TESTING THE SIGNIFICANCE OF CATEGORICAL PREDICTOR VARIABLES IN NONPARAMETRIC REGRESSION MODELS

JEFF RACINE
DEPARTMENT OF ECONOMICS, SYRACUSE UNIVERSITY
SYRACUSE, NY USA 13244-1020

JEFFREY HART
DEPARTMENT OF STATISTICS, TEXAS A&M UNIVERSITY
COLLEGE STATION, TX USA 77843-4228

QI LI
DEPARTMENT OF ECONOMICS, TEXAS A&M UNIVERSITY
COLLEGE STATION, TX USA 77843-4228

Abstract. In this paper we propose a test for the significance of categorical predictors in nonparametric regression models. The test is fully data-driven and employs cross-validated smoothing parameter selection while the null distribution of the test is obtained via bootstrapping. The proposed approach allows applied researchers to test hypotheses concerning categorical variables in a fully nonparametric and robust framework, thereby deflecting potential criticism that a particular finding is driven by an arbitrary parametric specification. Simulations reveal that the test performs well, having significantly better power than a conventional frequency-based nonparametric test. The test is applied to determine whether OECD and non-OECD countries follow the same growth rate model or not. Our test suggests that OECD and non-OECD countries follow different growth rate models, while the tests based on a popular parametric specification and the conventional frequency-based nonparametric estimation method fail to detect any significant difference.

Date: October 29, 2003.
The authors would like to thank but not implicate Essie Maasoumi for his helpful comments. A preliminary version of this paper was presented at the 2002 International Conference on Current Advances and Trends in Nonparametric Statistics held in Crete, and we would like to thank numerous conference participants for their valuable input. Hart's research was supported by NSF Grant DMS 99-71755. Li's research is supported by the Bush Program in the Economics of Public Policy, and the Private Enterprises Research Center, Texas A&M University. Racine would like to thank the Center for Policy Research at Syracuse University for their ongoing support.

1. Introduction

Though traditional nonparametric kernel methods presume that underlying data types are continuous in nature, it is common to encounter a mix of continuous and categorical data types in applied data analysis. Such encounters have spawned a growing literature on semiparametric and nonparametric kernel estimation in the presence of mixed data types, beginning with the seminal work of Aitchison and Aitken (1976) on through work by Hall (1981), Grund and Hall (1993), Scott (1992), Simonoff (1996), and Li and Racine (2003), to mention only a few.

The 'test of significance' is probably the most frequently used test in applied regression analysis, and is often used to confirm or refute theories. Sound parametric inference hinges on the correct functional specification of the underlying data generating process (DGP); however, the likelihood of misspecification in a parametric framework cannot be ignored, particularly in light of the fact that applied researchers tend to choose parametric models on the basis of parsimony and tractability. Significance testing in a nonparametric kernel framework would therefore have obvious appeal given that nonparametric techniques are consistent under much less restrictive assumptions than those required for a parametric approach. Fan and Li (1996), Racine (1997), and Chen and Fan (1999) have considered nonparametric tests of significance of continuous variables in nonparametric regression models. While it is possible to extend these tests to the case of testing the significance of a categorical variable using the conventional nonparametric frequency estimation method, such a test is likely to suffer finite-sample power loss because this conventional frequency approach splits the sample into a number of 'discrete cells' or subsamples, and only uses those observations within each cell to generate a nonparametric estimate. This efficiency loss is unfortunate because, under the null hypothesis, some discrete variables are irrelevant regressors and should therefore be removed from the regression model, i.e., the corresponding discrete cells should be 'smoothed out' or 'pooled' as opposed to

splitting the sample into different cells. The sample splitting method also suffers the unfortunate drawback that, when the number of discrete cells is large relative to the sample size, the conventional frequency approach may even become infeasible.

In this paper we smooth both the discrete and continuous variables, and we propose a test for the significance of categorical variables in nonparametric regression models. The test employs cross-validated smoothing parameter selection, while the null distribution of the test is obtained via bootstrapping methods (Efron (1983), Hall (1992), Beran (1988), Efron and Tibshirani (1993)). This approach results in a nonparametric test that is robust to functional specification issues, while the sampling distribution of the statistic under the null is also obtained in a nonparametric fashion, that is, there are neither unknown parameters nor functional forms that need to be set by the applied researcher. The paper proceeds as follows: Section 2 presents the proposed test statistic, Section 3 outlines a resampling approach for generating the test's null distribution, Section 4 examines the finite-sample performance of the statistic, Section 5 presents an application of the method to the question of whether or not 'convergence clubs' exist, an issue which arises in the economics of growth literature, while Section 6 concludes.

2. The Test Statistic

We consider a nonparametric regression model with mixed categorical and continuous regressors, and we are interested in testing whether some of the categorical regressors are 'irrelevant.' Let z denote the categorical variables that might be redundant, let x be the remaining explanatory variables in the regression model, and let y denote the dependent variable. Then the null hypothesis can be written as

$$(1) \qquad H_0: \qquad E(y|x,z) = E(y|x) \text{ almost everywhere (a.e.).}$$

The alternative hypothesis is the negation of H_0. H_1: $E(y|x,z) \neq E(y|x)$ on a set with positive measure. We allow x to contain both categorical (discrete) and continuous variables. Let x^c and x^d denote the continuous and discrete components of x, respectively. We assume that $x^c \in R^q$ and x^d is of dimension $k \times 1$. We will first focus on the case where z is a univariate categorical variable. We discuss the multivariate z variable case at the end of this section.

It is well known that bandwidth selection is of crucial importance for nonparametric estimation. The test statistic proposed in this paper depends on data-driven cross-validated smoothing parameter selection for both the discrete variable z and the mixed variables x. Given its importance, we briefly discuss the cross-validation method used herein.

Let $g(x) = E(y|x)$ and $m(x,z) = E(y|x,z)$. The null hypothesis is $m(x,z) = g(x)$ a.e. Suppose that the univariate z takes c different values: $\{0, 1, 2, \ldots, c-1\}$. If $c = 2$ then z is a 0-1 dummy variable, which is probably the most commonly encountered case in practice.

We assume that some of the discrete variables are ordinal (having a natural ordering), examples of which would include preference orderings (like, indifference, dislike), health conditions (excellent, good, poor) and so forth. Let $\tilde{x}_i^d$ denote a $k_1 \times 1$ vector (say, the first k_1 components of x_i^d, $0 \leq k_1 \leq k$) of discrete regressors that have a natural ordering, and let $\bar{x}_i^d$ denote the remaining $k_2 = k - k_1$ discrete regressors that are only nominal (no natural ordering). We use $x_{i,t}^d$ to denote the tth component of x_i^d $(t = 1, \ldots, k)$. It should be mentioned that Ahmad and Cerrito (1994) and Bierens (1983, 1987) also consider the case of estimating a regression function with mixed categorical and continuous variables, but they did not study the theoretical properties of the resulting estimator when using data-driven methods such as cross-validation to select smoothing parameters.

For an ordered categorical variable, we use the following kernel function:

$$(2) \qquad \tilde{l}(\tilde{x}_{i,t}^d, \tilde{x}_{j,t}^d, \lambda) = \begin{cases} 1, & \text{if } \tilde{x}_{i,t}^d = \tilde{x}_{j,t}^d, \\ \lambda^{|\tilde{x}_{i,t}^d - \tilde{x}_{j,t}^d|}, & \text{if } \tilde{x}_{i,t}^d \neq \tilde{x}_{j,t}^d, \end{cases}$$

where λ is a smoothing parameter. Note that (i) when $\lambda = 0$, $l(\tilde{x}_{i,t}^d, \tilde{x}_{j,t}^d, \lambda = 0)$ becomes an indicator function, and (ii) when $\lambda = 1$, $l(\tilde{x}_{i,t}^d, \tilde{x}_{j,t}^d, \lambda = 1) = 1$ is a uniform weight function. These two properties are of utmost importance when smoothing discrete variables. Property (i) is indispensable because otherwise the smoothing method may lead to inconsistent nonparametric estimation, and (ii) is indispensable as it results in a kernel estimator having the ability to smooth out (remove) an irrelevant discrete variable.

All of the existing (discrete variable) kernel functions satisfy (i), but many of them do not satisfy (ii). For example, when $\tilde{x}_t \in \{0, 1, \ldots, c_{t-1}\}$, Aitchison and Aitken (1976) suggested the weighting function: $l(\tilde{x}_{i,t}^d, \tilde{x}_{j,t}^d, \lambda) = \binom{c_t}{m}(1-\lambda)^{c_t - m}\lambda^m$ if $|\tilde{x}_{i,t}^d - \tilde{x}_{j,t}^d| = m$ $(0 \leq m \leq c_t)$. This kernel satisfies (i), but, it is easy to see that it cannot give a uniform weight function for any choice of λ when $c_t \geq 3$. Thus, it lacks the ability to smooth out an irrelevant discrete variable.

For an unordered categorical variable, we use a variation on Aitchison and Aitken's (1976) kernel function defined by

$$(3) \qquad \bar{l}(\bar{x}_{i,t}^d, \bar{x}_{j,t}^d) = \begin{cases} 1, & \text{if } \bar{x}_{i,t}^d = \bar{x}_{j,t}^d, \\ \lambda, & \text{otherwise.} \end{cases}$$

Again $\lambda = 0$ leads to an indicator function, and $\lambda = 1$ gives a uniform weight function.

Let $\mathbf{1}(A)$ denote an indicator function that assumes the value 1 if the event A occurs and 0 otherwise. Combining (2) and (3), we obtain the product kernel function given by

$$(4) \qquad L(x_i^d, x_j^d, \lambda) = \left[\prod_{t=1}^{k_1} \lambda^{|\tilde{x}_{i,t}^d - \tilde{x}_{j,t}^d|}\right] \left[\prod_{t=k_1+1}^{k} \lambda^{1-\mathbf{1}(\bar{x}_{i,t}^d = \bar{x}_{j,t}^d)}\right] = \lambda^{\tilde{d}_{x_i,x_j} + \bar{d}_{x_i,x_j}} = \lambda^{d_{x_i,x_j}},$$

where $\tilde{d}_{x_i,x_j} = \sum_{t=1}^{k_1} |\tilde{x}_{i,t}^d - \tilde{x}_{j,t}^d|$ is the distance between $\tilde{x}_i^d$ and $\tilde{x}_j^d$, $\bar{d}_{x_i,x_j} = k_2 - \sum_{t=k_1+1}^{k} \mathbf{1}(\bar{x}_{t,i}^d = \bar{x}_{j,t}^d)$ is the number of disagreement components between $\bar{x}_i^d$ and $\bar{x}_j^d$, and $d_{x_i,x_j} = \tilde{d}_{x_i,x_j} + \bar{d}_{x_i,x_j}$.

It is fairly straightforward to generalize the above result to the case of a k-dimensional vector of smoothing parameters λ. As noted earlier, for simplicity of presentation, only the scalar λ case is treated here. Of course in practice one needs to allow each different discrete variable (each component of z_i) to have a different smoothing parameter just as in the continuous variable case. For the simulations and application conducted herein we allow λ to differ across variables.

Since we have assumed that z is a univariate categorical variable, the kernel function for z is the same as (2). If z is an ordinal categorical variable, i.e.,

$$(5) \qquad l(z_i, z_j, \lambda_z) = \begin{cases} 1, & \text{if } z_i = z_j, \\ \lambda_z^{|z_i - z_j|}, & \text{if } z_i \neq z_j, \end{cases}$$

where λ_z is the smoothing parameter. If z is nominal, then

$$(6) \qquad l(z_i, z_j, \lambda_z) = \begin{cases} 1, & \text{if } z_i = z_j, \\ \lambda_z, & \text{otherwise.} \end{cases}$$

We use a different notation λ_z for the smoothing parameter for the z variable because under H_0, the statistical behavior of λ_z is quite different from λ, the smoothing parameter associated with x^d.

We use $W(\cdot)$ to denote the kernel function for a continuous variable and h to denote the smoothing parameter for a continuous variable. We will use the shorthand notations

$W_{h,ij} = h^{-q}W((x_i^c - x_j^c)/h)$, $L_{\lambda,ij} = L(x_i^d, x_j^d, \lambda)$, and $l_{\lambda_z,ij} = l(z_i, z_j, \lambda_z)$. Then a leave-one-out kernel estimator of $m(x_i, z_i) \equiv m(x_i^c, x_i^d, z_i)$ is given by

$$\hat{m}_{-i}(x_i, z_i) = \frac{\sum_{j \neq i} y_j W_{h,ij} L_{\lambda,ij} l_{\lambda_z,ij}}{\sum_{j \neq i} W_{h,ij} L_{\lambda,ij} l_{\lambda_z,ij}}. \tag{7}$$

We point out that Ahmad and Cerrito (1994) consider using more general discrete kernel functions, which include the kernel function used in (7) as a special case.

We choose (h, λ, λ_z) to minimize the following objective function:

$$CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{m}_{-i}(x_i, z_i)]^2, \tag{8}$$

where $\hat{m}_{-i}(x_i, z_i)$ is defined in (7).

We use $(\hat{h}, \hat{\lambda}, \hat{\lambda}_z)$ to denote the cross-validation choices of (h, λ, λ_z) that minimize (8). When H_1 is true (i.e., H_0 is false), Racine and Li (2003) have shown that $\hat{h} = O_p(n^{-1/(4+q)})$, $\hat{\lambda} = O_p(n^{-2/(4+q)})$, and $\hat{\lambda}_z = O_p(n^{-2/(4+q)})$. All the smoothing parameters converge to 0 under H_1. Intuitively this is easy to understand as the nonparametric estimation bias is of the order of $O(h^2 + \lambda + \lambda_z)$. Consistency of the nonparametric estimator requires that h, λ, and λ_z should all converge to 0 as $n \to \infty$. However, when H_0 is true, it can be shown that $\hat{h}$ and $\hat{\lambda}$ tend to zero in probability as $n \to \infty$, but $\hat{\lambda}_z$ has a high probability of being near its upper bound of 1, a fact confirmed by our simulations.[1] This is also easy to understand since, under H_0, the regression function is not related to z and therefore it is more efficient to estimate the regression function using $\lambda_z = 1$ rather than values of $\lambda_z < 1$. In such cases, the cross-validation method correctly selects large values of λ_z with high probability. Thus, our estimation method is much more efficient than the conventional sample splitting method, especially under the null hypothesis, because our method tends

[1]Hart and Wehrly (1992) observe a similar phenomenon with a cross-validation-based test for linearity with a univariate continuous variable. In their case h tends to take a large positive value when the null of linearity is true. For a sample size of $n = 100$, they observe that 60 percent of the time the smoothing parameter assumes values larger than $1,000$.

to smooth out the irrelevant discrete regressors while the conventional frequency method splits the sample into a number of subsets even when the discrete variable is irrelevant. We now discuss the test statistic that we use for testing the null hypothesis.

Note that the null hypothesis H_0 is equivalent to: $m(x, z = l) = m(x, z = 0)$ almost everywhere for $l = 1, \ldots, c-1$. Our test statistic is an estimator of

$$(9) \qquad I = \sum_{l=1}^{c-1} E\left\{[m(x, z = l) - m(x, z = 0)]^2\right\}.$$

Obviously $I \geq 0$ and $I = 0$ if and only if H_0 is true. Therefore, I serves as a proper measure for testing H_0. A feasible test statistic is given by

$$(10) \qquad \hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{c-1} [\hat{m}(x_i, z_i = l) - \hat{m}(x_i, z_i = 0)]^2,$$

where

$$(11) \qquad \hat{m}(x_i, z_i = l) = \frac{\sum_{j=1}^{n} y_j W_{\hat{h},ij} L_{\hat{\lambda},ij} l_{z_j, z=l, \hat{\lambda}_z}}{\sum_{j=1}^{2} W_{\hat{h},ij} L_{\hat{\lambda},ij} l_{z_j, z=l, \hat{\lambda}_z}}.$$

It is easy to show that $\hat{I}_n$ is a consistent estimator of I. Therefore, $\hat{I}_n \to 0$ in probability under H_0 and $\hat{I}_n \to I > 0$ in probability under H_1. In practice one should reject H_0 if $\hat{I}_n$ takes 'too large' a value.

It is straightforward to generalize the test statistic (10) to handle the case where z is a multivariate categorical variable. Suppose z is of dimension r. Let z_l and $z_{l,i}$ denote the lth components of z and z_i, respectively, and assume that z_l takes c_l different values in $\{0, 1, \ldots, c_l - 1\}$ $(l = 1, \ldots, r)$. For multivariate z the test statistic $\hat{I}_n$ becomes

$$(12) \qquad \hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{z} [\hat{m}(x_i, z) - \hat{m}(x_i, z_1 = 0, \ldots, z_r = 0)]^2,$$

where $\sum_z$ denotes summation over all possible values of $z \in \prod_{l=1}^{r} \{0, 1, \ldots, c_l - 1\}$. The definition of $\hat{m}(x_i, z)$ is similar to (11) except that the univariate kernel $\bar{l}(z_i, z_j, \hat{\lambda}_z)$

should be replaced by the product kernel $\prod_{s=1}^{r} l(z_{s,i}, z_{s,j}, \hat{\lambda}_{z,s})$, and the $\hat{\lambda}_{z,s}$'s are the cross-validated values of $\lambda_{z,s}$, the smoothing parameters associated with z_s $(s = 1, \ldots, r)$. We now turn our attention to using bootstrap methods to approximate the finite-sample null distribution of the test statistic.

3. A Bootstrap Procedure

A conventional approach for determining when $\hat{I}_n$ assumes 'too large' a value involves obtaining its asymptotic distribution under H_0 and then using this to approximate the finite-sample null distribution. However, in a nonparametric setting this approach is known to be unsatisfactory. To see why, note first that for smoothing-based tests in which the bandwidth tends to 0 as $n \to \infty$, the limiting distribution of the statistic is usually completely free of the bandwidth. However, in finite-sized samples, the distribution *does* depend on the bandwidth of the smoother. Indeed, a number of authors have noted that the distribution is quite sensitive to bandwidth choice. For example, Robinson (1991) states that, for a kernel smoother-based test, "substantial variability in the [test statistic] across bandwidths was recorded."

In contrast to the case just discussed, the data-driven bandwidth $\hat{\lambda}_z$ converges in distribution to a nondegenerate random variable under H_0. This means that $\hat{\lambda}_z$ has a *first order* effect on the limiting distribution of $\hat{I}_n$. Determining this effect precisely is a daunting theoretical problem. But even if it were straightforward to derive this limiting distribution, one would still be skeptical of its accuracy as a small-sample approximation. Therefore, we will forgo asymptotics altogether and instead use the bootstrap in order to approximate critical values for our test.

Resampling, or bootstrap, methods (Efron (1983)) have been successfully used for approximating the finite-sample null distributions of test statistics in a range of settings, both parametric and nonparametric. These methods have been shown to account remarkably

well for the effect of bandwidth on the null distribution of test statistics (Racine (1997)). Note that in our testing problem one should not simply resample from $\{y_i, x_i, z_i\}_{i=1}^n$ since doing so does not impose H_0. We therefore propose two bootstrap procedures, both of which approximate the null distribution of $\hat{I}_n$.

3.1. **Bootstrap Method I.**

(1) Randomly select z_i^* from $\{z_j\}_{j=1}^n$ with replacement, and call $\{y_i, x_i, z_i^*\}_{i=1}^n$ the bootstrap sample.

(2) Use the bootstrap sample to compute the bootstrap statistic $\hat{I}_n^*$, where $\hat{I}_n^*$ is the same as $\hat{I}_n$ except that z_i is replaced by z_i^* (using the same cross-validated smoothing parameters of $\hat{h}$, $\hat{\lambda}$ and $\hat{\lambda}_z$ obtained earlier).

(3) Repeat steps 1 and 2 a large number of times, say B times. Let $\{\hat{I}_{n,j}^*\}_{j=1}^B$ be the ordered (in an ascending order) statistic of the B bootstrap statistics, and let $\hat{I}_{n,(\alpha)}^*$ denote the αth percentile of $\{\hat{I}_{n,j}^*\}_{j=1}^B$. We reject H_0 if $\hat{I}_n > \hat{I}_{n,(\alpha)}^*$ at the level α.

The advantage of 'Bootstrap Method I' above is that it is computationally simple. This follows simply because one does not re-compute the cross-validated smoothing parameters for each bootstrap sample. The second method outlined below, 'Bootstrap Method II,' is computationally more intensive than Bootstrap Method I outlined above.

3.2. **Bootstrap Method II.**

(1) Randomly select z_i^* from $\{z_j\}_{j=1}^n$ with replacement, and call $\{y_i, x_i, z_i^*\}_{i=1}^n$ the bootstrap sample.

(2) Use the bootstrap sample to find the least squares cross-validation smoothing parameter $\hat{\lambda}_z^*$, i.e., choose $\hat{\lambda}_z^*$ to minimize

$$CV(\lambda_z) = \sum_{i=1}^{n} [y_i - \hat{g}_{-i}(x_i, z_i^*)]^2, \tag{13}$$

where $\hat{g}_{-i}(x_i, z_i^*) = \sum_{j \neq i} y_j K_{\hat{h}_x, \hat{\lambda}_x} L(z_i^*, z_j^*, \lambda_z) / \sum_{j \neq i} K_{\hat{h}_x, \hat{\lambda}_x} L(z_i^*, z_j^*, \lambda_z)$. Compute the bootstrap statistic $\tilde{I}_n^*$ in the same way as $\hat{I}_n$ except that z_i and $\hat{\lambda}_z$ are replaced by z_i^* and $\hat{\lambda}^*$, respectively.

Note that, in the above cross-validation procedure, only λ_z varies since $\hat{h}_x$ and $\hat{\lambda}_x$ are obtained at the initial estimation stage.

(3) Repeat steps (i) and (ii) a large number of times, say B times. Let $\{\tilde{I}_{n,j}^*\}_{j=1}^B$ be the ordered (in an ascending order) statistic of the B bootstrap statistics, and let $\tilde{I}_{n,(\alpha)}^*$ denote the αth percentile of $\{\tilde{I}_{n,j}^*\}_{j=1}^B$. We reject H_0 if $\hat{I}_n > \tilde{I}_{n,(\alpha)}^*$ at the level α.

Bootstrap Method I is computationally more attractive than Bootstrap Method II because the latter uses the cross-validation method to select λ_z^* at each bootstrap resample. Results of Hall and Kang (2001) seem to suggest that there would be little (if any) benefit to using the more computationally burdensome Bootstrap Method II. Their results show that Edgeworth expansions of the distributions of kernel density estimators $\hat{f}(x|h_0)$ and $\hat{f}(x|\hat{h})$ have the same first terms, where h_0 and $\hat{h}$ are optimal and (consistent) data-driven bandwidths, respectively. An implication of this result is that computing $\hat{h}^*$ on each bootstrap sample has no impact on first-order accuracy of the bootstrap. However, in contrast to the setting of Hall and Kang (2001), $\hat{\lambda}_z$ has a nondegenerate asymptotic distribution, and we thus anticipate a marked improvement (at least asymptotically) from using Bootstrap Method II. Simulations using both methods will be conducted in the next section.

Finally, we note that the above two bootstrap procedures are not (asymptotically) pivotal. However, the simulation results presented below show that both testing procedures work well even for small to moderate sample sizes. We have also computed a standardized

version of our test statistic as follows:

$$\hat{t}_n = \frac{\hat{I}_n}{s(\hat{I}_n)}, \tag{14}$$

where $s(\hat{I}_n)$ is the estimated standard error of $\hat{I}_n$, which is itself obtained via nested bootstrap resampling (e.g., Efron and Tibshirani (1993, page 162)). We discuss the finite sample performance of $\hat{t}_n$ in comparison to $\hat{I}_n$ in the next section.

4. Monte Carlo Simulations

For all simulations that follow, we consider sample sizes of $n = 50$ and 100, while 1,000 Monte Carlo replications are conducted throughout. When bootstrapping the null distribution of the test statistic, we employ 299 bootstrap replications. As noted, cross-validation is used to obtain all smoothing parameters, while 2 restarts of the search algorithm for different admissible random values of (h, λ, λ_z) are used throughout, with those yielding the lowest value of the cross-validation function being retained. A second-order Epanechnikov kernel was used for the continuous variable. Code was written in ANSI C and simulations were run on a 24-node Athlon-based Beowulf cluster (see Racine (2002) for further details).

4.1. **Size and Power of the Proposed Tests.** In this section we report on a Monte Carlo experiment designed to examine the finite-sample size and power of the proposed test. The data generating process (DGP) we consider is a nonlinear function having interaction between a discrete and continuous variable and is given by

$$y_i = \beta_0 + \beta_1 z_{i1}(1 + x_i^2) + \beta_2 z_{i2} + \beta_3 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{15}$$

where Z_1 and Z_2 are both discrete binomial 0/1 random variables having $Pr[Z_j = 1] = 0.5$, $j = 1, 2$, $X \sim N(0, \sigma_x^2)$ with $\sigma_x = 1.0$, $\epsilon \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.1$, and $(\beta_0, \beta_1, \beta_2, \beta_3) = (1, \beta_1, 1, 1)$.

We consider testing the significance of Z_1. Under the null ($\beta_1 = 0$) that Z_1 is an irrelevant regressor, the DGP is $y_i = \beta_0 + \beta_2 z_{i2} + \beta_3 x_i + \epsilon_i$. We assess the finite-sample size and power of the test by varying β_1, so that when $\beta_1 = 0$ we can examine the test's size while when $\beta_1 \neq 0$ we can assess the test's power.

We consider the performance of the proposed test using both bootstrap methods outlined above. We vary β_1 in increments of 0.05 $(0.0, 0.05, 0.10, \dots)$ and compute the empirical rejection frequency at nominal levels $\alpha = (0.01, 0.05, 0.10)$. We then construct smooth power curves. Power curves are plotted in Figures 1 and 2, while rejection frequencies for $n = 50$ can be found in Tables 1 and 2.

We note from Figures 1 and 2 that the sizes of the test based on Bootstrap Method I are a bit high for $n = 50$, while those of Bootstrap Method II are closer to their nominal values. In fact, the empirical rejection rate for Bootstrap Method I is significantly higher than nominal size for each α and sample size (using a one-sided binomial test with level 0.05).[2] It is also worth noting that the empirical levels for Bootstrap Method II improve when n increases from 50 to 100, while those for Bootstrap Method I barely change.

The estimated powers for Bootstrap Method I and Bootstrap Method II are somewhat close to each other. The slightly better power of Bootstrap Method I is likely due to its being somewhat oversized in comparison with Bootstrap Method II. As expected, the power increases as either β_1 or n increases ($\beta_1 \neq 0$).

Based on the limited simulation results reported above, it appears that Bootstrap Method II has somewhat better level properties than Bootstrap Method I, while the power properties of the two are somewhat similar. When computational burdens are not an issue, we thus suggest use of Bootstrap Method II for moderate sample sizes. On the other hand, the excess size of Bootstrap Method I is not substantial for sample sizes of 100

[2]Letting $\hat{\alpha}$ denote the empirical rejection frequency associated with nominal level α, we tested the null H_0: $\hat{\alpha} = \alpha$ for Bootstrap Method I for $n = 50$, and obtained P-values of 0.04, 0.00, and 0.00 for $\alpha = 0.01$, 0.05, and 0.10 respectively, while for Bootstrap Method II, we obtained P-values of 0.14, 0.29, and 0.18 for $\alpha = 0.01$, 0.05, and 0.10 respectively.

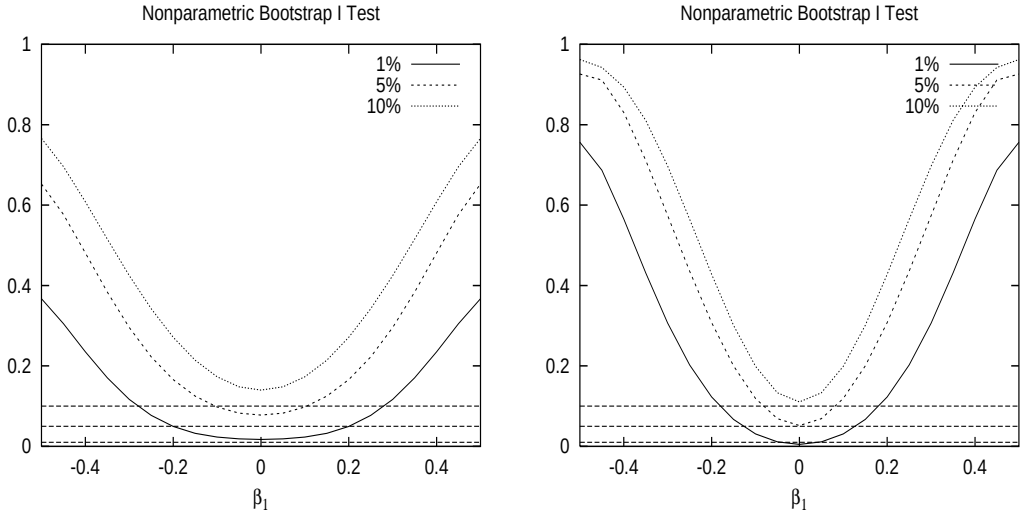

FIGURE 1. Empirical size and power of Bootstrap Method I, $n = 50$ (left) and $n = 100$ (right).

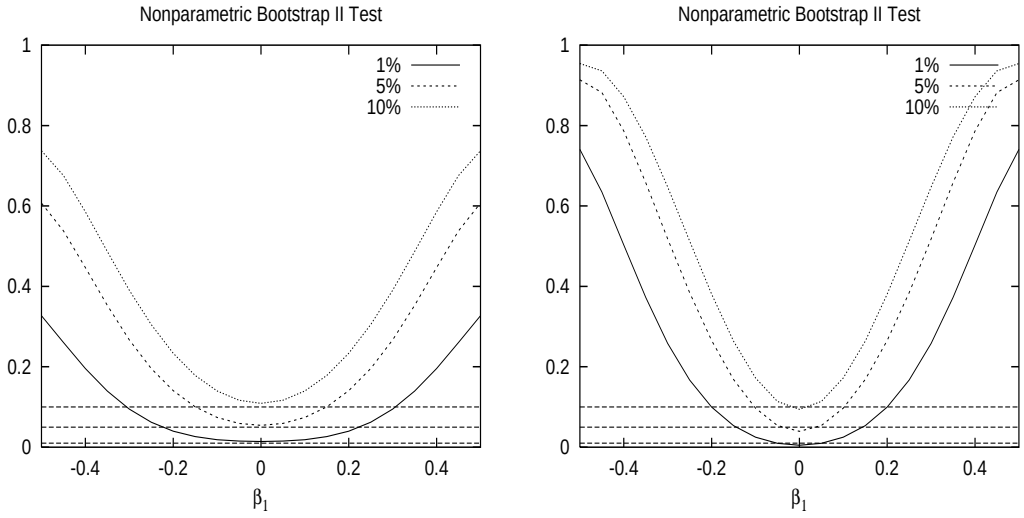

FIGURE 2. Empirical size and power of Bootstrap Method II, $n = 50$ (left) and $n = 100$ (right).

or more, and hence it is a reasonable method when it is important to reduce computation time.

TABLE 1. Empirical size ($\beta_1 = 0$) and power ($\beta_1 \neq 0$) of Bootstrap Method I, $n = 50$.

β_1	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
0.000	0.017	0.078	0.140
0.100	0.023	0.098	0.173
0.200	0.049	0.166	0.271
0.300	0.117	0.296	0.425
0.400	0.235	0.481	0.608
0.500	0.367	0.653	0.766

TABLE 2. Empirical size ($\beta_1 = 0$) and power ($\beta_1 \neq 0$) of Bootstrap Method II, $n = 50$.

β_1	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
0.000	0.014	0.054	0.109
0.100	0.019	0.074	0.140
0.200	0.040	0.141	0.234
0.300	0.095	0.267	0.391
0.400	0.196	0.447	0.586
0.500	0.327	0.607	0.737

4.2. **Size and Power of Our Tests Relative to the Conventional Nonparametric 'Frequency' Estimator.** We now consider the same DGP given in (15) above, but in this section our goal is to assess the finite-sample performance of the proposed 'smoothing' test relative to the conventional 'non-smoothing' ('frequency') test. This is accomplished simply by setting $\lambda = 0$ for all categorical variables in the model. Empirical size and power for both bootstrap methods are tabulated in Tables 1, 2, and 3 for the case of $n = 50$.[3]

It can be seen from Tables 1, 2, and 3 that the proposed method is substantially more powerful than the non-smoothed (frequency) approach, the loss in power increasing as β_1 increases for the range considered herein. It would appear therefore that optimal smoothing leads to finite-sample power gains for the proposed test relative to the non-smoothed version of the test.

[3]Qualitatively similar results were obtained for $n = 100$, thus we do not those tables for the sake of brevity.

TABLE 3. Empirical size ($\beta_1 = 0$) and power ($\beta_1 \neq 0$) of the non-smoothed test with $\lambda = 0$, $n = 50$.

β_1	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
0.000	0.013	0.054	0.110
0.100	0.018	0.071	0.136
0.200	0.035	0.127	0.216
0.300	0.072	0.235	0.349
0.400	0.139	0.390	0.522
0.500	0.243	0.541	0.690

4.3. **The Standardized Test.** In addition to the experiments reported above, we also implemented a standardized version of our test, the $\hat{t}_n$ test, as described at the end of Section 3. Compared with the non-standardized test $\hat{I}_n$ using Bootstrap Method I, the use of $\hat{t}_n$ in this case yields small improvements in nominal size as expected and also appears to lead to a small reduction in power.[4] Recall that the $\hat{I}_n$ test based on Bootstrap Method I is slightly oversized. Thus, this power reduction may reflect the difference in estimated sizes. Indeed simulations (not reported here) show that the size-adjusted powers of the $\hat{I}_n$ and the $\hat{t}_n$ tests are virtually identical. The use of $\hat{t}_n$ based on nested bootstrap procedure increases the computational burden of the proposed approach by an order of magnitude; hence, we conclude that standardizing the test does not appear to be necessary to achieve reasonable size and power in this setting.

5. Application - 'Growth Convergence Clubs'

Quah (1997) and others have examined the issue of whether there exist 'convergence clubs,' that is, whether growth rates differ for members of clubs such as the Organization for Economic Cooperation and Development (OECD) among others. We do not attempt to review this vast literature here; rather, we refer the interested reader to Mankiw et al. (1992), Liu and Stengos (1999), Durlauf and Quah (1999) and the references therein.

[4]These simulation results are not reported here to save space. The results are available from the authors upon request.

We apply the proposed test to determine whether OECD countries and non-OECD countries follow the same growth model. This is done by testing whether OECD membership (a binary categorical variable) is a relevant regressor in a nonparametric framework. The null hypothesis is that the OECD membership is an irrelevant regressor; thus, under the null, OECD and non-OECD countries' growth rates are all determined by the same growth model. The alternative hypothesis is the negation of the null hypothesis. That is, OECD and non-OECD countries have different growth rate (regression) models.

When using parametric methods, if the regression functional form is misspecified one may obtain misleading conclusions. By using methods that are robust to functional specification issues we hope to avoid criticism that findings are driven by a particular functional form presumed.

Following Liu and Stengos (1999), we employ panel data for 88 countries over seven (five-year average) periods (1960-1964, 1965-1969, 1970-1974, 1975-1979, 1980-1984, 1985-1989 and 1990-1994) yielding a total of $88 \times 7 = 616$ observations in the panel. We then construct our test based on the following model:

(16)

$$\text{growth}_{it} = m(\text{OECD}_{it}, \text{DT}_t, \ln(\text{inv}_{it}), \ln(\text{popgro}_{it}), \ln(\text{initgdp}_{it}), \ln(\text{humancap}_{it})) + \epsilon_{it},$$

where growth_{it} refers to the growth rate of income per capita during each period, DT_t the seven period dummies, inv_{it} the ratio of investment to Gross Domestic Product (GDP), popgro_{it} growth of the labor force, initgdp_{it} per capita income at the beginning of each period, and humancap_{it} human capital. Initial income estimates are from the Summers-Heston (1988) data base, as are the estimates of the average investment/GDP ratio for five-year periods. The average growth rate of per capita GDP and the average annual population growth rate for each period are from the World Bank. Finally, human capital

(average years of schooling in the population above 15 years of age) is obtained from Barro and Lee (2000).

Before we report results for our smoothing-based nonparametric test, we first consider some popular parametric methods for approaching this problem. A common parametric approach is to employ a linear regression model, with the OECD dummy variable being one possible regressor, and then to test whether the coefficient on this dummy variable is significant. We consider a parametric specification suggested by Liu and Stengos (1999), which contains dummy variables for OECD status and is nonlinear in the initial GDP and human capital variables.[5]

$$\begin{aligned} \text{growth}_{it} = {} & \beta_0 \text{OECD}_{it} + \sum_{s=1}^{7} \beta_s \text{DT}_s + \beta_8 \ln(\text{inv}_{it}) + \beta_9 \ln(\text{popgro}_{it}) \\ & + \sum_{s=1}^{4} \alpha_s [\ln(\text{initgdp}_{it})]^s + \sum_{s=1}^{3} \gamma_s [\ln(\text{humancap})_{it})]^s + \epsilon_{it}. \end{aligned} \tag{17}$$

Estimation results for model (17) are given in Table 4, while the t-statistic for the OECD dummy is -0.973 having a P-value of 0.33.[6] Thus, the parametric test fails to reject the null.

Next, we follow the conventional frequency approach and implement the nonparametric test, i.e., our estimation is based on model (16) with sample splitting on the OECD and the DT dummies. Using 999 bootstrap resamples we obtain a P-value of 0.113, and once again we fail to reject the null at the conventional 1 percent, 5 percent, and 10 percent levels.

We now report the results for our proposed smoothing-based nonparametric test. For each bootstrap test we employed 999 bootstrap resamples, while for cross-validation we

[5]We are grateful to Thanasis Stengos for providing data and for suggesting this parametric specification based upon his work in this area.

[6]R code and data needed for the replication of these parametric results are available from the authors upon request.

TABLE 4. Parametric Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
OECD	−0.0043	0.0044	−0.97	0.3311
d1965	6.5101	3.8180	1.71	0.0887
d1970	6.5102	3.8182	1.71	0.0887
d1975	6.5129	3.8183	1.71	0.0886
d1980	6.5028	3.8184	1.70	0.0891
d1985	6.4863	3.8183	1.70	0.0899
d1990	6.4965	3.8182	1.70	0.0894
d1995	6.4913	3.8180	1.70	0.0896
initgdp	−3.3940	2.0025	−1.69	0.0906
I(initgdp^2)	0.6572	0.3908	1.68	0.0931
I(initgdp^3)	−0.0558	0.0336	−1.66	0.0975
I(initgdp^4)	0.0018	0.0011	1.63	0.1043
popgro	−0.0172	0.0105	−1.63	0.1035
inv	0.0185	0.0023	7.93	0.0000
humancap	0.0007	0.0032	0.21	0.8366
I(humancap^2)	0.0011	0.0021	0.51	0.6084
I(humancap^3)	0.0005	0.0011	0.45	0.6512
Residual standard error: 0.026 on 599 degrees of freedom				
Multiple R-Squared: 0.5077, Adjusted R-squared: 0.4937				
F-statistic: 36.34 on 17 and 599 DF, p-value: $<$ 2.2e-16				

employed five restarts of the numerical search algorithm and retained those smoothing parameters that yielded the lowest value of the cross-validation function. The P-value generated from inverting the empirical CDF at $\hat{I}_n$ is 0.006 for Bootstrap Method I and 0.003 for Bootstrap Method II, which constitutes strong evidence against the validity of the null.

The inconsistency of the parametric test and our proposed nonparametric test also suggests that the parametric model is misspecified. We therefore applied a consistent nonparametric test for correct specification of the parametric model (see Hsiao, Li, and Racine (2003)). The P-value from this test was 0.0008 and we therefore reject the null of correct parametric specification.

The reason why the conventional frequency based nonparametric test also fails to reject the null is that it splits the sample into $2 \times 7 = 14$ parts (the number of discrete cells

from the discrete variables OECD and DT) when estimating the nonparametric regression functions; thus, the much smaller (sub) sample sizes lead to substantial finite sample power loss for a test based on the conventional frequency approach.

We conclude that there is robust nonparametric evidence in favor of the existence of 'convergence clubs,' a feature that may remain undetected when using both common parametric specifications and conventional nonparametric approaches. That is, growth rates for OECD countries appear to be different from those for non-OECD countries.

6. Conclusion

In this paper we propose a test for the significance of categorical variables for nonparametric regression. The test is fully data-driven and uses resampling procedures for obtaining the null distribution of the test statistic. Monte Carlo simulations suggest that the test is well-behaved in finite samples, having correct size and power that increases with the degree of departure from the null and with the sample size. Two resampling methods for generating the test statistic's null distribution are proposed, and both perform admirably. The test is more powerful (in finite-sample applications) than a conventional non-smoothing version of the test, indicating that optimal smoothing of categorical variables is desirable not only for estimation but also for inference. An application demonstrates how one can test economic hypotheses concerning categorical variables in a fully nonparametric and robust framework, thereby parrying the thrusts of critics who might argue that the outcome was driven by the choice of a parametric specification.

References

Ahmad, I. A. and P. B. Cerrito (1994), "Nonparametric estimation of joint discrete-continuous probability densities with applications," *Journal of Statistical Planning and Inference*, 41, 349-364.

Aitchison, J. and C. G. G. Aitken (1976), "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413-420.

Barro, R. and J. W. Lee (2000), "International Data on Educational Attainment: Updates and Implications," Working paper No. 42, Center for International Development, Harvard University.

Beran, R. (1988), "Prepivoting test statistics: A bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 83, 687-697.

Chen, X. and Y. Fan (1999), "Consistent Hypothesis Tests in Nonparametric and Semiparametric Models for Econometric Time Series," *Journal of Econometrics*, 91, 373-401.

Durlauf, S. N., and D. T. Quah (1999), "The New Empirics of Economic Growth," Chapter 4, of J. B. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics I*, Elsevier Sciences, 235-308.

Efron, B. (1983), The Jackknife, the Bootstrap, and Other Resampling Plans, Philadelphia, Society for Industrial and Applied Mathematics.

Efron, B. and R. J. Tibshirani (1993), An Introduction to the Bootstrap, New York, London, Chapman and Hall.

Fan, Y. and Q. Li (1996), "Consistent model specification tests: omitted variables and semiparametric functional forms," *Econometrica*, 64, 865-890.

Hall, P. (1992), The Bootstrap and Edgeworth Expansion, New York, Springer Series in Statistics, Springer-Verlag.

Hall, P. and K. H. Kang (2001), "Bootstrapping nonparametric density estimators with empirically chosen bandwidths," *Annals of Statistics*, 29, 1443-1468.

Hart, J. and T. E. Wehrly (1992), "Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models," *Journal of American Statistical Association*, 87, 1018-1024.

Hsiao, C., Q. Li and J. S. Racine (2003), "A Consistent Model Specification Test with Mixed Categorical and Continuous Data," revised and resubmitted to the *International Economic Review*.

Li, Q. and J. S. Racine (2003), "Nonparametric estimation of distributions with categorical and continuous data," *Journal of Multivariate Analysis*, 86, 266-292.

Liu, Z. and T. Stengos (1999), "Non-linearities in cross country growth regressions: a semiparametric approach," *Journal of Applied Econometrics*, 14, 527-538.

Mankiw, N., Romer, D. and D. Weil (1992), "A contribution to the empirics of economic growth," *Quarterly Journal of Economics*, 108, 407-437.

Quah, D. T. (1997), "Empirics for growth and distribution: stratification, polarization and convergence clubs," *Journal of Economic Growth*, 2, 27-59.

Racine, J. S. (1997), "Consistent significance testing for nonparametric regression," *Journal of Business and Economic Statistics*, 15 (3), 369-379.

Racine, J. S. (2002), "Parallel distributed kernel estimation," *Computational Statistics and Data Analysis*, 40 (2), 293–302.

Racine, J. S. and Q. Li (2003), "Nonparametric estimation of regression functions with both categorical and continuous data," forthcoming in *Journal of Econometrics*.

Robinson, P. M. (1991), "Consistent nonparametric entropy-based testing," *Review of Economic Studies*, 58, 437-453.

Summers, R. and A. Heston (1988), "A new set of international comparisons of real product and prices: Estimates for 130 countries," *Review of Income and Wealth*, 34, 1-26.