A Kernel Test for Neglected Nonlinearity

by
Ralph Bradley
Robert McClelland
Room 3105
US Bureau of Labor Statistics
2 Massachusetts Ave. NE
Washington, DC 20212
(202)-606-6573

The authors would like to thank Tim Erickson, Jeff Wooldridge and participants of seminars at the American Statistical Association, the Bureau of Labor Statistics, George Mason University and the 1994 Summer Econometric Society meetings. 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.

1. Introduction

Recent tests of the nonlinearity of the functional form a regression model have focused on attempts to develop consistent tests that do not specify a parametric alternative and encompassing forms. These tests are a special case of the broader conditional moment tests for misspecification, such as that of Bierens (1990), that have greater power by restricting the specification to the linear form. Examples of this are the tests of Lee, White and Granger (1993) (LWG), which is based on the Bierens test and which detects omitted nonlinearity through the use of a 'neural net', and Wooldridge (1992), which is based on the Davidson-Mackinnon test and uses a Fourier expansion. Like the test of Bierens, these tests are consistent against deviations from the null.

In this paper we adapt the consistent test of Bradley and McClelland (1994) to a test of nonlinearity in a manner analogous to the adaptation of the Bierens test by LWG. Our test is based on the idea of Newey (1985) that if the model is not properly specified then the vector of explanatory variables, x, can better predict the residuals from a linear regression than the residual's mean. Thus, while many x-measurable functions, such as a neural net, can be used to detect neglected nonlinearities an appealing function to use is the conditional expectation of the residuals given x. We implement this idea 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 neural net test with the kernel estimate in the first stage replacing the net.[1] By directly estimating the nonlinearity with a conditional expectation function, the limit of our test over the sample size is the largest of all conditional moment tests. Because the kernel regression is an estimate of the misspecification, the test can be used to better understand the nature of any misspecification.

Another advantage to our test is the potential for increased finite sample power through a constrained cross-validation of the window width parameter. Cross-validating with a 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 the idea 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. Lewbel (1993), Wooldridge (1992), and Bierens (1990) discuss the

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

problem that a cross-validation type mechanism can overfit the alternative model under the null hypothesis, forcing their tests to acquire undesirable or unknown distributions. Thus, none of these studies use cross-validation.

In our study we cross-validate but address the problem of overfitting by resampling and limiting the window width to a compact set of permissible values. This resampling also solves the problem caused by the dependence inherent in the residuals produced by a regression that includes a constant term. This, plus the absence of bias for the kernel regression under the null hypothesis also allows the distribution of our statistic to converge to a $\chi^2(1)$ distribution.

Finally, we compare Monte Carlo simulations of the results of using our test and the neural net and Ramsey RESET test. Because we use the same models studied in Wooldridge (1992), we can also make limited comparisons with the sieve test developed in that paper. These simulations show that the kernel test is about as powerful as the neural net test in many simulations and more powerful than it or the RESET test in others. 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 describes the test of Bradley and McClelland (1994) and then shows how this can be adapted to a test for the detection of nonlinearity in a regression. Section 3 discusses results from Monte Carlo experiments that compares our test to the neural net and Ramsey RESET test and conclusions are in Section 4. The assumptions on the random vectors of dependent and independent variables (y,x) are listed in Appendix A and the proof of theorem 1 is in Appendix B.

2. A Misspecification Test

In this section we describe a test in the set of conditional moment tests that use a direct estimate of the misspecification as the weighting function on the estimated residuals. Relying upon the work of Bierens (1982, 1987, 1990) and described in detail in Bradley and McClelland (1994), this test is consistent against all deviations from the null hypothesis and has the greatest asymptotic power of all tests in set of conditional moment tests in the sense that the probability limit of the kernel estimate is the function that maximizes plim (W/n), where W is defined as an element in the set of conditional moment tests and n is the sample size.

We also increase the finite sample size power of our test over a fixed window width by using a cross-validated window width. Only two parameters need to be set before using the test: bounds on the window width and the size of the bootstrap sample. Lastly, we can use the estimate of the misspecification to gather insight into the nature of any problem detected by the test. This general specification test can also be modified to test the more restrictive null of linearity. By following the same intuition as that guiding LWG, we construct a Lagrange multiplier test of linearity that can be constructed from a simple R^2 statistic.

To focus our discussion upon the use of a direct estimate of the misspecification as a weighting system and to make our test as compatible as possible with the sieve test of Wooldridge, 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 $\Re\times\Re^k$. Further, we assume that $E(y^2) < \infty$ and that:

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

$E(y|x) = f(x,\theta).$

Suppose we use the functional form $f(x,\theta)$ to estimate $E(y|x)$ and define $\hat{u}_i = y_i - f(x_i, \hat{\theta})$. A direct estimate of the misspecification is then $g(x) \equiv E(\hat{u}|x)$. Under H_o, we assume that the true value of θ which we denote as θ_o satisfies:

(1.) $\theta_o = \text{argmin}_{\theta\in\theta} \; E([y_i - f(x_i,\theta)]^2)$ 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. If H_o in (1.) is correct, then

(2.) $\text{plim}\; E(\hat{u}*h(x)) = 0$

for all $h(x) \in H(X)$, the set of x-measurable functions on X, the domain of x. Under H_1, there exists some function

(3.) $\text{plim}\; E(\hat{u}*h(x)) > 0.$

A reasonable set of tests for misspecification would then be

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

The Rao-Blackwell theorem tells us that under H_1 :

(4.) $\lim E(\hat{u}-g(x))^2 < \lim E(\hat{u} - E(\hat{u}))^2$

which implies $\lim E(\hat{u}g(x))=\lim E(g(x)^2) > 0$. Under H_o, since $\text{plim}\;\hat{\theta}=\theta_o$, the inequality in (4.) becomes an equality so that $\lim E(\hat{u}E(\hat{u}|x)) = 0$. The function g(x) therefore satisfies both conditions (2.) and (3.) and our test uses a nonparametric estimator of g(x). The features in (1.) through (4.) suggest the following four steps to test the null hypothesis 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 $g(x)=E(\hat{u}_i|x_i)$ with a kernel regression by selecting a random subsample with replacement of size $n' = \mathrm{int}(n^{3/2}\gamma_n{}^{k/2})$ 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'. We then define the cardinal 'bootstrapping' variable S(A) as the number of occurrences of event A, 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:

(5.)
$$\hat{g}_i(x_i,\gamma_i) = \frac{\sum_{'} S(j \in N_i^{'})\hat{u}_j K(\frac{x_i - x_j}{\gamma_i})}{\sum S(j \in N_i^{'})K(\frac{x_i - x_j}{\gamma_i})}$$

and γ_i is the bandwidth of the kernel function $K(\cdot)$ with the following properties:

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

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

(P.3) $\int \prod_j^k u_j^{i_j} du_j = 0$ for $\sum_j^k i_j < m$, where m is the kernel order such that $n\gamma_n{}^k \rightarrow \infty$ and $n\gamma_n{}^{2m+k/2} \rightarrow 0$.

(P.4) $0 = \arg\max_u K(u)$.

(iv) Calculate the statistic:

(6.)
$$\hat{W}(\gamma^*, \hat{\theta}) = n\, \hat{T}^2(\gamma^*, \hat{\theta})/\hat{s}^2(\gamma_i),$$

where

(7.)
$$\hat{T}(\gamma^*, \hat{\theta}) = (1/n)\left[\sum_{i=1}^{n} \hat{u}_i \hat{g}_i(x_i, \gamma_i)\right],$$

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

and $\hat{s}^2(\gamma_i)$ is a consistent variance estimate.

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

$$\gamma_i = \arg\max_{C_n < \gamma < B_n < \infty}\left[\sum_{k \in Ni'} \hat{u}_k - \hat{\hat{g}}_{-k}(x_k, \gamma))^2\right]$$

where

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

and the bound B_n to have two properties:

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

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

(P.7) $C_n = O(n\gamma_n^{2m+k/2})$ for $C_n < B_n$ for all n.

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. 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. It also prevents the window width from becoming so large that the test statistic degenerates to a constant. The lower bound C_n is used to prevent overfitting under the null.

$\hat{W}(\gamma^* ,\hat{\theta})$ is essentially a Wald Test for the restriction that $E(\hat{u}[E\hat{u}|x]) = 0$. The following theorem is the main result of Bradley and McClelland (1994):

*Theorem 1 (*Bradley and McClelland (1994))*: Let the assumptions in Appendix A hold. The statistic $\hat{W}(\gamma^* ,\hat{\theta})$ generated in equation (6.) is asymptotically distributed as $\chi^2(1)$ under H_o.*

This theorem shows that the limiting distribution of $\hat{W}(\gamma^* ,\hat{\theta})$ under the null is a χ^2 distribution with one degree of freedom. Given the properties (P.1) through (P.4) Bradley and McClelland (1994) test is consistent against all alternative hypotheses and additionally that $\hat{W}(\gamma^* ,\hat{\theta})$ is the most asymptotically powerful test in the set $\{W_h\}$. The reason for this 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 (6.) and (7.) is an estimator of $E\{\hat{u}\ E(\hat{u}|x)\}$ and therefore the numerator in equation (5.) uses an estimator that maximizes the value of $E\{\hat{u}h(x)\}$.

We now turn to a possible complication. 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, partially because there is finite sample dependence among observations of $\hat{u}_i$ arising from the first moment restriction, $\sum_{i=1}^{n} \hat{u}_i = 0$ 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 from the literature on U-statistics 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 of 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).

3. The Kernel Test

Given the statistic $\hat{W}(\gamma * ,\hat{\theta})$, we can construct a test for neglected nonlinearity in the same manner that Lee, White and Granger (1993) use the results of Bierens (1990) to construct the neural net test for nonlinearity. The null that we wish to test is now more restrictive than (1). For the linearity case, the assumptions in Appendix A. ensure that under H_o plim $\hat{\theta}$-θ_o= 0 and under H_1 plim $\hat{\theta}$ exists.

(8.) H_o: there exist $\theta \in \Theta \subset \Re^k$ such that $\Pr(E(y|x)=x'\theta) = 1$

(9.) H_1: $\Pr(E(y|x)=x'\theta) < 1$ for all $\theta \in \Re^k$

Here the null is that the true conditional expectation function is linear, which allows us to offer a Lagrange Multiplier version of the statistic $\hat{W}(\gamma * ,\hat{\theta})$. To form this statistic, we derive $\hat{\theta}$ from an ordinary least squares (OLS) regression of y_i on x_i. We calculate the estimated residuals, $\hat{u}_i$, and we estimate the kernel density $\hat{g}_i(x_i,\gamma_i)$ as outlined in (5.). Finally, we regress $\hat{u}_i$ on x_i and $\hat{g}_i(x_i,\gamma_i)$.

In order to use the statistic, we need the following theorem for x-measurable functions h(x):

Theorem 1)

Under the linear null hypothesis in (8), and homoskedasticity, define the statistic W_n *as follows*:

$$W_n = n \frac{\left[h' \hat{u}_i\right]^2}{\hat{u}_i' \, h(I - x(x' x)^{-1}x)h' \, \hat{u}_i}.$$

Then $W_n - nR^2 \xrightarrow{p} 0$*, where* R^2 *is the coefficient of determination of* $\hat{u}_i$ *regressed on* x *and* h(x) *where* h *is the column vector of* $h(x_i)$.

Proof *See Appendix B.*

Because $\hat{g}_i(x_i,\gamma_i)$ is an x-measurable function, we can apply this theorem to it. We then have a test statistic that avoids explicit computation of $\hat{W}(\gamma * ,\hat{\theta})$ because

(10.) $nR^2 \xrightarrow{d} \chi^2(1)$

where R^2 is the uncentered squared multiple correlation from an ordinary least squares regression of $\hat{u}_i$ on x_i and $\hat{g}_i(x_i,\gamma_i)$.

If H_1 is true, then the probability limit of the regression coefficient on $\hat{g}_i(x_i,\gamma_i)$ should be one. If one rejects the null that the coefficient is equal to one then there is evidence that $\hat{g}_i(x_i,\gamma_i)$ is not picking up the misspecification E(u|x). Since $\hat{g}_i(x_i,\gamma_i)$ is an estimator for $E(\hat{u}|x_i)$, it can provide direction in which the investigator can go to better specify the model if the coefficient is not significantly different from one. Although not proved in this article, under misspecification the OLS coefficient on $\hat{g}_i(x_i,\gamma_i)$

converges in probability to one. If the coefficient is not significantly different from one, then one can look at the confidence intervals of derivatives of $\hat{g}(x, \gamma_i)$ to determine the additions to the linear model that need to be added to achieve a correct specification.

As with the more general test we need to select two parameters: bounds on the bandwidth and the bootstrap sample size. For the simulations in the next section, we set the upper bandwidth bound to 19.5 + 1/n and the lower bound to $1.5n^{-1/\delta}$ where δ satisfies the inequality, $k<\delta<m-k/2$. These bounds on the cross validation serve to prevent overfitting under the null. Additionally, these limits force the cross validated bandwidth to converge to zero at a rate where the asymptotic bias of the kernel estimate will go to zero. The bootstrap sample size is trunc(.85n) and the estimated variance in the denominator equals the order of our statistic in the numerator so that the statistic does not degenerate under the null. In addition, we use the Epanechnikov kernel, which satisfies the properties P.1, P.2 and P.4..

4. Monte Carlo Experiments

We now briefly consider two Monte Carlo experiments discussed in Wooldridge (1992), allowing us to make some comparisons with his sieve test. In each, we compare W_n with the neural net test of LWG using ten nodes and the Ramsey RESET test using three powers of the estimated conditional expectation of y. We consider sample sizes of 50, 100 and 200 and a variety of misspecifications. We also show how the function $\hat{g}(x, \gamma_i)$ can be used to gather information about the nature of the misspecification.

Both the sieve test and the neural net test are also robust tests. Our test is more closely related to the neural net test in that both are variations of the Bierens 1990 test and both are based on the asymptotic distribution of nR^2 of the residuals regressed on the X values and an additional non-linear X-measurable function. Wooldridge uses the Davidson-Mackinnon test framework with a linear regression and a nonparametric sieve regression. In order to achieve consistency under the null his truncation parameter has to expand at a rate that is an inverse function of the rate at which $\hat{y}$ converges to y. Because Wooldridge uses the entire sample for his sieve estimation, this bound on the expansion rate prevents him from cross-validating his truncation parameters.

The neural net test takes the inner product of the column vectors of x and q independent random k-vectors. These q random vectors are also known as "nodes". The q products $\tilde{x}$ are then transformed by the logistic cumulative density function $(1+e^{-\tilde{x}})^{-1}$. To reduce correlation among the q transformed products, only the first two principle components of the transformed products are taken. These components are then regressed on û. The resultant test statistic formed by the R^2 multiplied by the number of observations has an asymptotic $\chi^2(2)$ distribution.

The RESET test is designed to detect misspecifications of E(y|x) by modeling the misspecification as a power expansion of E(y|x). While the test is easy to calculate, its power depends upon the degree to which the misspecifications are well approximated by a low-order expansion. We implement the RESET test by checking the joint significance of α_2 , α_3 , α_4 in a regression of the form

$$y_i = x_i\theta + \sum_{j=2}^{4} \alpha_j (\hat{y}_i|x_i)^j + e_i.$$

Under the null hypothesis of a correct specification, all α_i's should equal zero. We use the LWG version of the test so that it has a $\chi^2(3)$ distribution under the null.

The data generating process for both experiments is essentially the same. Let v_1 and v_2 be two independent random variables with uniform distributions on $[0, 2\pi]$ and define

$z_1 = v_1$

$z_2=0.5v_1 + 0.5v_2$.

Then the vector of independent variables are defined as

$x=[1, z_1, z_2]$

and the coefficient vector is

$\theta=[1,1,1]'$.

Given homoskedastic residuals u drawn from a standard normal distribution, the data generating process under the null hypothesis is:

$y_i = x_i\theta + u_i$.

The first experiment considers the model:

(DGP1) $y_i = x_i\theta + \gamma(x_i\theta)^2 + u_i$.

Table 1 shows that all three tests are effective in detecting the misspecification. With 50 observations all three tests reject almost 100% of the iterations with γ equal to 0.20 or -0.20. With 100 observations the three tests reject when γ is equal to 0.15. Because the RESET actually nests the true misspecification, it should and does have the greatest power. Comparing the kernel test and the neural test to the RESET is worthwhile here because it represents the upper bound on power. Both the kernel and neural net tests seem to reject almost exactly the same number of cases although a cursory examination of the rejection patterns shows that different simulations are being rejected. As the sample size increases, the rejection rates increase towards 1.000 rejection rate.

One difficulty with the kernel test appears to be its high rejection rate under the null hypotheses (γ=0). This rate reflects the finite sample bias of the kernel estimate. As the sample size increases, however, the rejection rate decreases somewhat and should continue to decrease to the proper rejection rates.

These results may also be compared with the results of Wooldridge (1992) using the sieve test. With 100 observations and γ equal to 0.1, the sieve test rejects 0.576 and 0.418 of the cases at the ten and five percent points, respectively. With the same number of observations and γ equal to -0.1 the test rejects 0.999 of the cases at both the ten and five percent points. This shows that for this model the power of the sieve test is more sensitive to the direction of the misspecification than any of the three tests considered here.

The second experiment considers the model:

(DGP2) $y_i = (x_i\theta)^{\lambda} + u_i.$

For values of λ we consider a range of values from 0.5 to 1.5 and for -0.5. The results in table 2 are somewhat ambiguous. For values of λ between 1.2 and 1.5 the kernel and neural tests move together, although the neural test decreases in power from 1.2 to 1.3 more rapidly than the kernel test. As expected, the RESET test performs at a progressively greater rate as λ rises. When λ is equal to 1.5 with 100 observations we can make a direct comparison with the sieve test. Comparing the nearly uniform rejection from all three tests against the rejection rates of 0.804 and 0.693 at the 10% and 5% points shows tha t other nonparametric tests have greater power against this alternative.

For values between 0.8 and 0.5 there is little change in power either as the sample size increases or as λ decreases. The neural net and RESET tests appear to have essentially no power to detect the misspecification while the kernel test shows some power. When λ is equal to 0.5 with 100 observations we can make another comparison with the sieve test. Comparing the low power of all three tests against the rejection rates of 0.24 and 0.166 at the 10% and 5% points shows that the sieve test performs approximately as well as the kernel test and outperforms the neural and RESET test. It would be of interest to know how the sieve test exhibits the same constancy of rejection rates over the λ range of 0.8 to 0.5 as the three tests show in table 2.

Finally, the negative value of -0.5 for λ reveals nearly the same rejection rates as for 0.5. Again the neural and RESET test appear to have little power for any of the sample sizes. Interestingly, the RESET test could not be performed for sample sizes of 100 because of repeated failures of the matrices to invert at the second stage of the test. This appears to be due to the creation of a variable $\hat{y}_i$ that has no essentially no variation. In contrast, the sieve test has much greater power than any of the tests, rejecting 0.443 and 0.307 percent of the cases at the 10% and 5% points and this power appears to grow with the sample size, rejecting 0.646 and 0.489 percent of the cases when the sample size increases to 300.

Although these simulations shed some light on the relative advantages of the tests for the misspecifications in DGP1 and DGP2, an additional advantage of the kernel test is that an estimate of the misspecification is preserved. We can show how this information, obtained with the estimation of $\hat{g}(x, \gamma_i)$, can be used to help the researcher understand the nature of the misspecification. To do this, we plot the first regression from DGP2 with 200 observations and λ equal to 0.6. Figure one shows the kernel estimate and the true conditional expectation $\hat{u}_i$ given x_1 and holding x_2 constant at the mean. We also include the 95% confidence intervals to show how closely $\hat{g}(x, \gamma_i)$ follows $\hat{u}_i$.

As the figure shows, $\hat{u}_1$ has a concave form consistent with modeling $(x_i\theta)^{0.6}$ as $(x_i\theta)$. The kernel estimate $\hat{g}(x, \gamma_i)$ also have the same general shape, although it is much less smooth than $\hat{u}_1$. Still, this example shows how given only $\hat{g}(x, \gamma_i)$ and the confidence limits, a researcher may obtain information about the manner in which the proposed model differs from the true relationship.

4. Conclusion

In this paper, we adapt the test of Bradley and McClelland (1994) to a test of nonlinearity. Using the idea of Newey (1985) that if the model is not properly specified then the vector of explanatory variables can better predict the residuals from a linear regression than the residual's mean we use the conditional expectation of the residuals given the regressors. We implement this idea by estimating the residuals as a function of the regressors with a kernel regression. By directly estimating the nonlinearity with a conditional expectation function, the limit of our test over the sample size should be the largest of all conditional moment tests.

Our test also allows researchers to increase power through cross-validation of the window width. We avoid the problems of cross-validation discussed by other authors by resampling. This resampling also solves the problem caused by the dependence inherent in the regression residuals.

In several Monte Carlo simulations we compare our kernel test to that of the neural net test, the Ramsey RESET test and the sieve test. Although the choice of models essentially guarantees that the RESET test will outperform others in most simulations, our test compares favorably with the neural net test in most simulations and appears to be somewhat more powerful that either the neural net or the RESET tests in others. The sieve test does not appear to have as much power, with one the notable exception.   Lastly, we use the results from an arbitrary regression to show how the kernel regression estimated for the test can be used by the researcher to gather information about the nature of any specification errors detected.

TABLE I
Simulations of DGP1 for Various Levels of γ
Percent of 1,000 Trials Rejected at the Asymptotic 10 per cent and 5 per cen
Significance Level

Sample Siz	50		100		200	
Significance Level	10%	5%	10%	5%	10%	5%
γ=0.20						
Kernel	0.970	0.957				
Neural	0.952	0.932				
RESET	1.000	1.000				
γ=0.15						
Kernel	0.900	0.872	0.998	0.998		
Neural	0.912	0.870	0.985	0.973		
RESET	0.997	0.994	1.000	1.000		
γ=0.10						
Kernel	0.669	0.614	0.916	0.899	0.997	0.996
Neural	0.757	0.654	0.916	0.892	0.980	0.972
RESET	0.922	0.868	0.998	0.995	1.000	1.000
γ=0.05						
Kernel	0.334	0.251	0.525	0.472	0.742	0.681
Neural	0.372	0.243	0.553	0.421	0.795	0.706
RESET	0.454	0.327	0.712	0.594	0.952	0.901
γ=0.00						
Kernel	0.163	0.101	0.157	0.099	0.151	0.09
Neural	0.109	0.053	0.087	0.044	0.095	0.048
RESET	0.119	0.051	0.113	0.058	0.098	0.049
γ=-0.05						
Kernel	0.302	0.218	0.483	0.417	0.748	0.694
Neural	0.351	0.232	0.528	0.413	0.815	0.727
RESET	0.372	0.249	0.658	0.543	0.954	0.909
γ=-0.10						
Kernel	0.648	0.578	0.922	0.901	0.994*	0.994
Neural	0.740	0.629	0.922	0.883	0.980*	0.972
RESET	0.860	0.785	0.992	0.987	1.000*	1.000
γ=-0.15						
Kernel	0.893	0.869	0.991	0.989		
Neural	0.915	0.875	0.974	0.963		
RESET	0.997	0.989	1.000	1.000		
γ=-0.20						
Kernel	0.976	0.969				
Neural	0.953	0.933				
RESET	1.000	1.000				

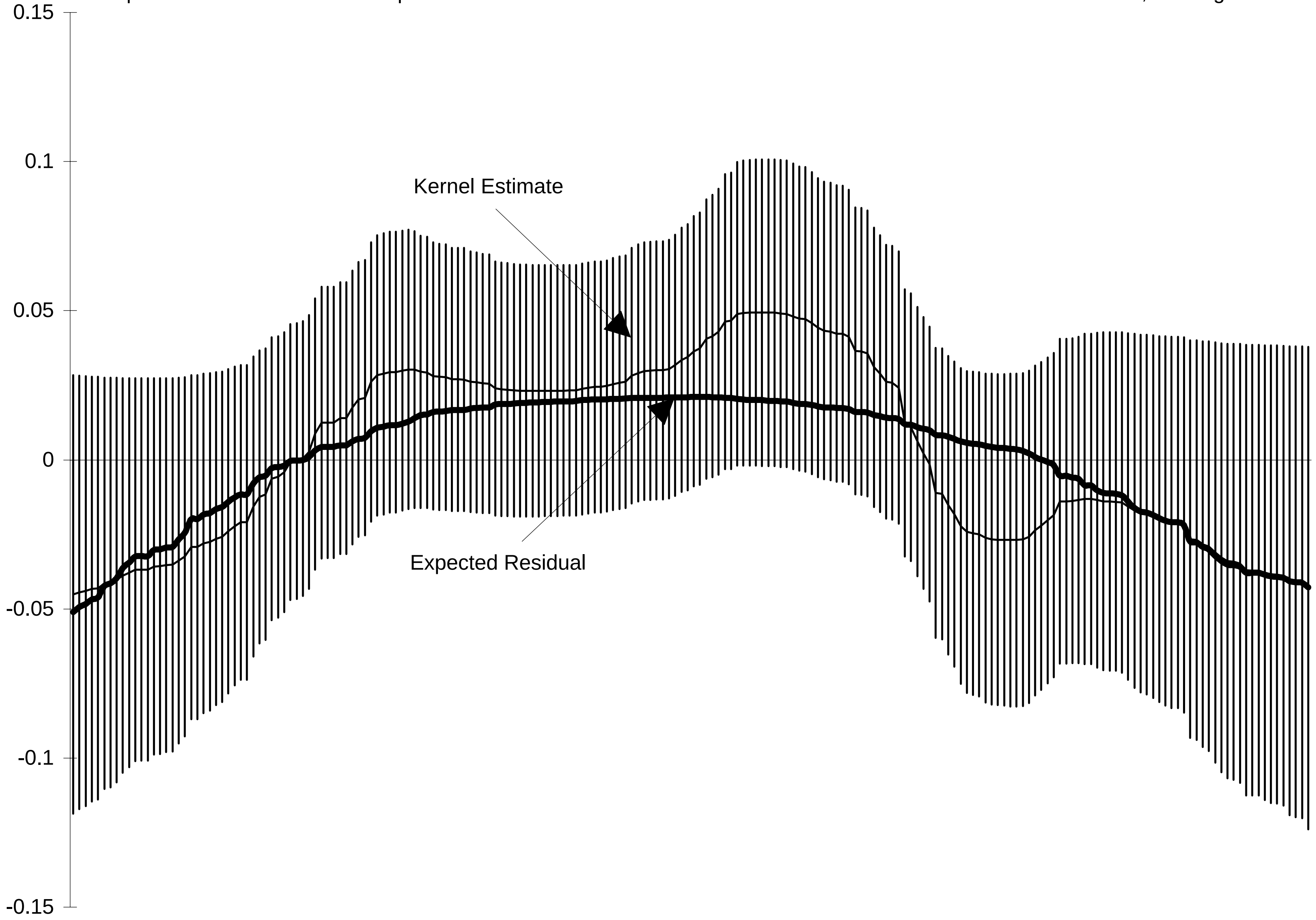
**Only 500 iterations are used in this simulation*

TABLE 2
Simulations of DGP2 for Various Levels of λ
Percent of 1,000 Trials Rejected at the Asymptotic 10 per cent and 5 percent Significance Level

Sample Size	50		100		200	
Significance Level	10%	5%	10%	5%	10%	5%
λ=1.5						
Kernel	0.896	0.875	0.995	0.995	1.000	1.000
Neural	0.921	0.884	0.980	0.971	0.998	0.992
RESET	0.993	0.987	1.000	1.000	1.000	1.000
λ=1.4						
Kernel	0.642	0.579	0.896	0.882	0.996*	0.994
Neural	0.717	0.596	0.908	0.853	0.976*	0.970
RESET	0.857	0.763	0.989	0.980	1.000*	1.000
λ=1.3						
Kernel	0.334	0.269	0.589	0.525	0.794	0.777
Neural	0.381	0.272	0.645	0.529	0.877	0.813
RESET	0.476	0.345	0.752	0.654	0.955	0.922
λ=1.2						
Kernel	0.267	0.198	0.290	0.237	0.418	0.339
Neural	0.192	0.118	0.254	0.155	0.408	0.304
RESET	0.221	0.134	0.291	0.194	0.486	0.368
λ=0.8						
Kernel	0.209	0.141	0.221	0.151	0.223	0.153
Neural	0.142	0.077	0.135	0.070	0.125	0.072
RESET	0.141	0.067	0.141	0.073	0.130	0.072
λ=0.7						
Kernel	0.196	0.139	0.205	0.142	0.280	0.202
Neural	0.119	0.049	0.135	0.068	0.147	0.064
RESET	0.115	0.051	0.141	0.070	0.145	0.083
λ=0.6						
Kernel	0.214	0.143	0.225	0.158	0.243	0.161
Neural	0.122	0.066	0.143	0.081	0.147	0.064
RESET	0.132	0.068	0.135	0.065	0.148	0.078
λ=0.5						
Kernel	0.200	0.126	0.222	0.147	0.229	0.167
Neural	0.114	0.052	0.111	0.063	0.140	0.077
RESET	0.116	0.063	0.108	0.054	0.138	0.073
λ=-0.5						
Kernel	0.189	0.122	0.218	0.153	0.230	0.165
Neural	0.114	0.051	0.106	0.061	0.101	0.057
RESET	0.118	0.058			0.111	0.066

**Only 500 iterations are used in this simulation.*

Figure 1
A Comparison of the Conditional Expectation of the Residual and the Kernel Estimate and 95% Confidence Interval, Holding X2 Constant



REFERENCES

Bradley, R. and R. McClelland: “An Improved Nonparametric Test for Misspecification of Functional Form,” manuscript, (1992)

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.

Lewbel, A.: "Consistent Tests with Nonparametric Components", manuscript (1993), University of Brandeis

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

Newey, W.K.: "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica* 53 (1985), 1047-1070

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.

Robinson, P.: "Root-N Consistent Semiparametric Regressions", *Econometrica*, 56 (1988), 931-954

White, H.: Asymptotic Theory for Econometricians. Orlando, Florida Academic Press, 1984

Wooldridge, J.M.: "A Test for Functional Form Against Nonparametric Alternatives", *Econometric Theory*, 8 (1992), 452-475

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 $i_1,i_2 = 1,..,m$,

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

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

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

(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

Proof of Theorem 1:

There are three steps:

i) show that W_n - Ω_n = 0

where Ω_n is the Wald test for γ =0 in model Y = θX + γh(x) + ε

ii) show that $\Omega_n - \Lambda_n \xrightarrow{p} 0$

where Λ_n is the Lagrange multiplier test for the above model

iii) show that Λ_n - nR^2 = 0

i) In the linear model we may write W_n as

$$W_n = n\frac{\left[h'\,\hat{u}\right]^2}{\hat{u}'\,h\left[I - X(X'\,X)^{-1}X'\right]h'\,\hat{u}}$$

$$= n\frac{\hat{u}'\,h\left[h'\left[I - X(X'\,X)^{-1}X'\right]h'\right]^{-1}h'\,\hat{u}}{\hat{u}'\,\hat{u}}$$

where h= $[h(x_1),...,h(x_n)]'$.

This is just a Wald test for γ=0:

$\hat{\gamma}\Gamma_n^{-1}\hat{\gamma}$,

where $\hat{\gamma}$ is an OLS estimator of γ, using x and h,

$\Gamma_n = R'\,(Z'\,Z\,/\,n)^{-1}R\,\dfrac{\hat{u}'\,\hat{u}}{n}$,

R=$[0_1,...,0_k, 1]'$

and

Z = [X, h].

ii) See White (1984)

iii) We begin with the Lagrange Multiplier statistic for the null γ=0. From the above,

E(Y|X) = θX + γh.

Let B=[θ, γ]'.

Then the null hypothesis is

RB=0,

where

R = [0_1,...0_k 1].

Let

1) $\ddot{\lambda} = 2[R(Z'\,Z\,/\,n)^{-1}R'\,]^{-1}R\hat{B}$

2) $\ddot{B} = \hat{B} - (Z'\,Z\,/\,n)^{-1}R\ddot{\lambda}\,/\,2$

$$= [\hat{\theta} \quad 0]'$$

The Lagrange Multiplier statistic is:

4) $L_n = n\ddot{\lambda}_n'\,\hat{\Lambda}_n^{-1}\ddot{\lambda}_n \longrightarrow \chi^2(1)$,

where

5) $\hat{\Lambda}_n = 4(R'\,(Z'\,Z\,/\,n)^{-1}R)^{-1}R(Z'\,Z\,/\,n)^{-1}\ddot{V}_n(Z'\,Z\,/\,n)^{-1}R'\,(R'\,(Z'\,Z\,/\,n)^{-1}R)^{-1}$,

where $\ddot{V}$ is an unbiased estimate of the variance of (Z'u/$n^{1/2}$).

We can simplify L_n to

6) $R\hat{B} = R(Z'\,Z\,/\,n)Z'\,Y\,/\,n$

We know that

7) $R\ddot{B} = 0$

Adding 7) to the righthand-side of 6):

8) $R\hat{B} = R(Z'\,Z\,/\,n)Z'\,(Y - Z'\,\ddot{B})\,/\,n$

$$= R(Z'\,Z\,/\,n)^{-1}Z'\,\hat{u}\,/\,n$$

Substituting the righthand-side of 8) into 1), we get

$\ddot{\lambda} = 2[R(Z'\,Z\,/\,n)^{-1}R'\,]^{-1}R(Z'\,Z\,/\,n)^{-1}Z'\,\hat{u}\,/\,n$.

Partitioning Z'Z^{-1}, we have.

$$(Z'\,Z)^{-1} = \begin{bmatrix} \left(X'\left(I - h(h'\,h)^{-1}h'\right)X\right)^{-1} & -E^{-1}X'\,h(h'\,h)^{-1} \\ -D^{-1}h'\,X(X'\,X)^{-1} & \left(h'\left(I - X(X'\,X)^{-1}X'\right)h\right)^{-1} \end{bmatrix},$$

where

$$E = X'(I - h'(h'h)^{-1}h)^{-1}h)X$$

$$D = h'(I - X(X'X)^{-1}X')^{-1}X)h.$$

Then

$$R(Z'Z)^{-1}R' = \left(h'\left(I - X(X'X)^{-1}X'\right)h\right)^{-1}$$

$$R(Z'Z)^{-1} = \left[-D^{-1}X(X'X)^{-1} \quad \left(h'\left(I - X(X'X)^{-1}X'\right)h\right)^{-1}\right]$$

Then

$$\left(R(Z'Z)^{-1}R'\right)^{-1}R(Z'Z)^{-1} = \left(h'\left(I - X(X'X)^{-1}X'\right)h\right)\left[-D^{-1}X(X'X)^{-1} \quad \left(h'\left(I - X(X'X)^{-1}X'\right)h\right)^{-1}\right]$$

$$= \left[-h'\left(I - X(X'X)^{-1}X'\right)hD^{-1}h'X(X'X)^{-1} \quad I\right]$$

$$= \left[-h'X(X'X)^{-1} \quad I\right].$$

Further,

$$\left(R(Z'Z)^{-1}R'\right)^{-1}R(Z'Z)^{-1}Z' = h'\left(I - X(X'X)^{-1}X'\right).$$

Then

9) $$\ddot{\lambda} = 2h'\left(I - X(X'X)^{-1}X'\right)\hat{u}/n$$

$$= 2h'\hat{u}/n$$

Given homoskedasticity, $\ddot{V}$ in (5) can be simplified as

10) $$\ddot{V} = \ddot{\sigma}(Z'Z/n)$$

$$= (\hat{u}'\hat{u}/n)(Z'Z/n)$$

Using (10) and substituting allows us to simplify (5) as

11) $$\hat{\Lambda}_n = 4(\hat{u}'\hat{u}/n)(Z'Z/n)^{-1}$$

Substituting (9) and (11) into (4), then

$$L_n = n\hat{u}'Z(Z'Z)^{-1}Z'\hat{u}/\hat{u}'\hat{u} = nR^2.$$