Testing the equality of nonparametric regression curves

Miguel A. Delgado

Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Spain

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 resembles in spirit the Kolmogorov–Smirnov statistic and it is easy to compute. It is powerful under alternatives that converge to the null hypothesis at a rate $n^{-1/2}$. The disturbance distributions are arbitrary and possibly unequal, and conditions on the regressors distribution are very mild. A Monte Carlo study illustrates the performance of the test in small and moderate samples. We also study extensions to multiple regression, and test the equality of several regression curves.

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

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; see 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 (see Barry and Hartigan, 1990, for an example), are computationally demanding, and their asymptotic properties are difficult to justify.

The test statistic proposed in this paper does not employ smoothing techniques. It is a weighted empirical process easy to compute, which resembles in spirit the Kolmogorov–Smirnov statistic.

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

$$E(Y \mid X = x) = g_Y(x) \quad \text{and} \quad E(Z \mid X = x) = g_Z(x).$$

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

The hypothesis to be tested is

$$H_0: \quad g_Y(x) = g_Z(x) \quad \text{for all } x \in \mathscr{X},$$

versus

$$H_1: \quad g_Y(x) \neq g_Z(x) \quad \text{for some } x \in \mathscr{X}.$$

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 has 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

$$\sup_{-\infty<t<\infty} \left| \int_{-\infty}^{t} (g_Y(x) - g_Z(x)) f(x)\, dx \right| = 0, \tag{2.1}$$

where $f(\cdot)$ 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 of (2.1), where $1(A)$ is the indicator function of the event A. Then, we propose the statistic

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

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 $X_{r_{1n}} > X_{r_{2n}} > \cdots > X_{r_{nn}}$. Then note that

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

Applying Kolmogorov's law of large numbers (KLLN),

$$(2n)^{-1} \sum_{i=1}^{n-1} (D_{j+1} - D_j)^2 \to \sigma^2 + \mathrm{Var}(g_Y(X) - g_Z(X)) \quad \text{w.p.1} \quad \text{as } n \to \infty \tag{2.2}$$

where 'w.p.1' means convergence with probability 1, $\sigma^2 = \sigma_Y^2 + \sigma_Z^2 - 2\sigma_{YZ}$, and $\sigma_{YZ} = E((Y - g_Y(X))(Z - g_Z(X)))$. If the regressors are fixed, and $\max_i(X_{i+1} - X_i) = 0$ as $n \to \infty$, the left-hand side of (2.2) converges to σ^2 w.p.1 as $n \to \infty$, under H_0 and H_1. In this case, the usual variance estimate

$n^{-1}\Sigma_{i=1}^{n}(D_i - n^{-1}\Sigma_{i=1}^{n}D_i)^2$ converges to a probabilistic limit which dominates σ^2 under H_1. The scale estimate on the left-hand side of (2.2) has been also used by Rice (1984), Hall and Hart (1990), and King et al. (1991).

By KLLN,

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

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 $\varepsilon_{Yi} = Y_i - g_Y(X_i)$ and $\varepsilon_{Zi} = Z_i - g_Z(X_i)$, $1 \leqslant i \leqslant n$. Since errors are independent of regressors, $(\varepsilon_{Yr_{in}} - \varepsilon_{Zr_{in}})$, $i \geqslant 1$, are i.i.d. with mean zero and variance σ^2, Donsker's theorem (see Billingsley, 1968) establishes that, under H_0,

$$\sup_{-\infty < t < \infty} \left| (n\sigma^2)^{-1/2} \sum_{i=1}^{n} D_i 1(X_i < t) \right| = \sup_{1 \leqslant k \leqslant n} \left| (n\sigma^2)^{-1/2} \sum_{i=1}^{k} (\varepsilon_{Yr_i} - \varepsilon_{Zr_i}) \right|$$
$$\xrightarrow{d} T = \sup_{0 \leqslant t \leqslant 1} |B(t)| \quad \text{as } n \to \infty,$$

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

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

(see Shorack and Wellner, 1986). Then $T_{0.1} = 1.96$, $T_{0.05} = 2.2414$, $T_{0.01} = 2.8074$ and $T_{0.001} = 3.4808$. Consider local alternatives

$$H_{1n}: \quad g_Y(x) - g_Z(x) = n^{-1/2} c\, |h(x)| \quad \text{for each } x \in [0, 1],$$

where $\Pr(X \in [0, 1]) = 1$, $h(\cdot)$ is a fixed function and c is a constant. Under H_{1n},

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

If $h(\cdot)$ and $f(\cdot)$ are continuous on (0, 1) and $f(\cdot)$ never vanishes,

$$\int_0^t |h(x)|\, f(x)\, dx = 0 \quad \text{for } 0 < t < 1 \quad \text{if and only if} \quad h(x) = 0 \quad \text{for all } x.$$

Then T_n diverges to ∞ as $|c| \to \infty$. Hence, the test is asymptotically powerful under alternatives converging to the null hypothesis at rate $n^{-1/2}$.

3. A Monte Carlo study

The first part of these simulations is based on the same design used by Hall and Hart (1990). The observations are generated according to the model

$$Y_i = g_Y(X_i) + \varepsilon_{Yi} \text{ and } Z_i = g_Z(X_i) + \varepsilon_{Zi}, \quad i = 1, \ldots, n. \tag{3.1}$$

Table 1
Proportion of rejections in 5000 replications in the first set of experiments ($X_i = i/n$) when $h(x) = g_Y(x) - g_Z(x)$

Error model	Alternative	α	n				
			15	20	30	50	100
(a)	(i)	0.05	0.0784	0.0680	0.0644	0.0554	0.0506
		0.01	0.0262	0.0210	0.0222	0.0168	0.0126
	(ii)	0.05	0.2922	0.3448	0.4422	0.6598	0.9242
		0.01	0.1586	0.1942	0.2668	0.4364	0.7972
	(iii)	0.05	0.7290	0.8514	0.9566	0.9974	1.0000
		0.01	0.5512	0.6894	0.8732	0.9878	0.9998
	(iv)	0.05	0.1538	0.1660	0.1928	0.2612	0.4698
		0.01	0.0660	0.0870	0.0706	0.1142	0.2536
	(v)	0.05	0.3678	0.4294	0.5468	0.7574	0.9640
		0.01	0.2016	0.2456	0.3340	0.5282	0.8782
(b)	(i)	0.05	0.0744	0.0704	0.0602	0.0574	0.0496
		0.01	0.0282	0.0228	0.0168	0.0130	0.0132
	(ii)	0.05	0.5944	0.7058	0.8612	0.9762	1.0000
		0.01	0.4044	0.5244	0.7080	0.9186	0.9992
	(iii)	0.05	0.9844	0.9982	1.0000	1.0000	1.0000
		0.01	0.9420	0.9870	1.0000	1.0000	1.0000
	(iv)	0.05	0.2798	0.3289	0.4232	0.6024	0.8716
		0.01	0.1428	0.1668	0.2344	0.3646	0.6996
	(v)	0.05	0.7124	0.8706	0.9300	0.9920	1.0000
		0.01	0.5120	0.6256	0.8048	0.9636	1.0000
(c)	(i)	0.05	0.0876	0.0762	0.0714	0.0570	0.0566
		0.01	0.0388	0.0302	0.0260	0.0162	0.0140
	(ii)	0.05	0.5998	0.7288	0.8894	0.9894	1.0000
		0.01	0.3696	0.4900	0.7212	0.9472	0.9998
	(iii)	0.05	0.9978	0.9998	1.0000	1.0000	1.0000
		0.01	0.9864	0.9988	1.0000	1.0000	1.0000
	(iv)	0.05	0.2460	0.2872	0.3968	0.5960	0.8932
		0.01	0.1072	0.1260	0.1936	0.3368	0.7186
	(v)	0.05	0.7466	0.8526	0.9576	0.9982	1.0000
		0.01	0.4988	0.7372	0.8412	0.9840	1.0000

Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two independent standard normal variables. The three choices for the distribution of the errors $(\varepsilon_{Yi}, \varepsilon_{Zi})$ were: (a) $(\mathcal{N}_1, \mathcal{N}_2)$, (b) $(|\mathcal{N}_1| - (2/\pi)^{1/2}, |\mathcal{N}_2| - (2/\pi)^{1/2})$, (c) $(|\mathcal{N}_1| - (2/\pi)^{1/2}, (2/\pi)^{1/2} - |\mathcal{N}_2|)$. All distributions have zero mean. In (a) and (b) the two error distributions are the same 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 $g_Y(x) - g_Z(x) = h(x)$, and (i) $h(x) = 0$, (ii) $h(x) = \frac{1}{2}$, (iii) $h(x) = 1$, (iv) $h(x) = \frac{1}{2}x$, and (v) $h(x) = x$.

Table 1 reports the proportion of rejections under (i)–(v) and errors distributions (a)–(c). The test performs well relative to the bootstrap test of Hall and Hart (1990) in the cases considered.

In a second set of experiments, observations are generated according to (3.1), but the design is random, $X_i \sim_{\text{i.i.d.}} \mathrm{N}(0, 1)$. Table 2 reports the proportion of rejections, in 5000 replications, under (i)–(v), and (vi) $h(x) = 2x$, 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 hypothesis 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 hypothesis than in the above set of

Table 2
Proportion of rejections in 5000 replications in the second set of experiments ($X_i \sim_{\text{i.i.d.}} N(0, 1)$) when $h(x) = g_Y(x) - g_Z(x)$

Error model	Alternative	α	n				
			15	20	30	50	100
(a)	(i)	0.05	0.0762	0.0616	0.0580	0.0502	0.0536
		0.01	0.0294	0.0214	0.0168	0.0142	0.0112
	(ii)	0.05	0.2778	0.3574	0.4766	0.6574	0.9242
		0.01	0.1496	0.1932	0.2810	0.4354	0.7954
	(iii)	0.05	0.7300	0.8598	0.9604	0.9974	1.0000
		0.01	0.5516	0.7038	0.8794	0.9876	1.0000
	(iv)	0.05	0.0882	0.0786	0.0952	0.1156	0.2230
		0.01	0.0310	0.0240	0.0272	0.0262	0.0592
	(v)	0.05	0.1074	0.1160	0.1758	0.2992	0.7176
		0.01	0.0376	0.0320	0.0488	0.0932	0.3490
	(vi)	0.05	0.1496	0.1912	0.3502	0.6632	0.9920
		0.01	0.0464	0.0574	0.1036	0.2764	0.8626

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 does not depend on the choice of a smoothing number. The Monte Carlo study results are encouraging.

The test can be implemented for testing the equality of several regression curves. Suppose we observe a random sample $\{(X_i, Y_i^{(1)}, \ldots, Y_i^{(p)}), i = 1, \ldots, n\}$ 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 $E(Y^{(k)} \mid X = x) = g_k(x)$, $k = 1, \ldots, p$.

We want to test the hypothesis

$$H_0: \quad g_k(x) = g_m(x) \quad \text{for all } m, k = 1, \ldots, p, \text{ and all } x \in \mathscr{X},$$

versus

$$H_1: \quad g_k(x) \neq g_m(x) \quad \text{for some } m \neq k, \ m, k = 1, \ldots, p, \text{ and some } x \in \mathscr{X}.$$

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

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

The test statistic is

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

Under the alternative hypothesis,

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

where

$$C^k(t) = \int_{-\infty}^{t} \left(g_k(x) - p^{-1} \sum_{j=1}^{p} g_j(x) \right) f(x)\, \mathrm{d}x,$$

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

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

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

Under H_{01}: $\max_k | E(g_Y(X) | X^k = x) - E(g_Z(X) | X^k = x)| = 0$ all x, the statistic

$$T_{nk} \xrightarrow{d} T \quad \text{as } n \to \infty.$$

Note that H_{01} is only a necessary condition for the equality of multiple regression curves. We may also try other functions of X, say $h : \mathbb{R}^p \to \mathbb{R}$, for testing $| E(g_Y(X) | h(X) = x) - E(g_Z(X) | h(X) = x)| = 0$ all x.

References

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

Barry, D. and J.A. Hartigan (1990), An omnibus test for departures from constant mean, *Ann. Statist.* **18**, 1340–1357.

Hall, P. and J.D. Hart (1990), Bootstrap test for difference between means in nonparametric regression, *J. Amer. Statist. Assoc.* **85**, 1039–1049.

Härdle, W. and J.S. Marron (1990), Semiparametric comparison of regression curves, *Ann. Statist.* **18**, 63–89.

Hong-zhi, A. and C. Bing (1991), A Kolmogorov–Smirnov type statistic with application to test for nonlinearity in time series, *Internat. Statist. Rev.* **59**, 287–307.

King, E.C. (1989), A test for the equality of two regression curves based on kernel smoothers, Ph.D. Dissertation, Dept. of Statist., Texas A&M Univ. (College Station, TX).

King, E.C., J.D. Hart and T.E. Wehrly (1991), Testing the equality of two regression curves using linear smoothers, *Statist. Probab. Lett.* **12**, 239–247.

Rice, J. (1984), Bandwidth choice for nonparametric regression, *Ann. Statist.* **12**, 1215–1230.

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