AN IMPROVED NONPARAMETRIC TEST FOR MISSPECIFICATION OF FUNCTIONAL FORM

By Ralph Bradley and Robert McClelland[1]

1. Introduction

In two recent papers, Bierens (1990,1982) derives a consistent conditional moment test of functional form that can be a powerful tool to verify the specification of a nonlinear model. Unlike other conditional moment tests that test for the nonzero product of the residual from an estimated regression and a function of the explanatory variables of that regression, the Bierens test does not specify an alternative and encompassing form. It is also robust in that the null is not restricted to any particular form, such as the linearity restriction of the notable tests of Ramsey (1969) and Lee, White and Granger (1993), and it does not require a pre-specified parametric alternative that can be consistently estimated, as is true of the test of Hausman (1978).

However, Bierens' test has several shortcomings. While the test is consistent against all deviations from the null, i e. in the limit the test statistic over the sample size is nonzero for all alternative hypotheses, this limit is not shown to be the largest in the class of conditional moment tests. Further, because a direct estimate of the misspecification is not used in the test there is a suspicion that it may be smaller than necessary. Instead, the test is a reduced form test that is based on the first moment of the residual from a regression multiplied by the anti-log of a bounded function of the explanatory variables. This also means that there is little information about the nature of any misspecification detected. Finally, there are important parameters that adversely affect the distribution of the test if chosen to maximize the power of the test. Instead the power is increased through selecting several penalty parameters and then solving a somewhat unintuitive constrained optimization problem.

In this paper, we attempt to develop a test that improves on Bierens' test while keeping its robustness. Our test is based on the intuition that if the model is not properly specified then the vector of explanatory variables, X, can better predict the residuals than the residual's mean. Thus, while many X-measurable functions, such as the antilog of a bounded transform of X, can be used to detect misspecification an appealing function to use is the conditional expectation of the residuals given X. We implement this intuition in the first stage of our test by estimating the residuals as a function of the explanatory variables. Because we obtain our estimate with a kernel regression, our test is nonparametric. The estimate is then a component of the second stage, which is similar to

the Bierens test with the misspecification estimate in the first stage replacing the antilog function.[2] By directly estimating the misspecification with a conditional expectation function, the limit of our test over the sample size is the largest of all conditional moment tests, has fewer arbitrary parameters than the Bierens test and can help researchers better understand the nature of any misspecification.

Another advantage of our test is in the potential for researchers to increase its finite sample power by increasing the predictive accuracy of the explanatory variables. To do this, we cross validate the bandwidth of a kernel regression, e.g., minimizing a quadratic loss function whose argument is the out-of-sample error of the kernel regression. The quadratic loss function is particularly appealing in our case because the expected value is minimized by predicting the residuals with their conditional expectation. This connects with our intuition that the regressors can better predict the residuals than the residual's unconditional mean in a misspecified model, so that procedures that increase this predictability increase the power of our test.

The minimization procedure we use does not impose as many arbitrary constraints as those imposed by the optimization procedure in Bieren's test. Instead we use fewer a priori constraints, attempting to allow the data itself to determine parameters such as the bandwidth. Therefore, our implementation of the test is robust to all misspecifications and is optimized solely on its ability to use the explanatory variables to predict the residuals. Yet under the null, the best predictor of the residuals is its unconditional mean when the loss function is quadratic, and we show that the kernel regression converges to the sample mean of the residuals. Therefore, we can show that our statistic has a asymptotic $\chi^2(1)$ distribution under the null.

It should be noted that even if the population has independently and identically distributed errors, there is no guarantee that the real size of our test will converge to its nominal size because of dependence inherent in the residuals produced by a regression. To solve this, we use standard bootstrapping techniques to obtain a subsample from the entire sample of residuals and regressors that ensures that under the null hypothesis, the real size of our test converges to its nominal size regardless of the choice of the window width.

Finally, we compare Monte Carlo simulations of our test and those in Bierens (1990). While both tests have approximately the correct distribution under the null hypothesis, our test rejects a greater proportion of the

simulations at all sample sizes when the hypothesis is false. We then choose one simulation at random under this alternative hypothesis and show how the kernel estimation in the first stage our test can be used to describe the nature of the misspecification.

The remainder of our paper is organized as follows. Section 2 briefly discusses Bierens test. Section 3 derives our test. Section 4 discusses results from Monte Carlo experiments that compares our test to the Bierens test and conclusions are in Section 5. The assumptions on the random vectors (y,x) are listed in Appendix A, and the Proofs and Lemmas are in Appendix B.

2. The Bierens Test

We first describe the calculation of the Bierens Test, and then briefly discuss the theory behind his test. Suppose the random vector (y,x), $y \in \Re$ and $x \in \Re^k$, has a joint probability distribution F(y,x) where the following holds:

$$E(y|x) = f(x)$$

$$y = f(x) + u,$$

where $u \in \Re$ is the error of the model. It is clear that E(u|x) = 0 for all x and E(u)=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 Θ is the parameter space and is a compact subset of $\Re^m$. Using the notation of Bierens (1990), we denote D(f) as the set of all probability distribution functions F(y,x) on $\Re \times \Re^k$ such that P(f(x)) = E(y|x) = 1. Therefore, we wish to test that a specific parameterization of f(x) which is denoted $f(x,\theta)$ satisfies the null hypothesis:

(1.) H_o: The distribution F(y,x) belongs to $D_0 = \bigcup_{\theta \in \Theta} D(f(.,\theta))$

In other words, the null states that there is at least one θ in Θ such for all x values in the support of x except for subsets with zero measure, $E(y|x) = f(x,\theta)$. The alternative hypothesis is that

(2.) H_1 : The distribution belongs to $D_1 = \bigcup_{f} D(f) \setminus D_0$

where the f is the union over all Borel measurable real functions, f, on $\Re^k$. Given a consistent estimator for θ which we denote as $\hat{\theta}$, Bierens (1990) proposes to test the null hypothesis based on the following statistic from the random sample $\{y_i, x_i\}$, i = 1,...,n and $t \in \Re^k$:

(3.a) $\hat{W}(t,\hat{\theta}) = n[\hat{M}(t,\hat{\theta})]^2 / \hat{s}^2(t),$

where

(3.b) $\hat{M}(t,\hat{\theta}) = (1/n)\sum_{j=1}^{n}(y_j - f(x_j,\hat{\theta}))\exp(t'\,\Phi(x_j)).$

In (3.a), $\hat{s}^2(t)$, is a consistent estimator of the asymptotic variance of $\sqrt{n}\,\hat{M}(t,\hat{\theta})$, and in (3.b), $\Phi(x)$ is a bounded x measurable function from $\Re^k$ to $\Re^k$ and there is finite vector $C \in \Re^k$ such that $\sup_x \Phi(x) < C$. Bierens uses the inverse tangent function where the *i*th element in the vector is $\tan^{-1}(x_i)$.

The test in (2.a) is a conditional moment test, which means that it is a member of the set of tests $\{W_h\}$, defined by

(4.) $$\{W_h\} = \left\{ \left[\frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(y_j - f(x_j,\hat{\theta}))h(x_j)}{\hat{s}} \right]^2 \middle| h(x) \in H(X) \right\},$$

where H(X) is the set of X-measurable functions and X is the support of x. One can use the probability limit of W/n, which is equal to $E(u(h(x))/E(u^2h(x)^2)$, to characterize the asymptotic power of the test.[3] Newey (1985) describes this test in which h(x) is equal to a finite sum of weights, each of which imposes a restriction upon the conditional moment. As Bierens points out, this finite number of moment conditions implies that the Newey test cannot be consistent against all possible alternative hypotheses.

In contrast, the use of $h(x) = \exp(t'\,\Phi(x))$ makes the test in (2.a) consistent against all deviations from the null hypothesis. The theoretical underpinning for this comes from Theorem 1 in Bierens (1982), where he shows that if $f(x,\theta)$ is misspecified, then $E\{u*\exp(t'\,\Phi(x))\}$is nonzero for some $t \in \Re^k$. Bierens further shows that $E\{[y-f(x,\theta)]\exp(t_o'\,\Phi(x))\}$is nonzero for some t_o in a neighborhood of zero. The expectation of the right-hand side of (3.b), which is used as an estimate of $E\{[y-f(x,\theta)]\exp(t_o'\,\Phi(x))\}$, will then be nonzero if the model is misspecified.

These results are used in Bierens (1990) Theorem 2 to show that $\hat{W}(t,\hat{\theta})$ has asymptotic power against the general alternative (2.). Unlike the Ramsey RESET test (1969), a wide range of nonlinear nulls can be tested for correct specification. Likewise, unlike the Hausman (1978) test, this test is robust against any departure from the null that is Borel-measurable.

Although Bierens shows that $\hat{W}(t,\hat{\theta})$ has asymptotic power against all alternatives, he says nothing about its asymptotic power relative to other conditional moment tests in the set $\{W_h\}$. Further, because the function $\exp(t'\Phi(x_j))$ does not directly estimate the misspecification, it is reasonable to suspect that a more powerful test exists. This suspicion is reinforced by the dependence of power of the test upon t, and as Bierens points out, one cannot choose t to maximize $\hat{W}(t,\hat{\theta})$ because the asymptotic distribution of $\max_t \hat{W}(t,\hat{\theta})$ converges to the supremum of a χ^2 process and not a χ^2 itself. In a sense, $\max_t \hat{W}(t,\hat{\theta})$ overfits the sample moments under the null.

In order for $\hat{W}(t,\hat{\theta})$ to achieve an asymptotic $\chi^2(1)$ distribution, Bierens implements the test by using random uniform draws of t from elements in a compact hypercube whose boundaries must be picked in an arbitrary manner. In other words, t is a vector of random variables whose parameters cannot be chosen to optimize the power of the test. Given these random draws, the statistic does not have a χ^2 distribution under the null unless the researcher further solves a constrained maximization problem, using constraints on $\hat{W}(t,\hat{\theta})$ formed by a priori penalty parameters. Because of the innovative nature of the problem there is little intuition to guide researchers as to the choice of these parameters. The dependence of the test upon the hypercube boundaries and these penalty parameters is therefore unfortunate because Bierens claims that the finite sample power of $\hat{W}(t,\hat{\theta})$ is sensitive to their selection.

Finally, the lack of a direct estimate of the misspecification means that there is little to direct the researcher toward a functional form that better specifies E(y|x). In order to gain insight into the nature of the misspecification when using (3.a), Bierens uses an additional model with the residuals as a dependent variable and $\exp(t'\Phi(x_j))$ as a right-hand side variable needs to be estimated. However, if the residual is a function of the explanatory variables, why not improve the efficiency of the test by incorporating an estimate of this function into the test itself?

3. The Kernel Test

In this section we propose a test in the set of conditional moment tests $\{W_h\}$ that uses a direct estimate of the misspecification $g(x)=E(y-f(x,\hat{\theta})|x)$ as the function h(x). Like the Bierens test, our test is consistent against all deviations from the null hypothesis. However, our test has the greatest asymptotic power of all tests in $\{W_h\}$in the

sense that the probability limit of the kernel estimate is g(x), which we show is the function that maximizes plim (W/n), where W is defined in (4.). We also increase its finite sample size power by using a cross validated window width. This cross validation increases the power of the test over a fixed window width even though the kernel regression estimator converges slowly to its asymptotic limit. In addition, the test presented here does not rely on random draws of t from a hypercube with arbitrary boundaries or the initial choice of nonintuitive penalty parameters. Instead, we use standard kernel regression and bootstrapping techniques that have been extensively studied and set two parameters: an upper bound on the window width and the size of the bootstrap sample. Finally, we can use the estimate of g(x) to gather insight into the nature of any problem detected by the test.

To start, we restrict ourselves to an independently and identically distributed (IID) random sample $\{y_i, x_i\}$, i=1,...,n from a distribution F(y,x) on $\Re\times\Re^k$ for which $E(y^2) < \infty$. As in Bierens, the extension to serially correlated series should be possible but is outside the bounds of this paper. Suppose we use the functional form $f(x,\theta)$ to estimate E(y|x). Under H_o, we assume that the true value of θ which we denote as θ_o satisfies:

(5.) $\theta_o = \text{argmin}_{\theta\in\theta} E([y_i - f(x_i,\theta)]^2)$ and $E([y_i - f(x_i,\theta_o)]|x_i) = 0$.

Let $\hat{\theta}$ be a consistent estimate of θ_o. This study is based on assumptions, outlined in Appendix A, that ensure that under H_o $\text{plim}\hat{\theta}-\theta_o = 0$ and under H_1 plim $\hat{\theta}$ exists. If H_o in (1.) is correct, then the asymptotic expectation of the product of $y-f(x,\hat{\theta})$ and any x-measurable function equals zero and

(6.) plim $E(y-f(x,\hat{\theta})|x) = 0$ for all x.

Under H_1 ,there is a set X of non zero measure such that:

7.) plim $E(y-f(x,\hat{\theta})|x) \neq 0$ for all $x \in X$.

since $Pr[\lim f(x,\hat{\theta}) - E(y|x)] < 1$.

Define $\hat{u}_i = y_i - f(x_i,\theta)$; then the Rao-Blackwell theorem tells us that under H_1 :

(8.) $\lim E(\hat{u}_i - E(\hat{u}_i|x_i))^2 < \lim E(\hat{u}_i - E(\hat{u}_i))^2$

which implies $\lim E(\hat{u}_i E(\hat{u}_i|x_i)) = \lim E(E(\hat{u}_i|x_i)^2) > 0$. Under H_o, since $\text{plim}\hat{\theta}=\theta_o$, the inequality in (8.) becomes an

equality. This implies that under H_o, lim $E(\hat{u}_i E(\hat{u}_i|x) = 0$. Therefore, our test is based on a nonparametric estimator for $E(\hat{u}_i E(\hat{u}_i|x)$. The features in (5.) through (8.) suggest the following four steps to test the null in (1):

(i) Estimate θ and denote this estimate as $\hat{\theta}$.

(ii) Generate $\hat{u}_i = y_i - f(x_i,\hat{\theta})$ for i = 1,..,n.

(iii) For each observation i, estimate $E(\hat{u}_i|x_i)$ with a kernel regression by selecting a random subsample with replacement of size n' < n from the existing sample. Specifically, for the original sample of size n, Let N={1,...,n} index the sample. We then choose for each sample point i a random subsample with replacement from N of size n' which we denote as N_i' and calculate for each i a kernel estimator denoted as $\hat{g}_i(x_i, \gamma_i)$ where:

(9.) $$\hat{g}_i(x_i,\gamma_i) = \frac{\sum_{j\in Ni'} \hat{u}(x_j)K(\frac{x_i - x_j}{\gamma_i})}{\sum_{j\in Ni'} K(\frac{x_i - x_j}{\gamma_i})}$$

and K(.) is a kernel function which has the following properties:

(P.1) $\int K(u)\,du = 1$

(P.2) $\int uK(u)du = 0$

(P.3) K(u) > 0 for all u.

(P.4) $0 = \arg\max K(u)$, K'(u) > 0 for u<0 and K'(u) < 0 for u>0.

(iv) Calculate the statistic:

(10.) $$\hat{W}(\gamma*,\hat{\theta}) = n\,\hat{T}^2(\gamma*,\hat{\theta})/\hat{s}^2(\gamma_i),$$

where

(11.) $$\hat{T}(\gamma*,\hat{\theta}) = (1/n)\left[\sum_{i=1}^{n} \hat{u}(x_i)\hat{g}_i(x_i,\gamma_i)\right],$$

$\gamma* = \{\gamma_1,\gamma_2,...,\gamma_n\}$

(12.) $$\hat{s}^2(\gamma^*) = \frac{1}{n}\sum_{i=1}^{n} \hat{u}(x_i)^2(\hat{g}(x_i,\gamma_i) - \hat{b}(\gamma^*)\hat{A}^{-1}(\partial/\partial\theta)f(x_i,\hat{\theta}))^2$$

(13.) $$\hat{b}(\gamma^*) = \sum_{i=1}^{n} (\partial/\partial\theta)f(x_i,\hat{\theta})\hat{g}(x_i,\gamma_i)$$

(14.) $$\hat{A} = \sum_{i=1}^{n} \{(\partial / \partial \theta')f(x_i, \hat{\theta})\}\{(\partial / \partial \theta) f(x_i, \hat{\theta})\}.$$

Given the properties (P.1) through (P.4) our test, like Bierens', is consistent against all alternative hypotheses. This follows from the fact that if $E\{\hat{u} \cdot \exp(t' \Phi(x))\}$ is nonzero, then $E\{\hat{u} \cdot \hat{g}(x,\gamma)\}$ is also nonzero (see Bierens (1987)). But because the first stage of our test estimates $E(\hat{u}|x)$ with $\hat{g}(x,\gamma)$, if $\hat{g}(x,\gamma)$ is consistent then the asymptotic power of our test should be at least as great as the Bierens test. The reason for this difference is that $E(\hat{u}|x)$ is the function h(x) in the set of bounded functions which maximizes $E\{\hat{u}h(x)\}$. In essence, $\hat{T}$ in equations (10.) and (11.) is an estimator of $E\{\hat{u}\, E(\hat{u}|x)\}$ and therefore the numerator in equation (10.) uses an estimator that maximizes the value of $E\{\hat{u}h(x)\}$.

To increase the predictability of $\hat{g}(x,\gamma)$ in finite samples, we choose γ_i as the cross validated window width by solving:

$$\gamma_i = \underset{0<\gamma<B_n<\infty}{\operatorname{argmin}} \left[\sum_{k \in Ni'} (\hat{u}(x_k) - \hat{\hat{g}}_{-k}(x_k, \gamma))^2 \right]$$

where

$$\hat{\hat{g}}_{-k}(x_k, \gamma) = \frac{\sum_{j \in Ni', j \neq k} \hat{u}(x_j) K(\frac{x_k - x_j}{\gamma})}{\sum_{j \in Ni', j \neq k} K(\frac{x_k - x_j}{\gamma})}$$

and the bound B_n to have two properties:

(P.5) $B_n \leq B_{n+1}$

(P.6) $\limsup B_n = B < \infty$.

The limit, B, on the window width γ_i is the only parameter we must choose a priori and has another important feature under H_o.[4] The bound B along with the property (P.4) for the kernel $K(\cdot)$ guarantees that $\hat{s}(\gamma^*)^2$ is always positive as long as the variance of x is greater than zero. In addition, searching for the optimal γ over a bounded set guarantees a solution in a finite time.

Given that this test uses standard a standard kernel regression to estimate the misspecification, it is perhaps surprising that this test has not been used earlier. Two possible complications may explain this. First, $\hat{g}_i(x_i,\gamma_i)$ is not an IID process even though the errors in the true model are IID under H_o, because there is finite sample dependence among observations of $\hat{u}_i$ arising from the first moment restriction, $\sum_{i=1}^{n} \hat{u}(x_i) = 0$ when there is a constant term. We solve this problem in step three by using a bootstrap from the original sample. This bootstrapping allows us to use a projection theorem that shows that the moments of $\{\hat{u}_i\hat{g}_i(x_i,\gamma_i)\}$, i=1,2,..n converge in probability to the moments an IID random process. The only parameter we must choose is the size of the bootstrap, which can be chosen according to standard bootstrap criteria, as long as the bootstrapped sample is of size O(n).

A second problem may be the bias inherent in a kernel estimate. However, the power of our test does not depend on the lack of bias in $\hat{g}_i(x_i,\gamma_i)$. Rather it depends upon the ability of $\hat{g}_i(x_i,\gamma_i)$ to predict $\hat{u}_i$. If any X measurable nonconstant function can better predict the residuals using a quadratic loss function, then by the Rao-Blackwell theorem $E(\hat{u}_i,|x_i)$ is non-constant and therefore we can detect the misspecification.

We now state the basic results of this paper:

THEOREM 1: Let the assumptions in Appendix A hold. The statistic $\hat{W}(\gamma^*,\hat{\theta})$ *generated in equation (10.) is asymptotically distributed as* $\chi^2(1)$ *under* H_o.

PROOF: See Appendix B.

$\hat{W}(\gamma^*,\hat{\theta})$ is essentially a Wald Test for the restriction that $E(\hat{u}[E\hat{u}|x]) = 0$. Because we are imposing one restriction the limiting distribution must have one degree of freedom.

While both $\hat{W}(t)$ and $\hat{W}(\gamma^*,\hat{\theta})$ are conditional moment tests consistent against all alternatives we now show that $\hat{W}(\gamma^*,\hat{\theta})$ is most asymptotically powerful in the set of statistics W. We do this by first showing that a function h(x) that is proportional to g(x) is the solution to the problem:

(15.) $$\max_{h \in H(X)} \operatorname{plim} \frac{W}{n} = \left[\frac{E(uh(x))}{E(u^2 h(x)^2)} \right]^2$$

and then observing that plim $\hat{g}_i(x_i,\gamma_i) = g(x)$ implies that plim$\hat{W}(\gamma^*, \hat{\theta})/n$ is the largest among from the set W.

PROPOSITION 1: Let {x,u} be a random vectors for some probability space where E(u)=0, and E(u^2) = $\sigma^2 < \infty$. *Denote* g(x) = E(u|x) *and* g(x) *is an* X *measurable continuous function. Then the solution to* (1.) *is proportional to* g(x).

PROOF: See Appendix B.

COROLLARY 1: Given the assumptions Appendix A, plim $\hat{W}(\gamma^*, \hat{\theta})/n$ *equals the maximand corresponding to the solution of (15.).*

4. Monte Carlo Experiments

In this section we briefly explore the finite sample size and convergence properties of our test by conducting an experiment identical to that conducted in Bierens (1990). We then give an example of how the information from $\hat{g}_i(x_i,\gamma_i)$ can be used to understand the nature of the misspecification.

In Bierens (1990) the following data generating mechanism is examined:

$$x_{1i} = z_i + v_{1i}\ ,\ v_{1i} \sim N(0,1)$$

$$x_{2i} = z_i + v_{2i}\ ,\ v_{2i} \sim N(0,1)$$

$$y_i = 1 + x_{1i} + x_{2i} + u_i\ .$$

The residual u_i is generated by either one of two data generating processes

(DGP1) $u_i = v_{1i}\, v_{2i} + e_i$

or

(DGP2) $u_i = \sqrt{2}\, e_i$.

where

$$e_i \sim N(0,1).$$

Both processes are fitted with the following model:

(16.) $y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + u_i$

The null is false for DGP1 and true for DGP2.

In Table I we compare the results of our test (labeled Kernel) with N_i=.95N and a Gaussian kernel on these to data generating processes with the results in the rows in Table 1 of Bierens (1990) where the penalty parameters γ and ρ yielded the most rejections of DGP1 while still maintaining reasonable size properties for DGP2. The hypercube chosen was [1,5]×[1,5] from which [N/10] - 1 draws were taken and the weighting function $\Phi(x)$ is set to $\tan^{-1}(x/2)$.

Our test rejects DGP1 more often than the Bierens test at all data sizes while offering similar size characteristics. In particular it does well at small sample sizes, rejecting the null hypothesis at the 5 per cent critical level in more than two thirds of the simulations using 50 observations. With one hundred observations, it rejects more than 90 per cent of the simulations and with two hundred observations it rejects all but three of the simulations at the 5 per cent critical value.

It should be pointed out that Bierens states that his choice of hypercube accidentally falls in one of the worst areas in terms of power. A better choice may yield more rejections for DGP1. Unfortunately, choosing the hypercube to maximize the statistic is not allowed. This itself points out a problem with the test: it crucially depends upon a parameter which the researcher cannot choose to increase the power of the statistic.

As an example of how to use the information from our test to better understand the nature of the misspecification, we more closely examine the final simulation of DGP 1 in which $\hat{W}(\gamma^*, \hat{\theta})$ is equal to 20.40. In figure one we show the true conditional mean of the residual as a function of the random variables x_1 and x_2. It is clear from the parabolic shape that a quadratic term is missing from the original specification. In Figure 2, we plot the true conditional mean of the residual against the realizations of x_1. The parabolic shape in figure one is clearly reflected in this figure as well. The plot of the kernel estimate also reflects the parabolic shape, although the finite sample bias of the kernel estimate causes an upward shift. Figure 3 plots the same two variables against the realizations of x_2. Again the parabolic shape is evident in both the expectation of the residual and the kernel estimate.

While it would seem useful to plot the kernel estimate and its confidence intervals with the expected residuals in a diagram similar to figure one, the complexity of such a figure renders it practically useless. Instead Figure 4 plots the kernel estimate and its 95 per cent confidence intervals against the expected residual, holding x_2 fixed at the arbitrary level of 0.5. In this figure the kernel estimate hovers around expected residual, and the residual is within the confidence intervals in about 95 per cent of the cases. Combining the information provided by the kernel estimates in the figures, a researcher might suspect that a quadratic term is missing from the regression.

As a final step we run two additional regressions. In the first, we include the missing quadratic components as explanatory variables. In the second, we include the kernel regression as an explanatory variable. Running our test on these new regressions yields the insignificant values of 0.12 and 3.30, respectively.

5. Conclusions

The test of Bierens (1990) is an important advance in attempt to detect misspecification of the functional form through a conditional moment test. In this paper we describe an improvement upon the Bierens test that directly estimating E(û|x) with a standard kernel regression on a bootstrapped sample of the data. It retains all of the robustness of Bierens' test, but does not requires the arbitrary selection of penalty parameters and boundaries and needs fewer parameters to be determined by the researcher. Through the use of E(û|x) our test is the most powerful in the class of conditional moment tests. Because E(û|x) is estimated, researchers obtain an estimate of any misspecification. In addition, cross-validation of the bandwidth in the kernel regression can be used to increase the power of the test, which we conjecture is at least as great as the Bierens test. We show the results of a Monte Carlo simulation study that suggests that our test has good power against misspecification. Finally, we use the results from an arbitrary simulation to demonstrate how the estimate of E(û|x) can be used to understand the nature of the misspecification.

US Bureau of Labor Statistics

APPENDIX A

Assumptions for the random vector (y,x)

(A.1) $\{y_i,x_i\}$, i=1,...,n are a simple random sample from a continuous probability distribution on $\Re\times\Re^k$ with

$E(y_i^2) < \infty$.

(A.2) Under H_o, the parameter space Θ is a compact and convex subset of $\Re^m$ and $f(x,\theta)$ is a Borel measurable real function on real k and for each k vector x is a twice continuously differentiable real function on θ. $E[\sup_{\theta\in\Theta} f(x_i,\theta)^2] < \infty$ and for $i_1,i_2 = 1,..,m$,

$$E\left[\sup_{\theta\in\Theta}\left|\{(\partial/\partial\theta_{i_1})f(x_1,\theta)\}\{(\partial/\partial\theta_{i_2})f(x_1,\theta)\}\right|\right] < \infty$$

$$E\left[\sup_{\theta\in\Theta}\left|\{y_1 - f(x_1,\theta)\}^2\{(\partial/\partial\theta_{i_1})f(x_1,\theta)\}\{(\partial/\partial\theta_{i_2})f(x_1,\theta)\}\right|\right] < \infty$$

$$E\left[\sup_{\theta\in\Theta}\left|\{y_1 - f(x_1,\theta)\}(\partial/\partial\theta_{i_1})(\partial/\partial\theta_{i_2})f(x_1,\theta)\right|\right] < \infty$$

(A.3) $E(y_1 - f(x_1,\theta))$ takes on a unique minimum on Θ at θ_o. Under H_o, the parameter vector θ_o is an interior point of Θ.

(A.4) The matrix A defined in (13.) is nonsingular.

APPENDIX B

Proofs

In the standard literature on kernel regression estimation the optimal window with γ_i is $O(n^{-1/(k+4)})$; however under H_o, we show:

LEMMA 1: Let v *be a random variable satisfying* $E|v| < \infty$ *and* $E(v) = 0$. *Let* x *be a bounded random vector in* $\Re^k$. *Suppose that* $\hat{h}_{-i}(x,\gamma)$ *is the kernel regression of v on x with the i-th observation excluded from the calculation of the kernel regression:*

$$\hat{h}_{-i}(x_i) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} v_j K(\frac{x_i - x_j}{\gamma})}{\sum_{\substack{j=1 \\ j\neq i}}^{n} K(\frac{x_i - x_j}{\gamma})}.$$

Denote g_n^* *as the* γ *that minimizes sample variance,*

$$S_n(\gamma) = \sum_{i=1}^{n} (v_i - \hat{h}_{-i}(x_i,\gamma))^2 / n.$$

If $P[E(v|x)=0]=1$, *then plim* $g_n^* = \infty$.

PROOF:

$g_n^* = \text{argmin}_\gamma S_n(\gamma)$. If $Pr(E(v|x) = E(v)= 0)=1$, we show that plim $g_n^*=\infty$.

Our proof uses the following facts:

i) Let μ be a constant and $\bar{S} = E(v - \mu)^2$. Then $\bar{S}$ is minimized when $\mu = 0$, and $\partial S / \partial \mu$ is continuous in μ.

ii) $\hat{h}_{-i}(x_i,\gamma) = \bar{v}$ at $\gamma = \infty$ where $\bar{v}$ is the sample average of v.

iii) $S_n(\gamma)$ converges in probability uniformly in γ to a nonstochastic function, $S(\gamma)$, and $\min_\gamma S(\gamma)= S(\infty)=\bar{S}$.

Let ψ_n be the half closed interval $[n,\infty)$ and let $\delta_n = \min_{\gamma\notin\psi_n} S(\infty) - S(\gamma)$. Define A_n to be the event

$|S_n(\gamma)-S(\gamma)|<\delta_n/2$ for all γ. Then

(B.1) $A_n \Rightarrow S(g_n^*) < \delta_n/2 + S_n(g_n^*)$

and by definition of $S_n(g_n^*)$, we know:

(B.2) $S_n(g_n^*) \leq S_n(\infty)$

(B.3) $A_n \Rightarrow S_n(\infty) < \delta_n/2 + S(\infty)$.

Combining (B.1), (B.2), and (B.3) we obtain:

(B.4) $A_n \Rightarrow S(g_n^*) < S(\infty) + \delta_n$

Therefore $A_n \Rightarrow g_n^* \in \psi_n$ which implies,

(B.5) $P(A_n) \leq P(g_n^* \in \psi_n)$.

By (iv) above $P(A_n) \rightarrow 1$

Then since inf $\psi_n \rightarrow \infty$, plim $g_n^* = \infty$.

QED

LEMMA 2

Given H_0, the assumptions in Appendix A, and the properties (P.1) through (P.6),

(i) $$\lim_{n\to\infty} E\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{u}_i g(x_i, \gamma_i) \right) = 0$$

where $\hat{g}(x_i, \gamma_i, \hat{\theta})$ is the regression defined by (8.) based on resampling $\{x_i, \hat{u}(x_i)\}$ and by errors validation is with replacement, where the size of the replacement is O(n). Furthermore,

(ii) $$\frac{1}{\sqrt{n}} \sum \hat{u}_i \hat{g}(x_i, \gamma_i, \hat{\theta}) \text{ is } O_p(1)$$

which implies

(iii) $$\text{plim} \frac{1}{n} \Sigma \hat{u}_i \hat{g}(x_i, \gamma_i, \hat{\theta}) = 0,$$

PROOF: For simplicity we drop the argument $\hat{\theta}$ and write $\hat{g}\left(x_i, \gamma_i, \hat{\theta}\right)$ as $\hat{g}(x_i, \gamma_i)$. To prove (i) first note that

$$\text{(B.6)} \qquad \frac{1}{\sqrt{n}} \sum \hat{u}_i \hat{g}(x, \gamma_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n'} \hat{u}_i \sum_{j \in N_i} \hat{u}_j w_{ij},$$

where

$$\text{(B.7)} \qquad w_{ij} = \frac{K\left(\frac{x_i - x_j}{\gamma_i}\right)}{\sum_{k=1}^{n'} \left(\frac{x_i - x_k}{\gamma_i}\right)},$$

and K(.), N_i and n' are as defined in the text. Then to prove (i), it is sufficient to show that

$$\text{(B.8)} \qquad \lim_{n \to \infty} E\left[\hat{u}_i \hat{u}_j w_{ij}\right] = 0 \ \forall\ i \in \{1,...,n\},\ j \in N_j .$$

Based on sampling with replacement

$$E\left(\hat{u}_i \hat{u}_j w_{ij}\right) = E\left(\hat{u}_i E(\hat{u}_j | \hat{u}_i) w_{ij}\right).$$

Because

$$\sum_{i=1}^{n} \hat{u}_i = 0,$$

we know

$$-\sum_{j \neq i} \hat{u}_j = \hat{u}_i ,$$

and

$$E\left(\hat{u}_j | \hat{u}_i\right) = \begin{cases} -\hat{u}_i/(n-1) & \text{if } j \neq i \\ \hat{u}_i & \text{if } j = i \end{cases}.$$

Therefore, based on resampling, the assumptions in appendix A and (P.1) to (P. 6),

$$\text{(B.8)} \qquad E\left(\hat{u}_i \hat{u}_j w_{ij}\right) = \Pr(i = j) E\left(\hat{u}_i^2 w_{ii}\right) - \Pr(i \neq j) E\left(\frac{\hat{u}_i^2}{n-1} w_{ij}\right)$$

or

$$\text{(B.9)} \qquad E\left(\hat{u}_i \hat{u}_j w_{ij}\right) = \frac{1}{n} E\left\{\hat{u}_i^2 \left(w_{ii} - w_{ij}\right)\right\}.$$

It can be shown that

$$\lim_{n \to \infty} E\left[\hat{u}_i^2 \left(w_{ii} - w_{ij}\right)\right] = 0 ,$$

and therefore

$$\lim_{n\to\infty} E\left[\hat{u}_i \hat{u}_j w_{ij}\right] = 0.$$

We can prove (ii) by showing that

$$n\ \mathrm{Var}\left(\frac{1}{n}\sum \hat{u}_i \hat{g}(x_i, \gamma_i)\right) = O(1)$$

or equivalently that

$$nE\left(\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{g}(x_i, \gamma_i)\right)^2 = O(1).$$

To show this first note that because $\hat{u}_i\hat{g}(x_i,\gamma_i)$ is identically distributed,

$$nE\left(\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{g}(x_i, \gamma_i)\right)^2 = E\left(\hat{u}_i^2 \hat{g}^2(x_i, \gamma)\right) - \frac{n(n-1)}{n} E\left[\left(\frac{\hat{u}_i^2}{n-1}\right)\hat{g}(x_i, \gamma_i)\hat{g}(x_j, \gamma_j)\right].$$

Since

$$\hat{u}_i^2 = O_p(1),$$

it suffices to show that under H_o,

$$\hat{g}(x_i, \gamma_i)^2 = O_p(1)$$

and

$$\hat{g}(x_i, \gamma_i)\hat{g}(x_j, \gamma_j) = O_p(1).$$

Note that

$$\text{(B.10)} \quad \hat{g}(x_i, \gamma_i)^2 = \left(\sum_{j\in N_1} \hat{u}_j w_{ij}\right)^2 = \sum_{j\in N_i}\sum_{k\in N_i} \hat{u}_j \hat{u}_k w_{ij} w_{ik}.$$

Given the properties (P.1) through (P.6), it follows:

$$\text{(B.11)} \quad \left(\sum_{j\in N_i} w_{ij}\right)^2 = \sum_{j\in N_i} w_{ij}^2 + \sum_{\substack{k,j\in N_i \\ k\neq j}} w_{ik} w_{ij} = 1.$$

From (B.11) we can show that $w_{ij}w_{ik}$ is $O(n'^{-2})$ for all but finitely many i, j and k. With the additional condition that x is continuously distributed if there is at least one $w_{ij}w_{ik}$ is greater than $O(n'^{-2})$, then there exists some neighborhood around x_j with an infinite number of elements such that $w_{ij}w_{ik}$ is greater than $O(n'^{-2})$, contradicting the limit of (B.11) as n goes to infinity. This implies that (B.10) is

$$n^2 O_p\left(n^{-2}\right) = O_p(1).$$

To show that $\hat{g}(x_i, \gamma_i)\hat{g}\left(x_j, \gamma_j\right)$ is no greater than $O_p(1)$, we rewrite

$$\hat{g}(x_i, \gamma_i)\hat{g}\left(x_j, \gamma_j\right) = \sum_{k \in N_i} \hat{u}_k w_{ik} \sum_{h \in N_j} \hat{u}_j w_{jh}$$
$$= \sum_{k \in N_i} \sum_{h \in N_j} \hat{u}_k \hat{u}_h w_{ik} w_{jh}.$$

If $k \neq h$ then

$$E\left(\hat{u}_k \hat{u}_h w_{ik} w_{jh}\right) = E\left(\hat{u}_k \frac{\hat{u}_k}{n-1} w_{ik} w_{jh}\right).$$

Because $w_{ik}w_{jh}$ is $O(n^{-2})$, we know that $\hat{u}_k\hat{u}_h w_{ik}w_{jh}$ is $O(n^{-3})$ for all but finitely k and h. If k=h then

$$E\left(\hat{u}_k \hat{u}_h w_{ik} w_{jh}\right) = \hat{u}_j^2 w_{ik} w_{jh}$$

is $O(n^{-2})$ for all but finitely many k. Therefore, defining n^j as card $(N_i \cap N_j)$

$$E\left(\sum_{k \in N_i} \sum_{h \in N_j} \hat{u}_k \hat{u}_h w_{ik} w_{jh}\right) = E\left[n^j\left(\hat{u}_k{}^2 w_{ik} w_{jh}\right)\right] - \left((n')^2 - n^j\right)\left(\frac{\hat{u}_k}{n-1} w_{ik} w_{jk}\right)$$

Since n^j is O(n') the order of

$$E\left(\sum_{k \in N_i} \sum_{h \in N_j} \hat{u}_k \hat{u}_h w_{ik} w_{jh}\right)$$

is of order

$$O(n')O\left(n^{-2}\right) + O\left(n^2\right)O\left(n^{-3}\right) \leq O(1)$$

(i) and (ii) imply (iii) by Chebychev's inequality. This completes the proof.

LEMMA 3: (Projection Theorem for Sampling with Replacement).

Let z_i be a random sequence of IID variables. And define

(B.11) $$T = \frac{1}{n}\sum_{i=1}^{n} E\left[\sum_{j \in N_i} p\left(z_i, z_j\right) \middle| z_i\right]$$

where N_i is a sample with replacement from the set of integers {1, ... , n} when each point has a probability of (1/n), n' = card $(N_i) \leq 0(n)$ and p(.,.) is a k-dimensional symmetric kernel. Define

(B.12) $$\hat{T} = \frac{1}{n}\sum_{i=1}^{n} \sum_{j \in N_i} p\left(z_i, z_j\right).$$

If

$$E\left[p(z_i, z_j)^2 \right] = o(n^{-1})$$

and

$$E\left[p(z_i, z_j) p(z_k, z_h) \right] = O(n^{-2}) \text{ for } i,j \neq k,h$$

then

$$\sqrt{n}\left|\hat{T} - T\right| = o_p(1).$$

PROOF:

Define

$$q(z_i, z_j) = p(z_i, z_j) - \frac{1}{n'} E\left[\sum_{j \in N_i} p(z_i, z_j) \middle| z_i \right]$$

so that

$$\hat{T} - T = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in N_i} q(z_i, z_j),$$

$$E(q(z_i, z_j)) = 0$$

and

$$E(\hat{T} - T)^2 = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{j \in N_i} \sum_{h \in N_k} E q(z_i, z_j) q(z_k, z_h)$$

$$= \left(\frac{1}{n}\right)^2 E \sum_{i=1}^{n} \sum_{j \in N_i} q(z_i, z_j)^2 + \left(\frac{1}{n}\right)^2 E\left(2 \sum_{i=1}^{n} \sum_{k=i+1}^{n} \sum_{j \in N_i} \sum_{h \in N_k} q(z_i, z_j) q(z_k, z_h) \right)$$

The expectations are such that

$$O\left(q(z_i, z_j)^2 \right) = O\left(p(z_i, z_j)^2 \right)$$

and

$$O(q(z_i, z_j) q(z_k, z_h)) = O(p(z_i, z_j) p(z_i, z_h)).$$

Since

$$E(q(z_i, z_j) q(z_k, z_h)) = 0$$

when i, j, k and h are not equal to each other, and based on sampling with replacement card ($N_i \cap N_j$) is O(n'),

implying that

$$\sum_{j \in N_i} \sum_{h \in N_k} I(j = h) = O(n').$$

Then $nE(\hat{T} - T)^2$ is of order

$$nO\left(n^{-2}\right)o\left(n^{-1}\right)+nO\left(n^{-2}\right)O(n')O\left(n^{-2}\right)= o(1),$$

which implies that $\sqrt{n}|\hat{T} - T|$ is $o_p(1)$.

QED

THEOREM 1: Let the assumptions in Appendix A hold. The statistic $\hat{W}(\gamma^*, \hat{\theta})$ *generated in equation (9.) is asymptotically distributed as* $\chi^2(1)$ *under* H_o.

PROOF:

Under H_0:

$$y = f(x,\theta) + u$$

is the true data generating process. Define the function $\hat{T}(\gamma^*, \hat{\theta})$ as in the text. Note that $\hat{T}(\gamma^*, \hat{\theta})$ uses the estimated function $\hat{g}\left(x_i, \gamma_i, \hat{\theta}\right)$ and residuals $\hat{u}$. Now define a function that uses $\hat{g}\left(x_i, \gamma_i, \theta\right)$, the true residuals u, $b(\theta,\gamma)$, A and $f(x,\theta)$:

$$\text{(B.14)} \quad \sqrt{n}T\left(\gamma^*, \theta, \hat{g}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i \left\{\hat{g}\left(x_i, \gamma_i, \theta\right) - b\left(\theta, \gamma^*\right)A^{-1}\frac{\partial f(x,\theta)}{\partial \theta}\right\}$$

where

$$b(\theta,\gamma) = E\left[\frac{\partial f(x,\theta)}{\partial \theta}\hat{g}(x,\gamma,\theta)\right],$$

$\hat{g}\left(x_i, \gamma_i, \hat{\theta}\right)$ is defined in (8.) with $\hat{u}$, $\hat{g}\left(x_i, \gamma_i, \theta\right)$ is analogously define with u (the true residuals) and

$$A = E\left(\frac{\partial f(x,\theta)}{\partial \theta}\frac{\partial f(x,\theta)}{d\theta'}\right).$$

We show

$$\text{(i)} \qquad \sqrt{n}\ \left|\hat{T}\left(\gamma^*, \hat{\theta}\right) - T\left(\gamma^*, \theta_0, \hat{g}\right)\right| = O_p(1)$$

$$\text{(ii)} \qquad \sqrt{n}\ \left|T\left(\gamma^*, \theta_0, \hat{g}\right) - T\left(\gamma^*, \theta_0, g\right)\right| = o_p(1)$$

where

$$g = E\left(\hat{g}(x_i, \gamma_i, \theta)|x_i\right).$$

Note that $T(\gamma^*, \theta_0, g)$ is the sum of IID random variables with finite variance. This, the Levy Lindeberg Central Limit Theorem and the fact that γ^* converges to B implies that

$$\sqrt{n}\ T(\gamma^*, \theta_0, g) \xrightarrow{d} N\left(0, S^2(B)\right)$$

where

$$S^2(B) = E\left[u_i^2 \left\{ g(x_i, \gamma, \theta) - b(B) A^{-1} \frac{\partial f(x_i, \theta)}{\partial \theta} \right\} \right].$$

To show (i), we use the Mean Value Theorem

$$\text{(B.16)} \quad \sqrt{n}\ T\left(\gamma^*, \hat{\theta}\right) - \sqrt{n}\ T\left(\gamma^*, \theta_0, g\right) = \sqrt{n}\ \partial \frac{T\left(\gamma^*, \tilde{\theta}\right)}{\partial \theta} \left(\hat{\theta} - \theta_0\right)$$

where

$$\tilde{\theta} \in \left\{ \theta : (\theta - \theta_0)' (\theta - \theta_0) \le \left(\hat{\theta} - \theta_0\right)' \left(\hat{\theta} - \theta_0\right) \right\}.$$

Let $\hat{b}\left(\theta, \gamma^*\right)$ be defined as in the text. Then

$$\hat{b}\left(\tilde{\theta}, \gamma^*\right) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{g}\left(x_i, \gamma^*, \tilde{\theta}\right) \frac{\partial f\left(x_i \tilde{\theta}\right)}{\partial \theta} + \tilde{u}_i \frac{\partial g\left(x_i, \gamma^*, \tilde{\theta}\right)}{\partial \theta} \right\}$$

and

$$\tilde{u}_i = y_i = f\left(x_i, \tilde{\theta}\right).$$

By the triangle inequality

$$\text{(B.17)} \quad \underset{n \to \infty}{\text{plim}} \ \sup_{\gamma < B} \left| \hat{b}\left(\tilde{\theta}, \gamma^*\right) - E\left\{ -\hat{g}\left(x, \gamma^*, \theta_0\right) \frac{\partial f(x, \theta_0)}{\partial \theta} \right\} \right|$$

$$\le \underset{n \to \infty}{\text{plim}} \ \sup_{\gamma < B} \left\{ \left| \left(\frac{1}{n} \sum_{i=1}^{n} \hat{g}\left(x_i, \gamma^*, \tilde{\theta}\right) \frac{\partial f\left(x_i, \tilde{\theta}\right)}{\partial \theta} \right) + E\left(\hat{g}\left(x_i, \gamma^*, \theta_0\right) \frac{\partial f(x, \theta_0)}{\partial \theta} \right) \right| \right.$$

$$\left. + \left| \frac{1}{n} \sum \tilde{u}_i \frac{\partial \hat{g}\left(x_i, \gamma^*, \tilde{\theta}\right)}{\partial \theta} \right| \right\} = 0$$

By the uniform law of large numbers and assumption (A.1) and Lemma 2 the first term goes to zero, and the last term is bounded by:

$$\underset{n \to \infty}{\text{plim}} \ \sup_{\gamma < B} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_i \frac{\partial \hat{g}\left(x_i, \gamma^*, \tilde{\theta}\right)}{\partial \theta} - \frac{1}{N} \sum_{i=1}^{n} \hat{u}_i \frac{\partial \hat{g}\left(x_i, \gamma^*, \hat{\theta}\right)}{\partial \theta} \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i \frac{\partial \hat{g}\left(x_i, \gamma^*, \hat{\theta}\right)}{\partial \theta} \right| \right\}$$

The first term goes to zero from assumption (A.1) and

$$\text{plim}\left(\hat{\theta}-\theta_0\right)'\left(\hat{\theta}-\theta_0\right)=0,$$

and the second goes to zero by Lemma 2. Since

$$b(\theta_0,\gamma)=E\left(\hat{g}(x_j,\gamma^*,\theta_0)\frac{\partial f(x,\theta_o)}{\partial\theta}\right)$$

it follows from B.16 and B.17 that

(B.18) $$\underset{n\to\infty}{\text{plim}}\ \sup_{\gamma<B}\ \sqrt{n}T\left(\gamma^*,\hat{\theta}\right)-\sqrt{n}T\left(\gamma^*,\theta_0,g\right)+b(\theta_0,\gamma)\sqrt{n}\left(\hat{\theta}-\theta_0\right)=0.$$

From standard nonlinear least squares theory, it follows that

(B.19) $$\underset{n\to\infty}{\text{plim}}\left|\sqrt{n}\left(\hat{\theta}-\theta_0\right)-A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\frac{\partial f(x_i,\theta_0)}{\partial\theta}\right|=0$$

where

$$u_i=y_i-f(x_i,\theta_0).$$

Substituting the second argument on the right-hand side of (B.19) into ($\hat{\theta}$ - θ) in (B.18), we get

$$\underset{n\to\infty}{\text{plim}}\ \sup_{\gamma^*<B}\left|\sqrt{n}\ T\left(\gamma^*,\hat{\theta}\right)-\sqrt{n}T\left(\gamma^*,\theta_0,g\right)\right|=0$$

or

$$\sqrt{n}\ \left|\hat{T}\left(\gamma^*,\hat{\theta}\right)-T\left(\gamma^*,\theta_0,\hat{g}\right)\right|=O_p(1).$$

This completes part (i) of the proof. To prove part (ii), we apply the results of Lemma 3. Using the definitions of Lemma 4, we set

$$z_i=\{u_i,\ x_i\}$$

and

(B.20) $$p\left(z_i,z_j\right)=u_iu_jw_{ij}.$$

It follows from the definitions above that Lemma 3 can be applied to ($T(\gamma^*, \theta_o, \hat{g}) - T(\gamma^*, \theta_o, g)$). Therefore, to prove (ii) it suffices to show that

(a) $$E\left(u_iu_jw_{ij}\right)^2=o(n^{-1})$$

(b) $$E(u_iu_jw_{ij})(u_ku_hw_{kh})=o(n^{-2})$$

To show (a) note that

$$E\left(u_i^2u_j^2w_{ij}^2\right)<\sigma^4w_{ii}^2.$$

It is therefore sufficient to note that $E(w_{ii}^2)$ can be shown to be $O(n^{-2})$ by the method in Lemma 2.

To show (b) note that there are two cases. Either the four indexes i, j k, h have two sets of equalities, e.g. i=k and j=h, or they do not, e.g. $i \neq j \neq k \neq h$. In the first case, we know from (a) that $E(u_i u_j w_{ij})(u_k u_h w_{kh})$ is at most $O(n^{-2})$.

In the second case

$$E(u_i u_j w_{ij})(u_k u_h w_{kh}) = 0 .$$

Together these imply that (b) holds. This completes part (ii).

Since $\sqrt{n}T(\gamma^*, \theta, g)$ is the sum of an IID. sequence of random variables, we get our final results by applying the Lindberg Levy Central Limit Theorem to $\sqrt{n}T(\gamma^*, \theta, g)$.

PROPOSITION 1: Let {x,u} be a random vectors for some probability space where E(u)=0, and $E(u^2) = \sigma^2 < \infty$. *Denote* g(x) = E(u|x) *and* g(x) *is an* X *measurable continuous function. Then the solution to* (1.) *is proportional to* g(x)

Proof: We first show that g(x) is the solution to :

$$(17.) \quad \max_{h \in H(X)} \frac{E(uh(x))}{E(u^2 h(x)^2)}$$

By the conditions of the Proposition, u can be decomposed as

(18.) $u=g(x)+\varepsilon$

Since $E(g(x)\varepsilon)=0$,

$E(u^2)= E(g(x)^2) + E\varepsilon^2$

$= \sigma_{ux}{}^2 + \sigma_{\varepsilon\varepsilon}{}^2$

Substituting (3.) into the denominator of (2.) and taking expectations over ε , we rewrite (2.) as:

$$(19.) \max_{h \in H(X)} \frac{E(uh(x))}{E(g(x)^2 h(x)^2 + \sigma_{\varepsilon\varepsilon}^2 h(x)^2)}$$

without loss of generality, we normalize (4.) by letting

$$D^2 = \frac{E(h(x)^2)}{E(g(x)^2)}$$

and define

$\tilde{h}(x) = h(x)/D$.

Since $E\tilde{h}(x)^2) = E(g(x)^2)=\sigma_{ux}$, the problem in (1.) is equivalent to the following:

(20.) $$\max_{\tilde{h}\in\{H(X):E(\tilde{h}^2(x)=E(g^2(x))\}} \frac{E(u\tilde{h}(x))}{E(g(x)^2\tilde{h}(x)^2+\sigma_{\varepsilon\varepsilon}^2\sigma_{ux}^2)}$$

The fraction in (5.) is bounded above by 1, and with the constant $\sigma_{ux}^2\sigma_{\varepsilon\varepsilon}^2$ in the denominator, the solution to (5.) is the same solution to:

(21.) $$\max_{\tilde{h}\in\{H(X):E(\tilde{h}^2(x)=E(g^2(x))\}} E(u\tilde{h}(x))$$

We can now show that (6.) is maximized by $\tilde{h}(x)=g(x)$. By the Rao Blackwell Theorem,

$E(u-\tilde{h}(x))^2 \geq E(u-g(x))^2$

which implies

$-E(2u\tilde{h}(x)) + E(\tilde{h}(x)^2) \geq -E(g(x)^2)$ since $E(g(x)^2) = E(ug(x))$.

Since $E(\tilde{h}(x)^2) = E(g(x)^2)$.

$E(ug(x)) \geq E(u\tilde{h}(x))$.

Therefore, g(x) is the solution to (2.). To be the solution to (1.), we need to show that there exists no function $h(x) \in H(X)$ such that

$$\left|\frac{E(uh(x))}{E(u^2h(x)^2)}\right| > \frac{E(ug(x))}{E(u^2g(x)^2)}$$

and

$$E(uh(x)) < 0.$$

To show this note that if E(uh(x)) is negative, then

$E(u(-h(x)))= -E(uh(x)) >0$

$$\left|\frac{E(u(-h(x)))}{E(u^2h(x)^2)}\right| = \frac{E(u(-h(x)))}{E(u^2h(x)^2)} > \frac{E(ug(x))}{E(u^2g(x)^2)},$$

contradicting the fact that g(x) maximizes (2.).

QED

REFERENCES

Bierens, H.J.:, "Consistent Model Specification Tests," *Journal of Econometrics*, 20 (1982), 105-134.

-----: "Kernel Estimators of Regression Functions," in *Advances in Econometrics*, Fifth World Congress, vol. 1, ed. T. R. Bewley, 1987.

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

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

Manski, C: "Nonparametric Estimation of Expectations in the Analysis of Discrete Choice Under Uncertainty" in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Cambridge, Mass.: Cambridge University Press, 1991.

Ramsey, J.B.: "Tests for Specification Errors in Classical Least Squares Regression Analysis," *Journal of the Royal Statistical Society Series B*, 31 (1969), 350-371.

TABLE I

Simulations of (DGP1) and (DGP2)
Percent of 1,000 Trials Rejected at the Asymptotic 10 per cent and 5 per cent Significance Level

Sample Size	50		100		200	
Significance Level	10%	5%	10%	5%	10%	5%
DGP1						
Bierens	24.4	12.8	35.2	25.2	57.6	46.2
Kernel	74.1	66.8	92.4	90.1	100.0	99.7
DGP2						
Bierens	12.4	4.0	10.0	6.0	12.0	5.6
Kernel	11.3	6.7	10.8	5.3	11.1	6.0

ENDNOTES

[1]The authors would like to thank Tim Erikson and participants of seminars at the American Statistical Association, the Bureau of Labor Statistics and George Mason University. 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.

[2]For an example of a two-stage procedure for semi-parametrically estimating a binary choice model under uncertainty using estimated conditional expectations, see Manksi (1991).

[3]For example, see Bierens (1990) Theorem 2.

[4] Simulations suggest that our test is insensitive to the limit B.

AN IMPROVED NONPARAMETRIC TEST FOR MISSPECIFICATION OF FUNCTIONAL FORM

Ralph Bradley and Robert McClelland
US Bureau of Labor Statistics
2 Massachusetts Ave NE
Room 3105
Washington, DC 20212

In this paper, we develop a two-stage test for misspecification of a conditional mean that is similar to one recently developed by Bierens(1990). Our test uses the idea that in misspecified models the explanatory variables, X, better predict the residuals, $\hat{u}$, than their mean. Thus, an appealing function that detects misspecification is $E(\hat{u}|X)$. We estimate this with a kernel regression which is then a component of the second stage which tests for the nonzero correlation of $\hat{u}$ and the kernel estimate. By directly estimating the misspecification our approach should have good power while making fewer a priori restrictions than Bierens, while also helping researchers better understand the nature of any misspecification.

KEYWORDS: Consistent Tests, misspecification of functional form, conditional moment tests, kernel regression.

FIGURE 1 - Conditional Expectation of the Residual as a Function of X1 and X2

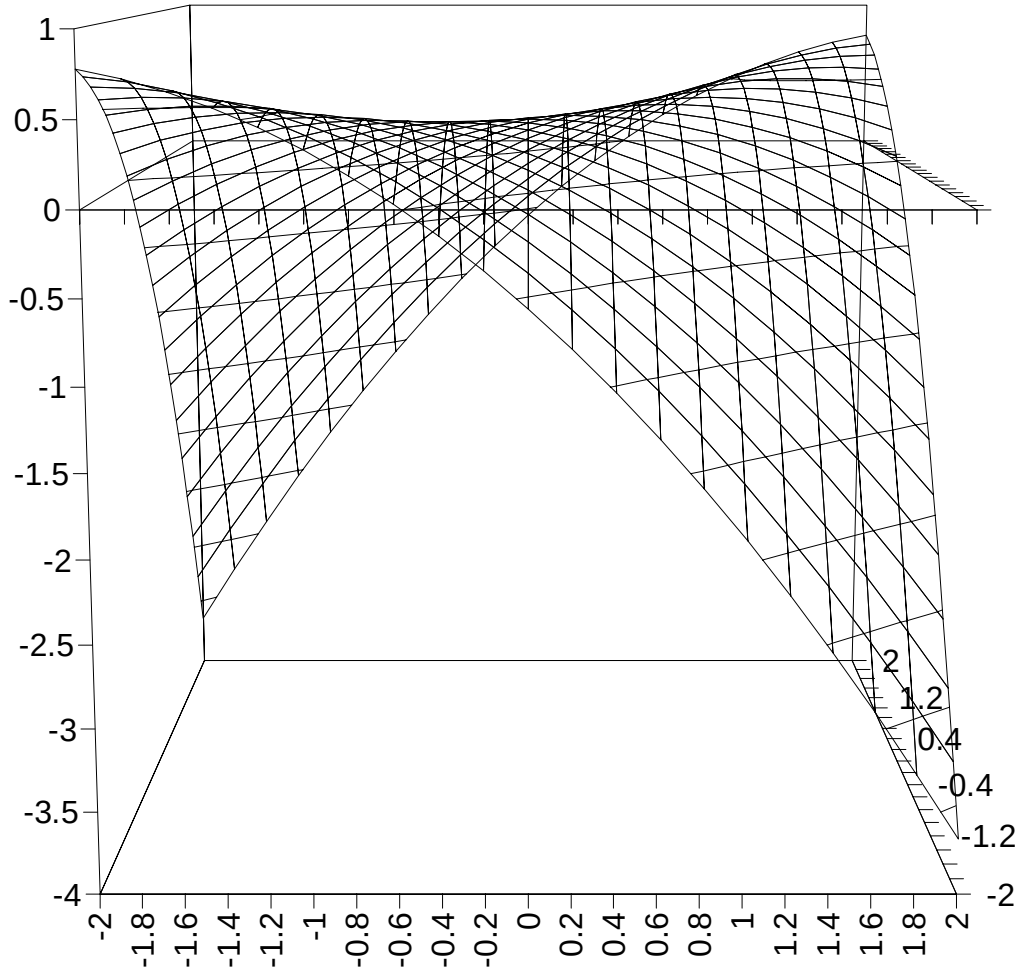

FIGURE 2 - A Comparison of the Conditional Expection of the Residual and the Kernel Estimate Plotted Against X1

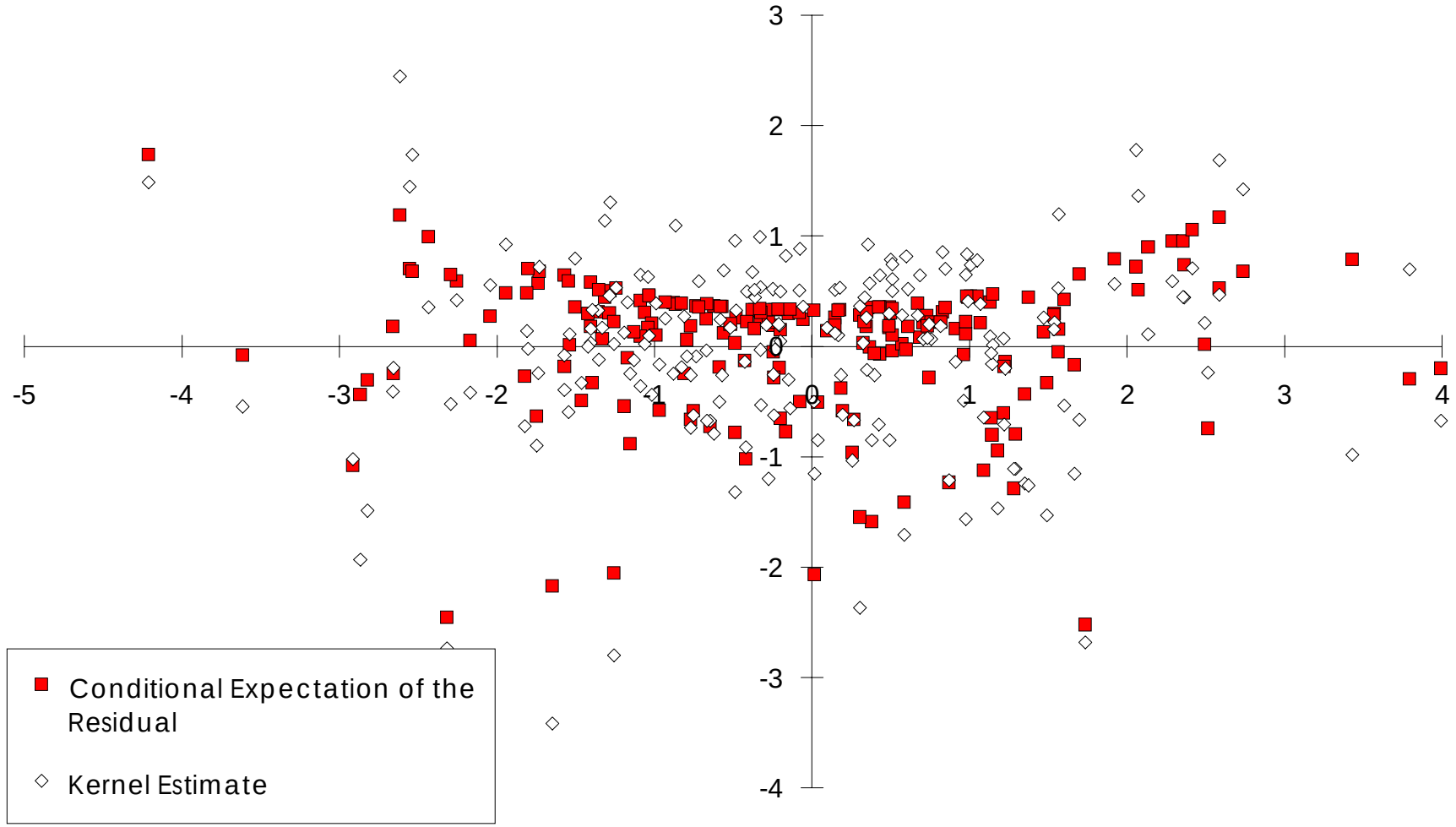

FIGURE 3 - A Comparison of the Conditional Expectation of the Residual and the Kernel Estimate Plotted Against X2

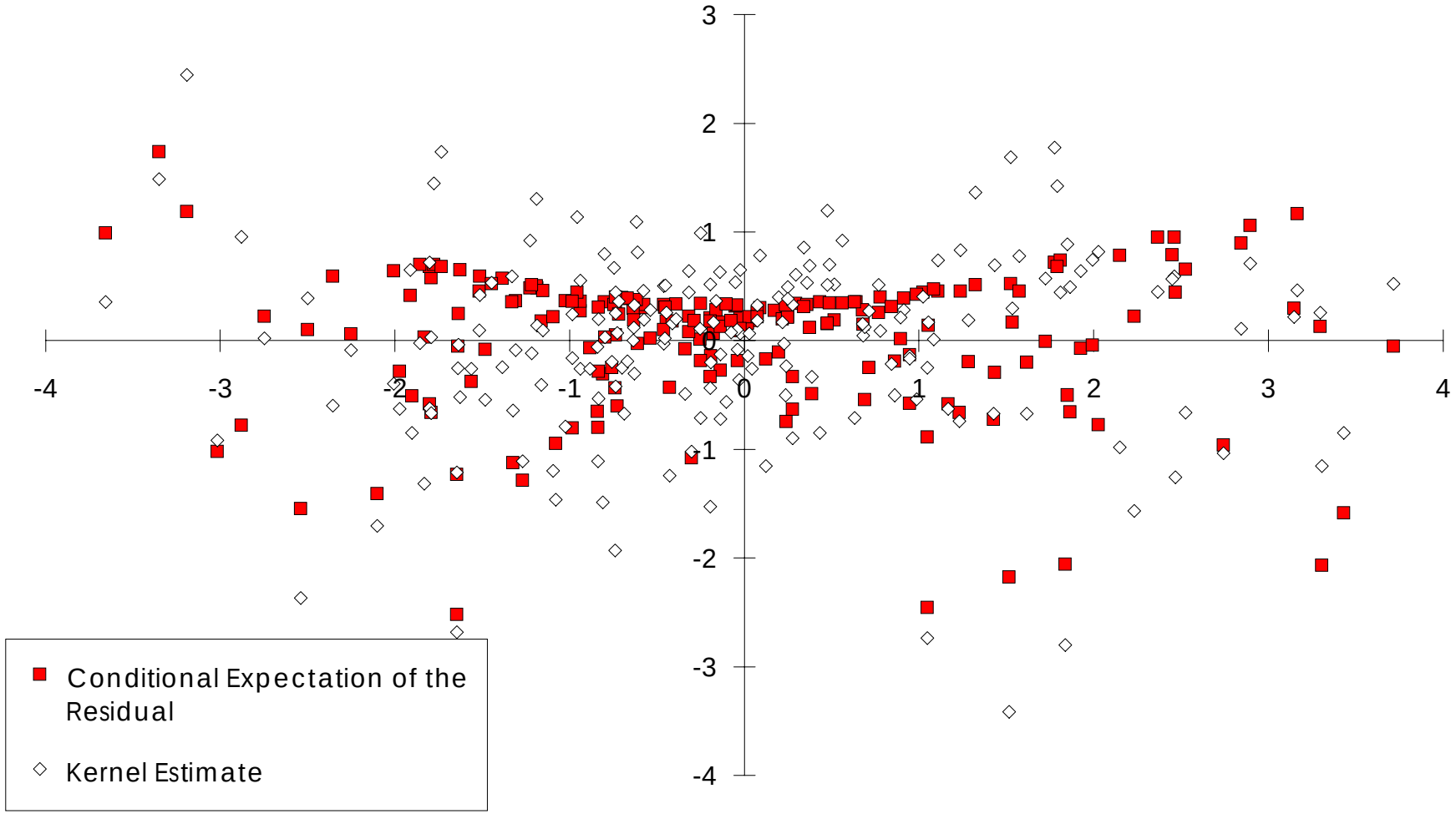

FIGURE 4 - A Comparison of the Conditional Expection of the Residual and the Kernel Estimate and Confidence Intervals, holding X2 Constant