Working Paper 92-15
April 1992

División de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9875

TESTING THE EQUALITY OF NONPARAMETRIC REGRESSION CURVES

Miguel A. Delgado*

Abstract

This paper proposes a test for the equality of nonparametric regression curves that does not depend on the choice of a smoothing number. The test statistic is a weighted empirical process easy to compute. It is powerful under alternatives that converge to the null at a rate $n^{-1/2}$. The disturbance distributions are arbitrary and possibly unequal, and conditions on the regressors distribution are very mild. A simulation study demonstrates that the test enjoys good level and power properties in small samples. We also study extensions to multiple regression, and testing the equality of several regression curves.

Key words:
Nonparametric testing; weighted empirical process; Donsker's invariance principle; Brownian motion; local alternatives.

* Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

1. INTRODUCTION.

This article proposes a test for the equality of regression curves of unknown functional form. The problem is well motivated and has been already investigated using smooth nonparametric estimates of the regression curve: King (1989), Härdle and Marron (1990), Hall and Hart (1990) and King et al (1991) among others. These tests need to choose a smoothing constant for constructing the nonparametric estimates, and their power properties generally depend on this choice. Statistics based on automatically chosen smoothing numbers are computationally demanding, and their asymptotic properties are difficult to justify.

Among the above mentioned papers, Hall and Hart (1990) (HH) is possibly the most attractive from the practical point of view. They proposed a bootstrap test quite insensitive to the choice of the smoothing number under the null hypothesis. The finite sample level and nominal level are almost identical for different choices of smoothing numbers and sample sizes as small as 15. The power of the test crucially depends on the smoothing constant choice.

The test statistic proposed in this paper resembles in spirit the Komolgorov-Smirnov. It is easy to compute and performs well in finite samples.

We observe a random sample $\{(X_i, Y_i, Z_i), i=1,\ldots,n\}$ of the random variable (X, Y, Z). The variables Y and Z are related to X according to the regression model

$$E(Y|X=x)=q_y(x) \text{ and } E(Z|X=x)=q_Z(x).$$

The marginal distribution function of X is continuous, and $q_Y(.)$ and $q_Z(.)$ are continuous on $\mathcal{X}$, where $Pr(X\in\mathcal{X})=1$. We also assume that the regression errors, $Z-q_Z(X)$ and $Y-q_Y(X)$, are independent of X and may have different distributions, $E|q_y(X)|^2<\infty$, $E|q_Z(X)|^2<\infty$, $0<\sigma_Y^2=E|Y-q_Y(X)|^2<\infty$ and $0<\sigma_Z^2=E|Z-q_Z(X)|^2<\infty$.

The hypothesis to be tested are

$$H_0: q_y(x)=q_Z(x) \text{ for all } x\in\mathcal{X},$$

versus

$$H_1: q_y(x)\neq q_Z(x) \text{ for some } x\in\mathcal{X}.$$

The regressors may be fixed. In this case, we assume that they are coming from the unit interval (or a bounded interval). It is also assumed that the regressors become dense in the observation interval as the sample size increases, the regression function have a bounded derivative in the observation interval, the regression errors are independent and do not depend on the regressors, and the error variances are bounded and positive.

Section 2 presents the test statistic and discusses its asymptotic properties. Section 3 reports the numerical results. Section 4 contains final remarks, including generalizations to multiple regression and testing the equality of several regression functions.

2. TEST STATISTIC

A necessary and sufficient condition for the null hypothesis to hold is that

$$\sup_{-\infty<t<\infty} \left| \int_{-\infty}^{t} (q_Y(x) - q_Z(x))\, f(x)\, dx \right| = 0, \qquad (2.1)$$

where $f(.)$ is the density function of X. Define $D_i = Y_i - Z_i$, the weighted empirical process

$$\sup_{-\infty<t<\infty} \left| n^{-1} \sum_{j=1}^{n} D_j\, 1(X_j < t) \right|,$$

consistently estimates the left hand side (L.H.S) of (2.1), where $1(A)$ is the indicator function of the event A. Then, we propose the test statistic

$$T_n = \left(\sum_{j=1}^{n-1} (D_{j+1} - D_j)^2 / 2 \right)^{-1/2} \sup_{-\infty<t<\infty} \left| \left(\sum_{j=1}^{n} D_j\, 1(X_j < t) \right) \right|,$$

which will take large values under H_1 and small values under H_0. A similar type of statistic has been used by Hong-zhi and Bing (1991) for testing linearity in regression models.

The statistic is easy to compute. Let $r_{1n},\ldots,r_{nn}$ be the antiranks of $X_1,\ldots,X_n$ defined as $X_{r_{1n}} > X_{r_{2n}} > \ldots > X_{r_{nn}}$. Then note that,

$$\sup_{-\infty<t<\infty} \left|\sum_{i=1}^{n} D_i\, 1(X_i< t)\right| = \sup_{1\le k\le n} \left|\sum_{i=1}^{k}(Y_{r_{in}} - Z_{r_{in}})\right|.$$

Applying Kolmogorov's law of large numbers (KLLN)

$$(2n)^{-1}\sum_{i=1}^{n-1} (D_{j+1} - D_j)^2 \longrightarrow \sigma^2 + \mathrm{Var}(q_Y(X) - q_Z(X)) \text{ w.p.1 as } n\to\infty \qquad (2.2)$$

where 'w.p.1' means convergence with probability 1, $\sigma^2 = \sigma_Y^2 + \sigma_Z^2 - 2\sigma_{YZ}$, and $\sigma_{YZ} = E((Y - q_Y(X))(Z - q_Z(X)))$. If the regressors are fixed, and $\max_i(X_{i+1} - X_i) = 0$ as $n\longrightarrow\infty$, the L.H.S of (2.2) converges to σ^2 w.p.1. as $n\longrightarrow\infty$, under H_0 and H_1. In this case, the usual variance estimate $n^{-1}\sum_{i=1}^{n}(D_i - n^{-1}\sum_{i=1}^{n} D_i)^2$ converges to a probabilistic limit which dominates σ^2 under H_1. The scale estimate on the L.H.S. of (2.2) has been also used by Rice (1984), Hall and Hart (1990), and King et. al. (1991).

By KLLN

$$n^{-1}\sum_{j=1}^{n} D_j\, 1(X_j< t) \longrightarrow C(t) = \int_{-\infty}^{t} (q_Y(x) - q_Z(x))\, f(x)\, dx \text{ w.p.1. as } n\to\infty.$$

Since $C(t) > 0$ for some t under the alternative hypothesis, T_n diverges to infinity at a rate $n^{1/2}$.

In order to investigate the asymptotic distribution of the test statistic under H_0, define $\varepsilon_{Yi} = Y_i - q_Y(X_i)$ and $\varepsilon_{Zi} = Z_i - q_Z(X_i)$, $1\le i\le n$. Since errors are independent of regressors, $(\varepsilon_{Yr_{in}} - \varepsilon_{Zr_{in}})$, $i > 1$, are iid with mean zero and variance σ^2, Donsker's theorem (see Billingsley 1968) establishes that, under H_0,

$$\sup_{-\infty<t<\infty} \left|(n\sigma^2)^{-1/2}\sum_{i=1}^{n} D_i\, 1(X_i< t)\right|$$

$$= \sup_{1\le k\le n} \left|(n\sigma^2)^{-1/2}\sum_{i=1}^{k}(\varepsilon_{Yr_i} - \varepsilon_{Zr_i})\right|$$

$$\xrightarrow{d} T = \sup_{0\le t\le 1} |B(t)| \text{ as } n\longrightarrow\infty,$$

where '$\xrightarrow{d}$' denotes weak convergence in distribution, and B(t) is a standard Brownian motion. Define T_α such that $\Pr(T > T_\alpha) = \alpha$, then

$$\lim_{n\to\infty} \Pr(T_n > T_\alpha) = \alpha \text{ under } H_0, \quad \text{and} \quad \lim_{n\to\infty} \Pr(T_n > T_\alpha) = 1 \text{ under } H_1.$$

The null hypothesis H_0 will be rejected at the α-level of significance when $T_n > T_\alpha$. The critical values and p-values can be easily obtained using the fact that

$$\Pr(T > b) = 1 - 4\,\pi^{-1} \sum_{j=0}^{\infty} (-1)^j (2j+1)^{-1} \exp\{-(2j+1)^2\,\pi^2/(8\,b^2)\},\ b > 0,$$

(see Shorack and Wellner 1986). Then, $T_{.1} = 1.96$, $T_{.05} = 2.2414$, $T_{.01} = 2.8074$ and $T_{.001} = 3.4808$.

Consider local alternatives in the fixed design case

$$H_{1n}: q_Y(x) - q_Z(x) = n^{-1/2}\, c\, |\ell(x)| \text{ for some } x \in [0, 1],$$

where $\ell(.)$ is a continuous fixed function and c is a constant. Under H_{1n}

$$T_n \xrightarrow{d} \sup_{0\le t\le 1} \left| c\,(\sigma^2)^{-1/2} \int_0^t |\ell(x)|\, f(x)\, dx + B(t) \right| \text{ as } n \to \infty.$$

and $\int_0^t |\ell(x)|\, f(x)\, dx = 0$ if and only if $\ell(x) = 0$ for all x. Then T_n diverges to ∞ as $|c| \longrightarrow \infty$. The test is asymptotically powerful under alternatives converging to the null at rate $n^{-1/2}$. This type of local alternatives does not have much sense when the regressors are random.

3. A MONTE CARLO STUDY

The fist part of these simulations are based on the same design employed by HH. The observations are generated according to the model

$$Y_i = q_Y(X_i) + U_{Yi} \text{ and } Z_i = q_Z(X_i) + U_{Zi},\ i = 1,\ldots,n. \tag{3.1}$$

Let N_1 and N_2 be two independent standard normal variables. The three choices for the distribution of the errors (U_{Yi}, U_{Zi}) were: (a) (N_1, N_2), (b) $(|N_1| - (2/\pi)^{1/2},\ |N_2| - (2/\pi)^{1/2})$, (c) $(|N_1| - (2/\pi)^{1/2},\ (2/\pi)^{1/2} - |N_2|)$. All distribution have zero mean. In (a) and (b) the two errors have the same distribution and in (c) the error distributions are skewed in opposite

directions. The regressors are fixed and evenly spaced, that is $X_i = i/n$. In each case, five sample sizes are used n= 15, 20, 30, 50, 100. For each sample size the proportion of rejections of the null hypothesis in 5000 replications is reported when $q_Y(x) - q_Z(x) = \ell(x)$, and (i) $\ell(x) = 0$, (i) $\ell(x) = 1/2$, (iii) $\ell(x) = 1$, (iv) $\ell(x) = x/2$, and (v) $\ell(x) = x$.

Table 1 reports the proportion of rejections under (i)-(v) and errors distributions (a)-(c). The level of the test is good when n= 50, 100. The bootstrap HH's test always performs better under the null. Under the alternative, our test also works very well for the different designs. For the three values of the smoothing constant chosen in HH, our test is at least as powerful as HH's test.

In a second set of experiments, observations are generated according to (3.1), but the design is random, $X_i \sim$ iid N(0,1). Table 2 reports the proportion of rejections, in 5000 replications, under (i)-(v), and (vi) $\ell(x) = 2x$, and under the error distribution (a), which has been the least favorable in terms of power. At each replication new regressors and errors are generated. The test performance is still good under the null and alternatives (ii) and (iii). Under alternatives (iv) and (v), power is lower than in Table 1, because (iv) and (v) are much closer to the null than in the above set of experiments. This is why we also report results for alternative (vi). For this alternative, the power of the test is reasonably good.

4. FINAL REMARKS

We have obtained an asymptotic test for detecting a difference between nonparametric regression curves that works well in small samples, and does not depend on the choice of a smoothing number.

The test can be implemented for testing the equality of several regression curves. Suppose we observe a random sample $\{(X_i, Y_1^{(1)},\ldots,Y_n^{(p)}), i=1,\ldots,n\}$ from the random variable $(X, Y^{(1)},\ldots,Y^{(p)})$. The variables $Y^{(1)},\ldots,Y^{(p)}$ are related to X according to the regression models $E(Y^{(k)} | X=x) = q_k(x)$, $k=1,\ldots,p$.

We want to test the hypothesis

$$H_0: q_k(x) = q_m(x) \text{ all } m,k = 1,\ldots,p, \text{ and all } x \in \mathcal{X},$$

versus

$$H_1: q_k(x) \neq q_m(x) \text{ some } m \neq k,\ m,k = 1,\ldots,p, \text{ and some } x \in \mathcal{X}.$$

Define $\bar{Y}_j = p^{-1} \sum_{j=1}^{p} Y^{(k)}$, $D_j^k = Y_j^{(k)} - \bar{Y}_j$, and

$$T_n^k = \left(\sum_{j=1}^{n-1} (D_{j+1}^k - D_j^k)^2/2 \right)^{-1/2} \sup_{-\infty < t < \infty} \left| \sum_{j=1}^{n} D_j^k \, 1(X_j < t) \right|.$$

The test statistic is

$$T_n = \max_{1 \leq k \leq p} T_n^k.$$

Under the alternative hypothesis

$$n^{-1} \sum_{j=1}^{n} D_j^k \, 1(X_j < t) \longrightarrow C^k(t) \text{ w.p.1. as } n \longrightarrow \infty,$$

where

$$C^k(t) = \int_{-\infty}^{t} (q_k(x) - p^{-1} \textstyle\sum_{j=1}^{p} q_j(x)) \, f(x) \, dx,$$

and $C^k(t) > 0$ for some k and some t. Under H_0, $T_n \xrightarrow{d} T$ as $n \to \infty$.

The statistic can be used for testing necessary conditions for the equality of multiple regression curves. Suppose that $\underset{\sim}{X} = (X^1,\ldots,X^p)$ is a p-dimensional random variable and we observe a random sample $\{(\underset{\sim}{X}_i, Y_i, Z_i), i=1,\ldots,n\}$ from the random variable $(\underset{\sim}{X}, Y, Z)$. Consider the statistic

$$T_{nk} = \left(\sum_{j=1}^{n-1} (D_{j+1} - D_j)^2/2 \right)^{-1/2} \sup_{-\infty < t < \infty} \left| \sum_{j=1}^{n} D_j \, 1(X_j^k < t) \right|.$$

This statistic is valid for testing the hypothesis

$$H_{01}: E(q_y(\underset{\sim}{X}) | X^k = x) = E(q_z(\underset{\sim}{X}) | X^k = x) \text{ for all } x \in \mathcal{X}^k,$$

versus

$$H_{11}: E(q_y(\underset{\sim}{X}) | X^k = x) \neq E(q_z(\underset{\sim}{X}) | X^k = x) \text{ for some } x \in \mathcal{X}^k,$$

where $\Pr(x \in \mathcal{X}^k) = 1$. Under H_{01}, the statistic

$$T_n^0 = \max_k T_{nk} \xrightarrow{d} T \text{ as } n \to \infty.$$

This statistic is valid for testing $\max_k |E(q_y(\underset{\sim}{X})|X^k = \alpha) - E(q_z(\underset{\sim}{X})|X^k = \alpha)| = 0$ all α. This is only a necessary condition for the equality of multiple regression curves. We may also try other functions of $\underset{\sim}{X}$, say h: $\mathbb{R}^p \longrightarrow \mathbb{R}$, for testing $|E(q_y(\underset{\sim}{X})|h(\underset{\sim}{X}) = \alpha) - E(q_z(\underset{\sim}{X})|h(\underset{\sim}{X}) = \alpha)| = 0$ all α.

REFERENCES

Billingsley, P. (1968): *Convergence of Probability Measures*, New York: Wiley

Hall, P. and J.D. Hart (1990): "Bootstrap test for difference between means in nonparametric regression", *Journal of the American Statistical Association* 85, 1039-1049.

Härdle, W. and Marron, J.S. (1990): "Semiparametric comparison of regression curves", *The Annals of Statistics*, 18, 63-89.

Hong-zhi, A. and Bing, C. (1991): "A Kolmogorov-Smirnov type statistic with application to test for nonlinearity in time series", *International Statistical Review* 59, 287-307.

King, E.C. (1989): *A Test for the Equality of Two Regression Curves Based on Kernel Smoothers*, Ph.D. Dissertation, Department of Statistics, Texas A&M Univ.

King, E.C, J.D. Hart, and T.E. Wehrly. (1991): "Testing the equality of two regression curves using linear smoothers", *Statistics & Probability Letters*, 12, 239-247.

Rice, J. (1984), "Bandwidth choice for nonparametric regression", *Annals of Statistics* 12, 1215-1230.

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

TABLE 1

Proportion of rejections in 5000 replications
in the first set of experiments ($X_i = i/n$) when $\ell(x) = q_Y(x) - q_Z(x)$.

Error model (a)

		n = 15	n = 20	n = 30	n = 50	n = 100
(i) $\ell(x) = 0$	$\alpha = .05$	.0784	.0680	.0644	.0554	.0506
	$\alpha = .01$	.0262	.0210	.0222	.0168	.0126
(ii) $\ell(x) = 1/2$	$\alpha = .05$	.2922	.3448	.4422	.6598	.9242
	$\alpha = .01$	.1586	.1942	.2668	.4364	.7972
(iii) $\ell(x) = 1$	$\alpha = .05$	.7290	.8514	.9566	.9974	1.0000
	$\alpha = .01$	.5512	.6894	.8732	.9878	.9998
(iv) $\ell(x) = x/2$	$\alpha = .05$	.1538	.1660	.1928	.2612	.4698
	$\alpha = .01$	.0660	.0870	.0706	.1142	.2536
(v) $\ell(x) = x$	$\alpha = .05$	.3678	.4294	.5468	.7574	.9640
	$\alpha = .01$	.2016	.2456	.3340	.5282	.8782

Error model (b)

		n = 15	n = 20	n = 30	n = 50	n = 100
(i) $\ell(x) = 0$	$\alpha = .05$	.0744	.0704	.0602	.0574	.0496
	$\alpha = .01$	.0282	.0228	.0168	.0130	.0132
(ii) $\ell(x) = 1/2$	$\alpha = .05$	.5944	.7058	.8612	.9762	1.0000
	$\alpha = .01$	.4044	.5244	.7080	.9186	.9992
(iii) $\ell(x) = 1$	$\alpha = .05$	.9844	.9982	1.0000	1.0000	1.0000
	$\alpha = .01$	.9420	.9870	1.0000	1.0000	1.0000
(iv) $\ell(x) = x/2$	$\alpha = .05$	.2798	.3289	.4232	.6024	.8716
	$\alpha = .01$	.1428	.1668	.2344	.3646	.6996
(v) $\ell(x) = x$	$\alpha = .05$	.7124	.8706	.9300	.9920	1.0000
	$\alpha = .01$	.5120	.6256	.8048	.9636	1.0000

TABLE 1 (cont.)

Error model (c)

		n = 15	n = 20	n = 30	n = 50	n = 100
(i) $\ell(x) = 0$	$\alpha=.05$	.0876	.0762	.0714	.0570	.0566
	$\alpha=.01$	.0388	.0302	.0260	.0162	.0140
(ii) $\ell(x) = 1/2$	$\alpha=.05$	.5998	.7288	.8894	.9894	1.0000
	$\alpha=.01$	.3696	.4900	.7212	.9472	.9998
(iii) $\ell(x) = 1$	$\alpha=.05$	.9978	.9998	1.0000	1.0000	1.0000
	$\alpha=.01$	.9864	.9988	1.0000	1.0000	1.0000
(iv) $\ell(x) = x/2$	$\alpha=.05$	.2460	.2872	.3968	.5960	.8932
	$\alpha=.01$	.1072	.1260	.1936	.3368	.7186
(v) $\ell(x) = x$	$\alpha=.05$	.7466	.8526	.9576	.9982	1.0000
	$\alpha=.01$	.4988	.7372	.8412	.9840	1.0000

TABLE 2

Proportion of rejections in 5000 replicationsin the second set of experiments (X_i ~ iid N(0, 1)) when $\ell(x) = q_Y(x) - q_Z(x)$.

Error model (a)

		n = 15	n = 20	n = 30	n = 50	n = 100
(i) $\ell(x) = 0$	$\alpha = .05$	.0762	.0616	.0580	.0502	.0536
	$\alpha = .01$	.0294	.0214	.0168	.0142	.0112
(ii) $\ell(x) = 1/2$	$\alpha = .05$	.2778	.3574	.4766	.6574	.9242
	$\alpha = .01$	.1496	.1932	.2810	.4354	.7954
(iii) $\ell(x) = 1$	$\alpha = .05$	.7300	.8598	.9604	.9974	1.0000
	$\alpha = .01$	.5516	.7038	.8794	.9876	1.0000
(iv) $\ell(x) = x/2$	$\alpha = .05$	.0882	.0786	.0952	.1156	.2230
	$\alpha = .01$	.0310	.0240	.0272	.0262	.0592
(v) $\ell(x) = x$	$\alpha = .05$	.1074	.1160	.1758	.2992	.7176
	$\alpha = .01$	.0376	.0320	.0488	.0932	.3490
(vi) $\ell(x) = 2x$	$\alpha = .05$	.1496	.1912	.3502	.6632	.9920
	$\alpha = .01$	.0464	.0574	.1036	.2764	.8626