta}_n\right)'\left[\hat{\theta}_n-\Theta_0\right]\right]$$

for some $\tilde{\theta}_n\in\left\{\theta:\left\|\theta-\theta_0\right\|\leq\left\|\hat{\theta}_n-\theta_0\right\|\right\}$.

The summand is equal to

$$\int\int n^{-1/2}\gamma_n^{-k/4}\Delta(x_i-\gamma_{nu})-\frac{df(x_i,\tilde{\theta}_n)'}{d\theta}\left[\hat{\theta}_n-\theta_0\right]S_{ij}K(u)\cdot\left[n^{-1/2}\gamma_n^{-k/4}\Delta(x_i)-\frac{df(x_i,\tilde{\theta}_n)'}{d\theta}\left[\hat{\theta}_n-\theta_0\right]p(x_i-\gamma_{nu})p(x_i)dx_i du\right]$$

$$=n^{-1/2}\gamma_n^{-k/2}\int\left[n^{-1/2}\gamma_n^{-k/4}\Delta(x_i)-\frac{df(x_i,\tilde{\theta}_n)'}{d\theta}\left[\hat{\theta}_n-\theta_0\right]\right]^2 p(x_i)dx_i+Op(1)$$

$$=n^{1/2}\gamma^{k/2}n^{-1}\gamma^{k/2}\cdot E\left(\Delta(x_i)-n^{1/2}\gamma_n^{1/4}\frac{df(x_i,\tilde{\theta}_n-\theta_0)'}{d\theta}\right)\left[\bar{\theta}_n-\theta_0\right]^2+Op(1)$$

$$=n^{-1/2}\mu$$

(iii) Since $E(V_{2n})=0$

$$V_{2n}=Op(n^{-1/2})$$

$$\sqrt{n}V_{2n}\to 0$$

QED

***Lemma 4.3***

An alternative estimator for $V_0$ is

$$\hat{\hat{V}}_n=4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i-x_j)\hat{\varepsilon}_i^2\hat{\varepsilon}_j^2$$

Proof:

Let

$z_i=(x_i,\varepsilon_i)$ and $\hat{z}_i=(x_i,\hat{\varepsilon}_i)$ where is defined in (3.15)

Therefore, taking a Taylor expansion around $\varepsilon_i$ yields:

$$\hat{\hat{V}}_n = 4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i - x_j)\varepsilon_j^2\varepsilon_i^2 + 8n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i - x_j)\bar{\varepsilon}_i\varepsilon_j^2(\hat{\varepsilon}_i - \varepsilon_i) + 8n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i - x_j)\bar{\varepsilon}_j\varepsilon_i^2(\hat{\varepsilon}_j - \varepsilon_j)$$

$$\hat{\hat{V}}_n = 4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i - x_j)\varepsilon_j^2\varepsilon_i^2 + o(1)$$

The last inequality comes from Bierens[1987] where $\hat{\varepsilon}_i - \varepsilon_i = O_p(n^{-1/2}\gamma_n^{-k/2})$.

Let $H_n(z_i, z_j) = \gamma_n^k K_n^2(x_i - x_j)\varepsilon_i^2\varepsilon_j^2$. Then, we can easily use Powel, Stock, and Stocker[1989] to show that $E[||H_n(z_i,z_j)||^2]=o(n)$. Further, the following holds:

$$E\{4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i - x_j)\varepsilon_j^2\varepsilon_i^2\} = \frac{n(n-1)}{2}\frac{1}{n^2}\int_X\int_X \frac{1}{\gamma_n^k}K(\frac{x_i - x_j}{\gamma_n})\sigma^2(x_i)\sigma^2(x_j)dx_i dx_j$$

Letting $u = (x_i - x_j)/\gamma_n$, we get:

$$E\{4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\gamma_n^k K_n^2(x_i - x_j)\varepsilon_j^2\varepsilon_i^2\} = \frac{n(n-1)}{2}\frac{1}{n^2}\int_T\int_X K(u)\sigma^2(x_i)\sigma^2(x_j - \gamma_n u)dx_i du = V_0 + o(1)$$

Therefore, the conditions of Powell, Stocker, and Stocker are satisfied.

***Theorem 4.3***

Let $K(z_1,z_2)$ have the properties list in assumptions (3.9).

Let

$$W_{nij} = \frac{\varepsilon_i K_n}{\gamma_n^{k_1}}(\frac{x_{1i} - x_{1j}}{\gamma_n}, \frac{x_{2i} - x_{2j}}{\gamma_n})\varepsilon_j S(j \in N_i)$$

$$F_{nij} = \frac{\varepsilon_i}{\gamma_n^{k_1}}K_1(\frac{x_{1i} - x_{1j}}{\gamma_n})I(x_{1j} = x_{2j})\varepsilon_j S(j \in N_i)$$

We have already shown that $n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(W_{nij}-F_{nij})=o(n^{-1/2})$. Therefore, it suffices to show that:

$\frac{\sqrt{n}}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}2F_{nij}\xrightarrow{d}N(0,V_0)$.

It is evident that $E(F_{nij}|\varepsilon_i,x_i)= E(F_{nij}|\varepsilon_j,x_{ij})=0$. Therefore, this is a degenerate statistic. Given the assumption (3.10) where $x_2$ can only take a finite number of values, the kernel $K_1(z_1,z_2)I(u_1=u_2)$ satisfies assumption (3.3) for $k=k_1$, and $t > 2k_1$. Assumption (3.4) is satisfied for $x_1$ and $k_1$. Therefore, we can still take the required Taylor Expansions with respect to $x_1$, and in assumption 3.5 we set:

$n' = O(n^{3/2}\gamma_n^{k_1/2})$

We satisfy assumption (3.6) with $k=k_1$. The standardization in assumption (3.7) needs to be done only on $x_1$. Therefore all the assumptions for Lemma 4.1 are satisfied. QED