Statistica Sinica **18**(2008), 601-615

TESTS FOR INDEPENDENCE IN NONPARAMETRIC REGRESSION

John H.J. Einmahl and Ingrid Van Keilegom

Tilburg University & *Université catholique de Louvain*

Abstract: Consider the nonparametric regression model $Y = m(X) + \varepsilon$, where the function m is smooth, but unknown. We construct tests for the independence of ε and X, based on n independent copies of (X, Y). The testing procedures are based on differences of neighboring Y's. We establish asymptotic results for the proposed tests statistics, investigate their finite sample properties through a simulation study and present an econometric application to household data. The proofs are based on delicate empirical process theory.

Key words and phrases: Empirical process, model diagnostics, nonparametric regression, test for independence, weak convergence.

1. Introduction

Let (X, Y) be a bivariate random vector where Y is the variable of interest and X is a covariate. We assume that X and Y are related via the nonparametric regression model

$$Y = m(X) + \varepsilon, \tag{1.1}$$

where m is the unknown regression curve and ε is the error. In order to avoid identification problems, we define m as follows. Let T be a given location functional, i.e., for any random variable Z and any $a > 0$ and b, we have $T(F_{aZ+b}) = aT(F_Z) + b$, where F_{aZ+b} is the distribution function of $aZ + b$. Now we define $m(x) = T(F(\cdot|\,x))$, with $F(\cdot|\,x)$ the conditional distribution function of Y, given $X = x$. As a consequence, $T(F_\varepsilon(\cdot|\,x)) = 0$, with $F_\varepsilon(\cdot|\,x)$ the conditional distribution function of ε, given $X = x$. In particular we can choose T to be the median (or a quantile), the mode, or the (trimmed) mean. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$, n independent replications of (X, Y), be our data.

In this paper we consider the problem of constructing omnibus tests for the submodel where

$$\varepsilon \text{ is stochastically independent of } X \tag{1.2}$$

or, in other words, where the conditional distribution of $Y - m(X)$, given $X = x$, does not depend on x. We propose procedures for testing the independence between ε and X that will detect any deviation from the null hypothesis. Although the nonparametric regression model (1.1) is a standard one, testing of (1.2) against the general alternative of dependence seems not to have been addressed in the literature. Model (1.1)$-$(1.2) is studied extensively in the literature, see, e.g., Akritas and Van Keilegom (2001), Neumeyer, Dette and Nagel (2006), and Van Keilegom, González Manteiga and Sánchez Sellero (2007) and the references therein.

In a number of papers (see, e.g., Lee (1992), Dette and Munk (1998), Liero (2003) and Cao and Gijbels (2005)) tests for homoscedasticity are developed. Instead of looking at the conditional variance only, in this paper we consider the full conditional distribution of ε given X. The motivation for considering this entire conditional distribution is as follows. Often, much better statistical inference can be made under (1.2) than when only homoscedasticity is assumed. To begin with, when estimating the conditional distribution of the error ε, given $X = x$, all the data can be used when (1.2) holds, see Akritas and Van Keilegom (2001), whereas only data with values of X around x can be used under homoscedasticity only. As a consequence, the same reasoning applies when estimating transformations of the conditional distribution of the error, such as the quantile function or the Lorenz curve, or functionals of this distribution, such as centered moments (skewness and kurtosis) or the extreme value index. When considering functionals (or transformations) of the conditional distribution of the response Y (instead of ε), given $X = x$, that can be written as functionals of the conditional error distribution (like the skewness), the above obviously remains applicable. When this is not the case, take e.g., a large quantile of Y, given $X = x$, we estimate it by the sum of this estimated quantile of the conditional error distribution and an estimator of $m(x)$. Now using (1.2) is in general again advantageous in comparison with using only homoscedasticity, since the quantile of the conditional error distribution can be better estimated. When the response Y is subject to random right censoring $-$ which is beyond the scope of this paper $-$ the use of (1.2) has even more advantages than in the uncensored case, considered here; see Van Keilegom and Akritas (1999). On the other hand, we like to emphasize that our tests detect heteroscedasticity very well.

Apart from being a goodness-of-fit test for the nonparametric model, the tests proposed in this paper can also serve for other purposes. Suppose e.g., that one wishes to know whether a certain random vector (X, Y) satisfies a parametric model $Y = m_{\boldsymbol{\beta}}(X) + \varepsilon$ (where ε is independent of X and $m_{\boldsymbol{\beta}}$ is a parametric regression curve, of which the form is still to be determined). In such a situation it might be useful to use the nonparametric tests proposed above. If the tests indicate that the independence between ε and X holds, one can then start searching for the particular form of the parametric regression curve.

Since the errors $\varepsilon_1, \ldots, \varepsilon_n$ are not observed, we cannot use them directly. We consider appropriate differences of Y's corresponding to neighboring X-values. Since m is smooth, m almost cancels out in these differences. The main difficulty is however that these differences are dependent, and hence the classical tests for independence available in the literature cannot be applied, since most tests assume that the pairs of observations are i.i.d. In this paper we focus on three tests, namely the Kolmogorov-Smirnov, the Cramér-von Mises and the Anderson-Darling test (see, e.g., Shorack and Wellner (1986)). We adapt these tests to the present setup and derive their asymptotic distributions. Difference-based procedures are widely used in nonparametric regression, especially for the estimation of the error variance (see e.g., Dette, Munk and Wagner (1998), Liero (2003), and Müller, Schick and Wefelmeyer (2003)).

Although the results in this paper will be presented for random design, they can easily be adapted to fixed design. Note that in that case, interest lies in whether or not the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are identically distributed.

This paper is organized as follows. In Section 2 we propose the test statistics, state and prove the main results. In Section 3 we investigate the finite sample performance of the tests in a simulation study, and in Section 4 we present an econometric application.

2. Main Results

Consider the model described in (1.1). We write F_X for the distribution function (df) of X and F_ε for the (unconditional) df of ε. Let $(X_1, \varepsilon_1), \ldots, (X_n, \varepsilon_n)$ be i.i.d. copies of (X, ε). We want to test

$$H_0 : \varepsilon \text{ is independent of } X$$

against the alternative of dependence, based on (X_i, Y_i), $i = 1, \ldots, n$, with $Y_i = m(X_i) + \varepsilon_i$. In this section we present certain test statistics and derive their asymptotic distribution under H_0. It should be noted that for the approach detailed below the actual choice of the location functional T (see Section 1) has, under H_0, no influence on the distribution of the test statistics below. If H_0 does not hold, the influence of the choice of T on the distribution of the test statistics is typically very minor. The method is rather robust in this sense.

Let $X_{1:n} \le \cdots \le X_{n:n}$ be the order statistics of the $X_i, i = 1, \ldots, n$, and denote with $Y_{[1:n]}, \ldots, Y_{[n:n]}$, the concomitants (or induced order statistics), the Y-values corresponding to the ordered X-values. Write

$$F_n(x, y) = \frac{1}{n} \sum_{j=1}^{n} I\Big(X_{j:n} \le x,\ Y_{[j-1:n]} - 2Y_{[j:n]} + Y_{[j+1:n]} \le y\Big). \tag{2.1}$$

(For notational convenience we relabel the original n by $n+2$ and take $j = i+1$ in order to have all the quantities properly defined; the final sample size is now n.) So F_n is the bivariate empirical df of the pairs $(X_j, Y_{[j-1:n]} - 2Y_{[j:n]} + Y_{[j+1:n]})$, $j = 1, \ldots, n$, i.e., we take an appropriate difference of three Y-values, corresponding to neighboring X-values. Set $\hat{F}_X(x) = F_n(x, \infty)$ and similarly $\hat{G}(y) = F_n(\infty, y)$. For our testing problem we consider the following test statistics:

$$T_{n,KS} = \sqrt{n} \sup_{x,y \in I\!R} \left| F_n(x,y) - \hat{F}_X(x)\hat{G}(y) \right|, \tag{2.2}$$

$$T_{n,CM} = n \iint (F_n(x,y) - \hat{F}_X(x)\hat{G}(y))^2 d\hat{F}_X(x) d\hat{G}(y), \tag{2.3}$$

$$T_{n,AD} = n \iint \frac{(F_n(x,y) - \hat{F}_X(x)\hat{G}(y))^2}{\hat{F}_X(x)\hat{G}(y)(1 - \hat{F}_{X-}(x))(1 - \hat{G}_-(y))} d\hat{F}_X(x) d\hat{G}(y). \tag{2.4}$$

(For a distribution function F, we denote by F_- its left-continuous version.) For bivariate i.i.d. random vectors, the first two statistics and the underlying process were introduced in Blum, Kiefer and Rosenblatt (1961); a statistic asymptotically equivalent to $T_{n,CM}$ dates back to Hoeffding (1948).

Remark 2.1. The choice of F_n in (2.1) needs explanation. Assume the third moment of the conditional error distribution is finite and, for convenience, let m be the conditional mean. Since we want m to vanish by using differences of Y's, taking the naive difference $Y_{[j-1:n]} - Y_{[j:n]}$ seems appropriate. Note however that we want our tests to improve on nonparametric tests for homoscedasticity. We want to detect conditional error distributions with equal variances, but with varying higher moments, in particular the third moment. The naive difference $Y_{[j-1:n]} - Y_{[j:n]}$ leads typically to the difference of two almost i.i.d. ε's, obviously has a third moment close to zero, and hence is useless for detecting a varying third moment. So next we take a linear combination of three Y-values: $aY_{[j-1:n]} + bY_{[j:n]} + cY_{[j+1:n]}$ $(a+b+c=0)$, where we choose the coefficients a, b, c such that the absolute value of the third moment of the corresponding linear combination of i.i.d. ε's is maximal, for fixed variance. This leads essentially to $a = c = 1$, $b = -2$, the coefficients used in (2.1). In this way we can detect a varying third moment easily. But this choice of coefficients has additional desirable properties, which the above naive difference lacks. If the class of distributions is such that all the moments exist and determine the distribution it can be readily shown, by an induction argument based on moments, that the distribution of $\varepsilon_l - 2\varepsilon_c + \varepsilon_r$ $(\varepsilon_l, \varepsilon_c, \varepsilon_r$ i.i.d) determines the distribution of ε_c. Therefore, when the conditional error distributions are in such a class of distributions, we can show consistency of our empirical process based tests (where m need not necessarily be the conditional mean). It is not clear if the df of $\varepsilon_l - 2\varepsilon_c + \varepsilon_r$ determines the df of ε_c in general, but we will see below that the tests perform well for various other alternatives.

All three test statistics are based on the process

$$\sqrt{n}\Big(F_n(x,y) - \hat{F}_X(x)\hat{G}(y)\Big), \quad x, y \in I\!R,$$

which we study first. In the remainder of this section we assume H_0 holds true. Let V_0 be a centered, bivariate Gaussian process with covariance structure

$$\begin{aligned}&E(V_0(x_1,y_1)V_0(x_2,y_2))\\&\quad= (F_X(x_1 \wedge x_2) - F_X(x_1)F_X(x_2))(G(y_1 \wedge y_2) + 2H_1(y_1,y_2) + 2H_2(y_1,y_2)\\&\qquad -5G(y_1)G(y_2)), \quad x_1, x_2, y_1, y_2 \in I\!R,\end{aligned}$$

where

$$\begin{aligned}&G(y) = P(\varepsilon_1 - 2\varepsilon_2 + \varepsilon_3 \le y),\\&H_1(y_1,y_2) = P(\varepsilon_1 - 2\varepsilon_2 + \varepsilon_3 \le y_1, \varepsilon_2 - 2\varepsilon_3 + \varepsilon_4 \le y_2)\ (= H_1(y_2,y_1)),\\&H_2(y_1,y_2) = P(\varepsilon_1 - 2\varepsilon_2 + \varepsilon_3 \le y_1, \varepsilon_3 - 2\varepsilon_4 + \varepsilon_5 \le y_2)\ (= H_2(y_2,y_1)).\end{aligned}$$

We have

$$G(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(1 - F_\varepsilon\Big(\frac{u+v-y}{2}\Big)\Big)dF_\varepsilon(u)dF_\varepsilon(v) \tag{2.5}$$

and, with g the density corresponding to G and f_ε the density of ε,

$$g(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2}f_\varepsilon\Big(\frac{u+v-y}{2}\Big)f_\varepsilon(u)f_\varepsilon(v)dudv. \tag{2.6}$$

Observe that V_0 is tied-down on all four sides, i.e., $V_0(x,y) = 0$ a.s. if $x = -\infty$ or $x = \infty$ or $y = -\infty$ or $y = \infty$. We now show that V_0 is the weak limit of $\sqrt{n}(F_n - \hat{F}_X\hat{G})$. Denote by D_X the support of X and by f_X its density. We assume that

$$D_X \text{ is a bounded interval and } \inf_{x\in D_X} f_X(x) > 0. \tag{2.7}$$

We also assume that m is differentiable, and that

$$\sup_{x\in D_X} |m'(x)| < \infty, \tag{2.8}$$

$$\sup_{y\in I\!R} f_\varepsilon(y) =: C < \infty. \tag{2.9}$$

We consider weak convergence on $D(D_X \times \bar{I\!R})$ endowed with the supremum norm metric and the σ-field generated by the open balls in $D(D_X \times \bar{I\!R})$.

Proposition 2.1. *Under H_0 and* (2.7), (2.8) *and* (2.9),

$$\sqrt{n}(F_n(x,y) - \hat{F}_X(x)\hat{G}(y)), \quad x \in D_X, y \in I\!R,$$

converges weakly to $V_0(x,y)$, $x \in D_X$, $y \in I\!R$.

Clearly, by the Continuous Mapping Theorem, Proposition 2.1 provides the weak convergence under H_0 of a myriad of possible test statistics. In Theorem 2.2 we deal with weak convergence of the test statistics in (2.2)−(2.4), using Proposition 2.1.

Proof of Proposition 2.1. $X_1,\ldots,X_n$ and $\varepsilon_1,\ldots,\varepsilon_n$ are two independent i.i.d. samples. Denote by $R_1,\ldots,R_n$ the ranks of $X_1,\ldots,X_n$. Observe that $X_1,\ldots,X_n$ and $\varepsilon_{R_1},\ldots,\varepsilon_{R_n}$ are also two independent i.i.d. samples. We consider $(X_1,\varepsilon_{R_1}),\ldots,(X_n,\varepsilon_{R_n})$. (Recall that n is actually $n+2$ here.) These are i.i.d. random vectors with independent components; clearly ε_{R_i} has df F_ε. Now we redefine our Y_i through $Y_i = m(X_i) + \varepsilon_{R_i}$. Obviously the new data have the same probability distribution as the original ones, so

$$\begin{aligned} F_n(x,y) = \frac{1}{n}\sum_{j=1}^{n} I\Big(&X_{j:n} \le x, m(X_{j-1:n}) - 2m(X_{j:n}) + m(X_{j+1:n}) \\ &+\varepsilon_{j-1} - 2\varepsilon_j + \varepsilon_{j+1} \le y\Big). \end{aligned}$$

First we show that $F_n(x,y)$ can be approximated by

$$\begin{aligned} \widetilde{F}_n(x,y) &= \frac{1}{n}\sum_{j=1}^{n} I(X_{j:n} \le x, \varepsilon_{j-1} - 2\varepsilon_j + \varepsilon_{j+1} \le y) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x, \varepsilon_{R_i-1} - 2\varepsilon_{R_i} + \varepsilon_{R_i+1} \le y). \end{aligned}$$

Using (2.7) we obtain $\max_{0\le j\le n}(X_{j+1:n} - X_{j:n}) = O_P(\log n/n)$. This in combination with (2.8) yields

$$\begin{aligned} &\max_{0\le j\le n} |m(X_{j+1:n}) - m(X_{j:n})| \\ &\quad = \sup_{x\in D_X} |m'(x)| O_P\Big(\frac{\log n}{n}\Big) = O_P\Big(\frac{\log n}{n}\Big) = o_P\Big(\frac{\log^2 n}{n}\Big). \end{aligned}$$

So with arbitrarily high probability for large n

$$F_n(x,y) \le \widetilde{F}_n\Big(x, y + \frac{\log^2 n}{n}\Big).$$

Set $F(x,y) = F_X(x)G(y)$. Then

$$\alpha_n(x,y) := \sqrt{n}(F_n(x,y) - F(x,y))$$

$$\begin{aligned}&\le \sqrt{n}\Big(\widetilde{F}_n\Big(x,y+\frac{\log^2 n}{n}\Big)-F\Big(x,y+\frac{\log^2 n}{n}\Big)\Big)\\&\quad+\sqrt{n}\Big(F\Big(x,y+\frac{\log^2 n}{n}\Big)-F(x,y)\Big)\\&=:\widetilde{\alpha}_n\Big(x,y+\frac{\log^2 n}{n}\Big)+\sqrt{n}F_X(x)\Big(G\Big(y+\frac{\log^2 n}{n}\Big)-G(y)\Big).\end{aligned}$$

From (2.9) and (2.5) we see that

$$\sup_{\substack{x\in D_X\\ y\in I\!R}}\sqrt{n}F_X(x)\Big(G\Big(y+\frac{\log^2 n}{n}\Big)-G(y)\Big)\le C\frac{\log^2 n}{\sqrt{n}}.$$

Then, with arbitrarily high probability for large n, and uniformly in x and y,

$$\left.\begin{aligned}\alpha_n(x,y)&\le\widetilde{\alpha}_n\Big(x,y+\frac{\log^2 n}{n}\Big)+C\frac{\log^2 n}{\sqrt{n}}\\ \alpha_n(x,y)&\ge\widetilde{\alpha}_n\Big(x,y-\frac{\log^2 n}{n}\Big)-C\frac{\log^2 n}{\sqrt{n}}\end{aligned}\right\},\tag{2.10}$$

where the latter inequality follows similarly.

We next consider the weak convergence of $\widetilde{\alpha}_n(x,y)=\sqrt{n}(\widetilde{F}_n(x,y)-F(x,y))$, where $\widetilde{F}_n(x,y)=n^{-1}\sum_{j=1}^n I(X_{j:n}\le x,V_j\le y)$, with $V_j=\varepsilon_{j-1}-2\varepsilon_j+\varepsilon_{j+1}$. Clearly the V_j are 2-dependent. Now $\widetilde{F}_n(x,y)=n^{-1}\sum_{j=1}^{n\hat{F}_X(x)}I(V_j\le y)$. Write

$$\hat{G}_x(y)=\frac{1}{\lfloor nx\rfloor}\sum_{j=1}^{\lfloor nx\rfloor}I(V_j\le y),$$

and observe that $\widetilde{F}_n(x,y)=\hat{F}_X(x)\hat{G}_{\hat{F}_X(x)}(y)$. Define also, for $0<z\le n$,

$$\begin{aligned}\widetilde{\alpha}_{2z}(y)&=\frac{1}{\sqrt{\lfloor z\rfloor}}\sum_{j=1}^{\lfloor z\rfloor}\Big(I(V_j\le y)-G(y)\Big)\quad(0/0=0),\\ Z_n(x,y)&=\sqrt{\frac{\lfloor nx\rfloor}{n}}\widetilde{\alpha}_{2nx}(y),\ 0\le x\le 1,\ y\in I\!R.\end{aligned}$$

Note that $Z_n(1,y)=\widetilde{\alpha}_{2n}(y)$. Since the V_j are 2-dependent, $Z_n(x,y)$ can be written as the sum of three dependent sequential empirical processes based on i.i.d. rv's. So Z_n is tight. It remains to prove the weak convergence of the finite dimensional distributions. Consider $(x_1,y_1),\dots,(x_k,y_k)$, with $x_1\le x_2\le\cdots\le x_k$. By the Cramér-Wold device it suffices to consider linear combinations, i.e.

$\sum_{r=1}^{k} a_r Z_n(x_r, y_r)$. Now using the Central Limit Theorem for triangular arrays of m-dependent rv's and the fact that

$$\begin{aligned}&\sum_{r=1}^{k} a_r Z_n(x_r, y_r)\\ &\quad= \sum_{r=1}^{k} a_r Z_n(x_1, y_r) + \sum_{r=2}^{k} a_r (Z_n(x_2, y_r) - Z_n(x_1, y_r))\\ &\qquad + \cdots + \sum_{r=k}^{k} a_r (Z_n(x_k, y_r) - Z_n(x_{k-1}, y_r)),\end{aligned}$$

where these k terms are almost independent, we see that $\sum_{r=1}^{k} a_r Z_n(x_r, y_r)$ converges weakly. In summary, since g is bounded (use (2.9) and (2.6)), Z_n converges weakly on $D([0,1] \times \bar{I\!R})$ to a centered, uniformly continuous, bounded Gaussian process Z with covariance structure

$$\begin{aligned}&E(Z(x_1, y_1)Z(x_2, y_2))\\ &\quad= (x_1 \wedge x_2)(G(y_1 \wedge y_2) + 2H_1(y_1, y_2) + 2H_2(y_1, y_2) - 5G(y_1)G(y_2)).\end{aligned}$$

So $\mathrm{Var}\,(Z(x, y)) = x(G(y) + 2H_1(y, y) + 2H_2(y, y) - 5G^2(y))$. Obviously $H_k(y, y) \le P(\varepsilon_1 - 2\varepsilon_2 + \varepsilon_3 \le y) = G(y)$, $k = 1, 2$. We have

$$\begin{aligned}\widetilde{\alpha}_n(x, y) &= \sqrt{n}\Big(\frac{1}{n}\sum_{j=1}^{n\hat{F}_X(x)} \Big(I(V_j \le y) - G(y)\Big)\Big)\\ &\qquad + G(y)\sqrt{n}(\hat{F}_X(x) - F_X(x)). \qquad (2.11)\end{aligned}$$

It is well known that $\sqrt{n}(\hat{F}_X - F_X)$ converges weakly to $B \circ F_X$, with B a Brownian bridge. We also have that $\sqrt{n}(\hat{F}_X - F_X)$ and Z_n are independent, and hence so are B and Z. Using the Skorohod construction (keeping the same notation for the new processes) we see that the right hand side of (2.11) is, almost surely,

$$\begin{aligned}&Z(\hat{F}_X(x), y) + G(y)B(F_X(x)) + o(1)\\ &\quad= Z(F_X(x), y) + G(y)B(F_X(x)) + o(1), \quad \text{uniformly in } x \text{ and } y.\end{aligned}$$

So $\{\tilde{\alpha}_n(x, y), x \in D_X, y \in I\!R\}$, converges weakly to

$$\Big\{Z(F_X(x), y) + G(y)B(F_X(x)), x \in D_X, y \in I\!R\Big\}.$$

Write $V(x, y) = Z(F_X(x), y) + G(y)B(F_X(x))$. Using this and the fact that V is uniformly continuous with respect to $d((x_1, y_1), (x_2, y_2)) = |F_X(x_1) - F_X(x_2)| + |y_1 - y_2|$, we see from (2.10) that α_n converges to the same limit, i.e., we have

$$\alpha_n \xrightarrow{d} V. \qquad (2.12)$$

In particular we have that

$$\sqrt{n}(\hat{F}_X - F_X) \xrightarrow{d} V(\cdot,\infty) \overset{d}{=} B \circ F_X, \tag{2.13}$$

and, similarly, with $\alpha_{2n}(y) = \alpha_n(\infty, y)$,

$$\alpha_{2n} \xrightarrow{d} V(\infty,\cdot) \overset{d}{=} Z(1,\cdot). \tag{2.14}$$

Since $\sqrt{n}(F_n(x,y) - \hat{F}_X(x)\hat{G}(y)) = \sqrt{n}(F_n(x,y) - F(x,y)) - G(y)\sqrt{n}(\hat{F}_X(x) - F_X(x)) - \hat{F}_X(x)\sqrt{n}(\hat{G}(y) - G(y))$, we obtain from (2.12), (2.13), (2.14), $\sqrt{n}(F_n - \hat{F}_X\hat{G}) \xrightarrow{d} V - G(y)V(\cdot,\infty) - F_X(x)V(\infty,\cdot) = Z(F_X,\cdot) - F_X Z(1,\cdot) =: V_0$.

Remark 2.2. Note that our testing procedure can in principle also be used for testing independence of ε and X in the nonparametric heteroscedastic model $Y = m(X) + \sigma(X)\varepsilon$, with σ an unknown, smooth, scale curve. To this end the expression $Y_{[j-1:n]} - 2Y_{[j:n]} + Y_{[j+1:n]}$ in (2.1) needs to be replaced by an expression where the function σ also vanishes for neighboring X-values, e.g., $(Y_{[j-1:n]} - Y_{[j:n]})/(Y_{[j+1:n]} - Y_{[j+2:n]})$.

Theorem 2.2. *Under* H_0 *and* (2.7), (2.8) *and* (2.9),

$$T_{n,KS} \xrightarrow{d} \sup_{\substack{x\in D_X \\ y\in I\!R}} |V_0(x,y)|, \tag{2.15}$$

$$T_{n,CM} \xrightarrow{d} \iint V_0^2(x,y)\,dF_X(x)\,dG(y), \tag{2.16}$$

$$T_{n,AD} \xrightarrow{d} \iint \frac{V_0^2(x,y)}{F_X(x)G(y)(1-F_X(x))(1-G(y))} dF_X(x)\,dG(y). \tag{2.17}$$

Proof. Statement (2.15) is immediate from Proposition 2.1 and statement (2.16) follows easily from Proposition 2.1 and the Helly-Bray Theorem.

The detailed proof of (2.17) is given in a supplement to this paper; here we just present an outline.

Set $V_{n,0} = \sqrt{n}(F_n - \hat{F}_X\hat{G})$. From (2.12) and Proposition 2.1, we have, using the Skorohod construction for (2.12) (but keeping the same notation),

$$\sup_{\substack{x\in D_X \\ y\in I\!R}} |\alpha_n(x,y) - V(x,y)| \to 0 \ \text{ a.s.}, \tag{2.18}$$

$$\sup_{\substack{x\in D_X \\ y\in I\!R}} |V_{n,0}(x,y) - V_0(x,y)| \to 0 \ \text{ a.s..} \tag{2.19}$$

Set $M(x,y) = F_X(x)G(y)(1-F_X(x))(1-G(y))$ and $\hat{M}(x,y) = \hat{F}_X(x)\hat{G}(y)(1 - \hat{F}_{X-}(x))(1-\hat{G}_-(y))$. Let $0 < \varepsilon < 1/4$ be arbitrary and let $\delta(\varepsilon) > 0$ be a function of ε such that $\lim_{\varepsilon\downarrow 0}\delta(\varepsilon)=0$. Denote by $q_{1\varepsilon}$ and $\widetilde{q}_{1\varepsilon}$ the $\delta(\varepsilon)$-th and $(1-\delta(\varepsilon))$-th quantiles of F_X, respectively, and by $q_{2\varepsilon}$, $\widetilde{q}_{2\varepsilon}$ the same quantiles of G. Write $S_\varepsilon=(q_{1\varepsilon},\widetilde{q}_{1\varepsilon})\times(q_{2\varepsilon},\widetilde{q}_{2\varepsilon})$. We have

$$\begin{aligned}
&\Big|\iint_{S_\varepsilon}\frac{V_{n,0}^2(x,y)}{\hat{M}(x,y)}d\hat{F}_X(x)d\hat{G}(y)-\iint_{S_\varepsilon}\frac{V_0^2(x,y)}{M(x,y)}dF_X(x)dG(y)\Big|\\
&\quad\le\iint_{S_\varepsilon}\frac{|V_{n,0}^2(x,y)-V_0^2(x,y)|}{\hat{M}(x,y)}d\hat{F}_X(x)d\hat{G}(y)\\
&\quad\quad+\iint_{S_\varepsilon}\frac{|M(x,y)-\hat{M}(x,y)|}{\hat{M}(x,y)M(x,y)}V_0^2(x,y)d\hat{F}_X(x)d\hat{G}(y)\\
&\quad\quad+\Big|\iint_{S_\varepsilon}\frac{V_0^2(x,y)}{M(x,y)}(d\hat{F}_X(x)d\hat{G}(y)-dF_X(x)dG(y))\Big|.
\end{aligned}$$

From (2.19) and (2.18) we now see that the first and second term on the right converge to 0 a.s. The a.s. convergence to 0 of the third term follows from the Helly-Bray Theorem.

Set $A_\varepsilon=I\!R^2\backslash S_\varepsilon$. In view of what we just proved, it is now sufficient for the proof of (2.17) to show that for large n and appropriate $\delta(\varepsilon)$

$$P\Big(\iint_{A_\varepsilon}\frac{V_{n,0}^2(x,y)}{\hat{M}(x,y)}d\hat{F}_X(x)d\hat{G}(y)\ge\varepsilon\Big)\le\varepsilon,$$

$$P\Big(\iint_{A_\varepsilon}\frac{V_0^2(x,y)}{M(x,y)}dF_X(x)dG(y)\ge\varepsilon\Big)\le\varepsilon.$$

The second inequality follows rather easily from $E(V_0^2(x,y))/M(x,y)\le 5$ and the Markov inequality. The first one requires a long proof using weighted empirical process theory.

3. Simulations

Suppose that X has a uniform-$(0,1)$ distribution and that $m(x)=x-0.5x^2$. The simulations are carried out for samples of size $n=200$ and 500, and the significance level $\alpha=0.05$. Each simulation consists of 2,000 replications for $n=200$, and of 1,000 replications for $n=500$.

To obtain the critical values for the test statistics $T_{n,KS},T_{n,CM}$ and $T_{n,AD}$, recall that $V_0(x,y)$ can be written as

$$V_0(x,y)=Z(F_X(x),y)-F_X(x)Z(1,y). \qquad (3.1)$$

To simulate an 'estimated' version of Z (for G, H_1 and H_2 are unknown), first partition the interval $[0,1]$ by means of r_x equidistant points $x_k=k/r_x$ $(k=$

$1, \ldots, r_x$) and use a grid of r_y points y_ℓ ($\ell = 1, \ldots, r_y$) on the real line. Then, simulate r_x i.i.d. r_y-variate normal random vectors $Z_k = (Z_{k1}, \ldots, Z_{kr_y})$ ($k = 1, \ldots, r_x$) with zero mean and covariance matrix

$$\begin{aligned}\mathrm{Cov}\,(Z_1) = \Big(r_x^{-1}\Big[&\hat{G}(y_i \wedge y_j) + \hat{H}_1(y_i, y_j) + \hat{H}_2(y_i, y_j) + \hat{H}_1(y_j, y_i) + \hat{H}_2(y_j, y_i)\\ &-5\hat{G}(y_i)\hat{G}(y_j)\Big]\Big)_{i,j=1}^{r_y},\end{aligned}$$

with $\hat{G}$ as in Section 2, and where

$$\begin{aligned}\hat{H}_1(y_1, y_2) &= \frac{1}{n-1}\sum_{j=1}^{n-1} I\Big(Y_{[j-1:n]} - 2Y_{[j:n]} + Y_{[j+1:n]} \le y_1,\\ &\qquad Y_{[j:n]} - 2Y_{[j+1:n]} + Y_{[j+2:n]} \le y_2),\\ \hat{H}_2(y_1, y_2) &= \frac{1}{n-2}\sum_{j=1}^{n-2} I\Big(Y_{[j-1:n]} - 2Y_{[j:n]} + Y_{[j+1:n]} \le y_1,\\ &\qquad Y_{[j+1:n]} - 2Y_{[j+2:n]} + Y_{[j+3:n]} \le y_2).\end{aligned}$$

Note that $Z_1, \ldots, Z_{r_x}$ can be simulated by using $Z_k = \sqrt{\mathrm{Cov}\,(Z_1)}(W_1^{(k)}, \ldots, W_{r_y}^{(k)})'$ ($k = 1, \ldots, r_x$), where $W_1^{(k)}, \ldots, W_{r_y}^{(k)}$ are independent standard normal random variables. The process Z is now approximated by the $(r_x \times r_y)$-variate random vector $\tilde{Z}(x_k, y_\ell) = \sum_{j=1}^{k} Z_{j\ell}$. Hence V_0 can be approximated by using the approximation of Z and by replacing F_X with $\hat{F}_X$ in (3.1). After repeating this procedure a large number of times, the critical values of the three tests can be approximated very well.

We consider four types of distributions. For the first three, the null model corresponds to a normal error term with zero mean and standard deviation equal to 0.1, and we take m to be the conditional mean of Y given X. In the fourth case, the error term has a standard Cauchy distribution under the null hypothesis; here m is the conditional median. Consider for the four cases the following alternative hypotheses:

$$H_{1,A} : \varepsilon \mid X = x \sim N\Big(0, \frac{1 + ax}{100}\Big),$$

with $a > 0$. Also, let

$$H_{1,B} : \varepsilon \mid X = x \overset{d}{=} \frac{W_x - s_x}{10\sqrt{2s_x}},$$

where $W_x \sim \chi^2_{s_x}$, $s_x = 1/(bx)$ and $b > 0$ controls the skewness of the distribution. Note that the first and second moment of the variable ε created in the latter fashion do not depend on x, and coincide with the respective moments under H_0. When b tends to 0, the distribution of $\varepsilon|X = x$ converges to its null distribution, since it is well known that a standardized χ^2_s-distribution converges to a normal distribution when $s \to \infty$. Next, let

$$H_{1,C} : \varepsilon \mid X = x \sim \frac{1}{10}\sqrt{1-(cx)^{1/4}}\, t_{2/(cx)^{1/4}},$$

where $0 < c \leq 1$ is a parameter controlling the kurtosis (which might be infinite) of the distribution. By construction, the conditional moments up to order three of ε given X are constant and coincide with the respective moments under the null hypothesis, while the fourth conditional moment does depend on X (note that the third and fourth moment do not need to exist). The distribution of ε under $H_{1,C}$ converges to the null distribution of ε when c tends to 0. The last type of error variables we consider follow a Cauchy distribution. Let

$$H_{1,D} : f_\varepsilon(v|x) = \frac{1}{(1+dx)\pi\{1+(\frac{v}{1+dx})^2\}},$$

where $d > -1$ controls the scale, and $f_\varepsilon(\cdot|x)$ represents the conditional density of ε given $X = x$. Clearly, the case $d = 0$ corresponds to the null hypothesis of a standard Cauchy distribution.

We compare the proposed tests with the test for homoscedasticity considered by Dette and Munk (1998). The latter test is suitable for detecting deviations from H_0 under alternative $H_{1,A}$ (heteroscedasticity), but not under the homoscedastic alternatives $H_{1,B}$ and $H_{1,C}$. Under $H_{1,D}$, the conditional variance of ε given X does not exist, and the test of Dette and Munk (1998) is not intended to work in this case.

Tables 1−4 show the results of the simulations under $H_{1,A}$, $H_{1,B}$, $H_{1,C}$, and $H_{1,D}$, respectively. We observe that the empirical α-levels (see $a, b, c, d = 0$) are reasonably close to their nominal value of 0.05, except for the Dette-Munk test which is conservative for the Cauchy distribution (see above), and except for the Anderson-Darling statistic which is conservative for the normal distribution (but the α-level does converge to the nominal level for large sample sizes - for $n = 800$ it is 0.048). Despite this conservatism, the power in Table 1 is highest for the Anderson-Darling statistic and is lowest for the Dette-Munk test. So, although the Dette-Munk test is a test for homoscedasticity and the proposed test is more of an omnibus test, the latter outperforms the former. For $H_{1,B}$ and $H_{1,C}$ the Crámer-von Mises test outperforms the other tests. Note that for the Dette-Munk test the null hypothesis of homoscedasticity holds.

Finally, for the (difficult) case of the Cauchy distribution, all three proposed tests perform well; the Crámer-von Mises test again performs best. The Dette-Munk test is not appropriate here.

Table 1. Power of $T_{n,KS}$, $T_{n,CM}$ and $T_{n,AD}$ and the test of Dette and Munk (DM) under $H_{1,A}$.

a	$n = 200$				$n = 500$			
	KS	CM	AD	DM	KS	CM	AD	DM
0	0.048	0.049	0.036	0.044	0.051	0.049	0.041	0.051
1	0.123	0.170	0.169	0.081	0.305	0.401	0.458	0.100
2.5	0.305	0.431	0.460	0.140	0.744	0.857	0.911	0.223
5	0.497	0.674	0.703	0.211	0.944	0.989	0.996	0.365
10	0.673	0.855	0.871	0.265	0.997	0.999	1.000	0.486
100	0.843	0.972	0.979	0.359	1.000	1.000	1.000	0.646

Table 2. Power of $T_{n,KS}$, $T_{n,CM}$ and $T_{n,AD}$ and the test of Dette and Munk (DM) under $H_{1,B}$.

b	$n = 200$				$n = 500$			
	KS	CM	AD	DM	KS	CM	AD	DM
0	0.048	0.049	0.036	0.044	0.051	0.049	0.041	0.051
1	0.105	0.166	0.112	0.070	0.300	0.397	0.292	0.067
2.5	0.259	0.417	0.286	0.069	0.727	0.870	0.772	0.081
5	0.467	0.701	0.569	0.064	0.936	0.994	0.982	0.054
10	0.701	0.893	0.826	0.056	0.996	1.000	0.998	0.045
100	0.932	0.999	0.998	0.051	1.000	1.000	1.000	0.033

Table 3. Power of $T_{n,KS}$, $T_{n,CM}$ and $T_{n,AD}$ and the test of Dette and Munk (DM) under $H_{1,C}$.

c	$n = 200$				$n = 500$			
	KS	CM	AD	DM	KS	CM	AD	DM
0	0.048	0.049	0.036	0.044	0.051	0.049	0.041	0.051
0.2	0.063	0.086	0.062	0.056	0.120	0.146	0.135	0.046
0.4	0.114	0.166	0.134	0.063	0.287	0.370	0.339	0.050
0.6	0.215	0.313	0.261	0.069	0.589	0.699	0.666	0.055
0.8	0.438	0.582	0.509	0.087	0.878	0.946	0.945	0.063
1.0	0.815	0.949	0.937	0.126	0.999	1.000	1.000	0.104

Table 4. Power of $T_{n,KS}$, $T_{n,CM}$ and $T_{n,AD}$ and the test of Dette and Munk (DM) under $H_{1,D}$.

d	$n = 200$				$n = 500$			
	KS	CM	AD	DM	KS	CM	AD	DM
0	0.035	0.045	0.037	0.020	0.048	0.046	0.054	0.013
1	0.139	0.193	0.153	0.022	0.340	0.440	0.401	0.013
2.5	0.364	0.516	0.430	0.023	0.822	0.903	0.863	0.018
5	0.573	0.753	0.688	0.026	0.965	0.991	0.989	0.019
10	0.739	0.884	0.849	0.025	0.996	1.000	1.000	0.020
100	0.901	0.988	0.975	0.027	1.00	1.000	1.000	0.020

Table 5. P-values for the household data.

Test	Y_1	Y_2
KS	0.027	0.980
CM	0.002	0.770
AD	0.002	0.561

4. Data analysis

The data we consider consist of monthly expenditures in Dfl. of Dutch households on several commodity categories, as well as a number of background variables (Dfl. = Dutch guilders, 1 Dfl. is about € 0.45). We use expenditures on food and total expenditures accumulated over the year from October 1986 through September 1987, and take households consisting of two persons; the sample size is 159. The data have been extracted from the Data Archive of the Journal of Applied Econometrics and have been analyzed in Adang and Melenberg (1995).

We want to regress two responses to the regressor $X = \log(\text{total expenditures})$, namely Y_1 = share of food expenditure in household budget and $Y_2 = \log(\text{expenditure on food per household})$, according to (1.1)−(1.2). In order to see if this model is appropriate we use our tests of Section 2. The P-values of the tests are presented in Table 5.

This table shows that model (1.1)−(1.2) is violated by Y_1, but not by Y_2. Hence this model can be used for further analysis of the log food expenditure data. Knowing the independence of X and ε for this case makes it possible to use statistical methods that outperform procedures that use only homoscedasticity.

Acknowledgement

We like to thank Arthur van Soest for pointing out the data on food expenditures, and for very helpful discussions concerning the corresponding data analysis.

References

Adang, P. J. M. and Melenberg, B. (1995). Nonnegativity constraints and intratemporal uncertainty in multi-good life-cycle models. *J. Appl. Econometrics* **10**, 1-15.

Akritas, M. G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549-568.

Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32**, 485-498.

Cao, R. and Gijbels, I. (2005). Testing homoscedasticity via the integrated conditional variance in nonparametric regression. In preparation.

Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *J. Roy. Statist. Soc. Ser. B* **60**, 693-708.

Dette, H., Munk, A. and Wagner, T. (1998). Estimating the variance in nonparametric regression - what is a reasonable choice? *J. Roy. Statist. Soc. Ser. B* **60**, 751-764.

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Statist.* **19**, 546-547.

Lee, B.-J. (1992). A heteroskedasticity test robust to conditional mean misspecification. *Econometrica* **60**, 159-171.

Liero, H. (2003). Testing homoscedasticity in nonparametric regression. *J. Nonparametr. Stat.* **15**, 31-51.

Müller, U., Schick, A. and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* **37**, 179-188.

Neumeyer, N., Dette, H. and Nagel, E.-R. (2006). Bootstrap tests for the error distribution in linear and nonparametric regression models. *Austral. N. Z. J. Statist.* **48**, 129-156

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Van Keilegom, I. and Akritas, M. G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.* **27**, 1745-1784.

Van Keilegom, I., González Manteiga, W. and Sánchez Sellero, C. (2007). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *TEST* (to appear).

Department of Econometrics & OR and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands.

E-mail: j.h.j.einmahl@uvt.nl

Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium.

E-mail: vankeilegom@stat.ucl.ac.be

(Received February 2006; accepted September 2006)