Journal of Statistial
Theary & Methods
1994, Vol. 5, pp.95 ~ 106

Goodness-of-Fit Tests in Regression via Nonparametric Function Techniques

Kim Jong Tae* and Moon Gyoung Ae**

ABSTRACT

A proposed test statistic is obtained by multiplying constant weights by the Neumann smooth type statistic discussed by Eubank and Hart(1993) in order to observe the effect of weight. It has very good results of power studies. Another advantage of this test is that it simultaneously provides an important diagnostic tools that can be used in many cases to determine how the model should be adjusted.

1. INTRODUCTION

Tests for no effect of a predictor x on the response have been considered by many authors : von Neumann(1941),von Neumann et al(1941), Munson and Jernigan(1989), Buckley(1991) and Eubank and Hart(1993) have dealt with testing for the hypothesis of no effect for a predictor in regression analysis.

* Department of Statistics, Pusan University of Foreign Studies, Pusan, Korea
** Department of Statistics, Kyungpook National University, Taegu, Korea

Assume that responses $y_1, y_2, \cdots, y_n$ are obtained from the model

$$y_i = \mu + f(x_i) + \epsilon_i \ , \quad i = 1, 2, \cdots, n \ , \tag{1.1}$$

at design points x_i of a predictor value x, where μ is an unknown constant, f is some unknown function satisfying $\int_0^1 f(t)dt = 0$ and the ϵ_i are independent and identically distributed($i.i.d.$) normal errors with zero mean and variance σ^2. Under this model we wish to test the hypothesis that x has no influence on the response, that is, we want to test for $H_0 \ : \ f = 0$.

Let $L_2[0,1]$ be the function space consisting of all functions f that satisfy $||f||^2 = \int_0^1 f^2(t)dt < \infty$ and $\int_0^1 f(t)dt = 0$. And let $\{\varphi_j\}_{j=1}^{\infty}$ be a complete orthonomal sequence(CONS) for $L_2[0,1]$ with norm $||\cdot||$. Assume that $f \in L_2[0,1]$, and define its Fourier coefficients corresponding to φ_j by $\beta_j = \int_0^1 f(x)\varphi_j(x)dx$. Then the unknown function f can be expressed as a Fourier series expansion

$$f(x) = \sum_{j=1}^{\infty} \beta_j \varphi_j(x) \ , \quad x \in [0,1].$$

One of strategies for finding $\widehat{f}$ relies on a sequence of numbers to optimize the estimator. Specifically, each coefficient, $\widehat{\beta}_{jn}$, $j = 1, 2, \cdots$, of the expansion $\widehat{f}$ is multiplied by a real number b_j(called multiplier) which decreases as j increases so that coefficients of $\widehat{f}$ can be gradually tapered to zero instead of being sharply truncated.

$$\widehat{f}(x) = \sum_{j=1}^{\infty} b_j \widehat{\beta}_{jn} \varphi_j(x_i) \ , \quad x \in [0,1] \ .$$

In this paper, we will propose the test based on a functional of each sample Fourier coefficient multiplied by a real number gradually tapered to zero in the expansion of $\widehat{f}$.

2. REVIEW OF OTHER TESTS

2.1. von Neumann Test

von Neumann(1941) suggested a test statistic based on the sum of

squares of first differences of the data to be reciprocal of

$$T_N = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{(n-1)\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2} \tag{2.1}$$

which rejects the null model for large values of (2.1). Recently, Munson and Jernigan(1989) have developed a modified version of the von Neumann (1941) ratio statistic, which is equivalent to the Durbin and Waston(1971) statistic when the regression is constant. Eubank and Hart(1993) showed that it is asymptotically equivalent to the von Neumann(1941) ratio statistic, which is given by

$$T_N = \frac{1}{\widehat{\sigma}^2}\sum_{j=1}^{n-1}\widehat{\alpha}_{jn}^2 \tag{2.2}$$

with $\widehat{\sigma}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(y_i - y_{i+1})^2$, where $\widehat{\sigma}^2$ is the consistent estimator of σ^2 proposed by Rice(1984) based on successive differences $y_{i+1} - y_i$.

2.2. Buckley Test

Buckley(1991) sought to detect any smooth variation in function f in viewpoint of Bayesian, not to specify a parametric alternatives, assuming that $f = h(n)g$ where $h(n)$ allows an arbitary departure from the null. With a particular prior distribution for $(g(x_1), g(x_2), \cdots, g(x_n))'$, a statistic for testing H_0 is proportional to

$$T_B = \frac{1}{n\widehat{\sigma}^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{j}(y_i - \overline{y})^2\right)^2 \tag{2.3}$$

with $\widehat{\sigma}^2$ given by (2.2), where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is given for testing H_0. This kind of statistics have been considered by Nair(1986) and Hirotsu(1986) against ordered alternatives. Eubank and Hart(1993) showed that T_B can be represented as follows :

$$T_B = \frac{1}{\widehat{\sigma}^2}\sum_{j=1}^{n-1}\frac{\widehat{\alpha}_{jn}^2}{\gamma_j} \tag{2.4}$$

with $\gamma_j = \left\{2n \sin(\frac{j\pi}{2n})\right\}^2$, according to the fact by Nair(1986).

2.3. Eubank and Hart Test

We have seen that T_B and T_N has some problems in the detection of alternatives with lower or higher frequency. In order to overcome difficulties arising when T_B and T_N are used as test statistics, Eubank and Hart(1993) have proposed some other tests. First, they considered a test derived by using a standard nonparametric estimator of g, $\widehat{g}_m(t) = \frac{\sqrt{2}}{\sqrt{n}} \sum_{j=1}^{m} \widehat{\alpha}_{jn} \cos(j\pi x)$, where $1 \leq m \leq n$ is an integer. Then

$$T_m = \frac{1}{\widehat{\sigma}^2} \sum_{j=1}^{m} \widehat{\alpha}_{jn}^2 , \tag{2.5}$$

which rejects the null hypothesis for large values of (2.5). Although, $T_m = T_N$ for $m = n-1$, m is often expected to be much smaller than n. The statistic T_m assigns weights 0 or 1 to the $\widehat{\alpha}_{jn}^2$ according to whether or not m is larger than j, while T_B downweights $\widehat{\alpha}_{jn}^2$ by $\frac{1}{\gamma_j}$.

3. THE PROPOSED TESTS

We now propose two new tests for testing $H_0 : f = 0$ under the model

$$y_i = \mu + h(n) g(\frac{2i-1}{2n}) + \epsilon_i , \quad i = 1, 2, \cdots, n, \tag{3.1}$$

where μ is an unknown parameter, g is some unknown function satisfying $\int_0^1 g(x)dx = 0$ and $h(n)$ is some function of the sample size that satisfies $h(n) \to 0$ as $n \to \infty$, and the ϵ_i are independent and identically distributed normal random errors with mean zero and variance σ^2. If the null hypothesis is true, the residuals $e_i = y_i - \overline{y}$ from fitting the null model (3.1) with $g = 0$ should have no pattern as a function of x. Thus, given some nonparametric regression fit $\widehat{g}$ to the residuals, we can base a test on quadratic functional form of g, for example, $\sum_{i=1}^{n} \widehat{g}^2(x_i)$. The derived test statistics comes from the above perspective.

For some integer $1 \leq m \leq n$, the Fourier cosine series estimator of g with m terms for residuals e_i can be given as follow :

$$\widehat{g}(x) = \sqrt{2}\sum_{j=1}^{m} b_j \widehat{a}_{jn} \cos(j\pi x), \tag{3.2}$$

with $b_j = 1 - \frac{j}{m+1}$, multiplier, and

$$\widehat{a}_{jn} = \frac{\sqrt{2}}{n}\sum_{i=1}^{n} y_i \cos(j\pi x_i)\ , \tag{3.3}$$

as an estimator of each Fourier coefficient $a_j = \sqrt{2}\int_0^1 g(x)\cos(j\pi x)dx$ in the expansion of $g(x)$. Under the assumption of model (3.1), define $a_{jn} = \frac{\sqrt{2}}{n}\sum_{i=1}^{n} g(x_i)\cos(j\pi x_i)$. Then the sample Fourier coefficient $\widehat{a}_{jn} = \frac{\sqrt{2}}{n}\sum_{i=1}^{n} y_i \cos(j\pi x_i)$ converges in distrbution to a normal random variable with mean $h(n)a_{jn}$ and variance $\frac{\sigma^2}{n}$.

Using the above estimator (3.2) of $g(x)$, we propose a new test statistic based on $\sum_{i=1}^{n}\widehat{g}^2(x_i)$ as follows:

$$Z_m = \frac{n\sum_{j=1}^{m} b_j^2 \widehat{a}_{jn}^2 - \widehat{\sigma}^2\sum_{j=1}^{m} b_j^2}{\widehat{\sigma}^2\left(2\sum_{i=1}^{m} b_j^4\right)^{1/2}}\ . \tag{3.4}$$

where $\widehat{\sigma}^2$ is any consistent estimator of σ^2.

To use this test statistic, Z_m, for the goodness-of-fit test, we first consider its asymptotic properties.

Theorem 3.1. Under the assumption of model (3.1), assume that $m \to \infty$, $n \to \infty$ in such a way that $\sup_{1\leq j\leq m} m^{1/2}\gamma(m,n) \to 0$, where $\gamma(m,n) = |a_{jn} - a_j|$. Then if $h(n) = m^{1/4}/\sqrt{n}$, $Z_m \xrightarrow{d} Z$, where Z is a normal random variable with unit variance and mean $\frac{\sqrt{5}||g||^2}{\sqrt{2}\sigma^2}$.

Proof. By using the following lemmas, we can find the asymptotic distribution of Z_m.

Lemma 3.1. For local alternative of the form (3.1),

(1) $|a_{jn} - a_j| = O(n^{-2})$ uniformly in $1 \le j \le m$
(2) Uniformly in j and q,

$$\mathrm{Cov}(\widehat{a}_{jn},\ \widehat{a}_{qn}) = \begin{cases} 0 & \text{if} \quad j \neq q \ , \\ \frac{\sigma^2}{n} & \text{if} \quad j = q \ . \end{cases}$$

Lemma 3.2. The matrix M_n for Z_m is $n \times n$ symmetric matrix that has i, k^{th} entry given by $\frac{2}{n}\sum_{j=1}^{m} {b_j}^2 \cos(j\pi x_i)\cos(j\pi x_k)$. Then,
(1) $tr(M_n) = \sum_{j=1}^{m} {b_j}^2$,
(2) $tr(M_n^2) = \sum_{j=1}^{m} {b_j}^4$,
(3) $f_n' M_n f_n = nh^2(n)\sum_{j=1}^{m} {b_j}^2 a_{jn}^2$,
(4) $f_n' M_n^2 f_n = nh^2(n)\sum_{j=1}^{m} {b_j}^4 a_{jn}^2$.

We use the Lindeberg-Feller theorem to prove that Z_m converges in distribution to normal. First, note that $\widehat{\sigma}^2$ is a consistent estimator of σ^2, $\sigma^2 - \widehat{\sigma}^2$ and $\frac{\sigma^2}{\widehat{\sigma}^2}$ converge to zero and unit as $n \to \infty$, respectively. Hence, by Slutsky's Theorem, Z_m has the same asymptotic distribution as

$$\frac{n\sum_{j=1}^{m} {b_j}^2 \widehat{a}_{jn}^2 - \sigma^2 \sum_{j=1}^{m} {b_j}^2}{\sigma^2 \left(2\sum_{j=1}^{m} {b_j}^4\right)^{1/2}} .$$

Now, observe that M_n is the matrix that has i, k^{th} entry as follows : $\frac{2}{n}\sum_{j=1}^{m} {b_j}^2 \cos(j\pi x_i)\cos(j\pi x_k)$. By the Lindeberg-Feller theorem, we note that

$$\begin{aligned} \frac{f_n' M_n^2 f_n}{tr(M_n^2)} &= \frac{nh^2(n)\sum_{j=1}^{m} {b_j}^4 a_{jn}^2}{\sum_{j=1}^{m} {b_j}^4} \\ &\le \frac{nh^2(n)(m\gamma(m,n)^2 + ||g||^2 + 2||g||m^{1/2}\gamma(m,n))}{\sum_{j=1}^{m} {b_j}^4} \\ &\longrightarrow 0 \end{aligned}$$

as $n \to \infty$, $m \to \infty$ and by Cauchy-Schwartz Inequality and the assumption. Finally, it remains only to show that

$$\frac{f_n' M_n f_n}{tr(M_n^2)^{1/2}} \to \sqrt{5}||g||^2 .$$

From Lemma 3.2, for $h(n) = m^{1/4}/\sqrt{n}$,

$$\frac{f_n' M_n f_n}{tr(M_n^2)^{1/2}} = \frac{\sqrt{m}\sum_{j=1}^{m} {b_j}^2 a_{jn}^2}{\left(\frac{m(6m^3+9m^2+m-1)}{30(m+1)^3}\right)^{1/2}} .$$

Thus, it is sufficient to show that $\sum_{j=1}^{m} {b_j}^2 a_{jn}^2 \to ||g||^2$, by the fact that as $m \to \infty$

$$\frac{\sqrt{m}}{\left(\frac{m(6m^3+9m^2+m-1)}{30(m+1)^3}\right)^{1/2}} \longrightarrow \sqrt{5} .$$

Then, $\sum_{j=1}^{m} {b_j}^2 a_{jn}^2 \leq \gamma^2(m,n)m + 2||g||m^{1/2}\gamma(m,n) + ||g||^2 \to ||g||^2$ since by Cauchy-Schwartz inequality and the assumption.

From the consequence of Theorem 3.1, we can see that Z_m is asymptotically standard normal under H_0 so that an approximate α-level test can be applied to the rejection of H_0 if Z_m exceeds the $100(1-\alpha)$ percentage point Z_α of the standard normal distribution. Since Z_m can only recognize alternatives converging at the rate $m^{1/4}/\sqrt{n}$ like T_m, it will not be capable of detecting alternatives that converges to the null as fast as the parametric rate $1/\sqrt{n}$. Therefore, it may seem that the price we have paid for using a test based on nonparametric estimators is the loss of the ability detecting alternatives that can be checked by parametric tests. But if we choose m to be of the orders $n^{1/3}$, then this test can detect alternatives that converges as fast as $n^{-5/12}$. Thus, if m is chosen correctly for analyzing the data obtained by the techniques such as generalized cross-validation, then this test may be able to detect alternatives converging at close to the parametric rate.

Corollary 3.1. Let $Z_{m\alpha}$ denote the upper α percentage points of Z_m. Assume that conditions of Theorem 3.1 and define $||g||^2 = \int_0^1 g^2(x)dx$. Then

$$\lim_{n\to\infty} P(Z_m \geq z_{m\alpha} \mid m^{1/4}g/\sqrt{n}) = 1 - \Phi\left(z_\alpha - \frac{\sqrt{5}||g||^2}{\sqrt{2}\sigma^2}\right) \tag{3.6}$$

where Φ is the standard normal distribution and $\Phi(Z_\alpha) = 1-\alpha$.

Corollary 3.1 gives another interesting consequences in that the asymptotic power of the Z_m test is monotone increasing function in the size of g, that is, $||g||^2$. This means that this test has the same asymptotic power against alternatives of the same size, and in particular, the power is not focused on any particular direction. This would seem to be a good property for such a test when no priori information is available of departures from the null. Specifically, against any alternative of size one, the test based on Z_m has the power

$$1 - \Phi\left(z_\alpha - \frac{\sqrt{5}}{\sqrt{2}\sigma^2}\right) > \alpha \ .$$

Each asymptotic power for α-percentile point of the standard normal distribution is listed in Table 3.1 with the assumptions of $||g|| = 1$ and $\sigma^2 = 1$ in Corollary 3.1.

Table 3.1 Asymptotic Power using The Standard Normal Distribution

Statistic	α 0.01	0.05	0.1
T_m	0.0521	0.1736	0.2810
Z_m	0.2296	0.4761	0.6179

As this result, we expect that our test statistic may have a good power for testing this null hypothesis. One of our interests for test is to see whether or not the test is consistent. The next theorem provides the consistency of the test.

Theorem 3.2 Define $\beta_{jn} = \frac{\sqrt{2}}{n}\sum_{i=1}^{n} f(x_i)\cos(j\pi x_i)$ and $\beta_j = \sqrt{2}\int_0^1 f(x)\cos(j\pi x)dx$. For any integer m assume that as $n,\ m \to \infty$, $m^{1/2}\sup_{1\le j\le m}|\beta_{jn} - \beta_j| \to 0$. Then for any fixed alternative f, the power of test based on Z_m tends to 1 if $\sqrt{m}/n \to 0$.

Proof. We will also assume that σ^2 is known as in Theorem 3.1. First note that $P(Z_m \ge z_{m\alpha}) = P(A_n + B_n + C_n \ge z_{m\alpha})$ with $\sigma_m = \sigma^2(2\sum_{j=1}^{m} b_j{}^4)^{1/2}$, $A_n = (n\sum_{j=1}^{m} b_j{}^2(\widehat{a}_{jn} - \beta_{jn})^2 - \sigma^2\sum_{j=1}^{m} b_j{}^2)/\sigma_m$, $B_n = 2n\sum_{j=1}^{m} b_j{}^2(\widehat{a}_{jn} - \beta_{jn})\beta_{jn}/\sigma_m$ and $C_n = n\sum_{j=1}^{m} b_j{}^2\beta_{jn}^2/\sigma_m$. A_n converges in distribution to a normal random variable with mean zero and unit

variance by the Lindeberg-Feller theorem. Now, $E(B_n) = 0$ and by Lemma 3.1,

$$Var(B_n) \leq O\left(\frac{n}{m}\right)(m\gamma^2(m,n) + 2||f||m^{1/2}\gamma(m,n) + ||f||^2).$$

Thus, Chebyshev's inequality shows B_n to $O_p\left(\left(\frac{n}{m}\right)^{1/2}\right)$. Since

$$C_n \leq \frac{1}{\sqrt{2}\sigma^2} O\left(\frac{n}{m}\right)(m\gamma^2(m,n) + 2||f||m^{1/2}\gamma(m,n) + ||f||^2)$$

by Cauchy-Schwartz inequality, it follows that $\frac{\sqrt{m}}{n} Z_m \xrightarrow{d} \frac{||f||^2}{\sqrt{2}\sigma^2}$, and theorem has been proved.

4. A MONTE CARLO SIMULATION

We examine the small sample properties of our test for fixed alternatives in order to investigate whether or not our asymptotic results can be extended in finite samples through the Monte Carlo simulation study in this section. The Monte Calro simulation was done by the samples of size equal to 20 and 40 which were generated from model (1.1) with the ϵ_i uncorrelated normal random errors. The variance of error σ^2 were assumed to be known and without loss of generality, to be equal to unit. And x_i were taken to be equally spaced.

For the function f in (1.1), we used

$$f_1(x) = b\left(e^{4x} - \frac{(e^4-1)}{4}\right)\left(\frac{e^8}{8} - \frac{(e^4-1)^2}{4}\right)^{-1/2}$$

and $f_2(x) = \gamma\cos(j\pi x)$, where b and γ determine the distance of alternatives from the null, while j is to be manipulated to obtain higher and lower frequency alternatives. We used $\gamma = 0.5,\ 1.0,\ 1.5$ and 2.0, and chose $j = 1,\ 3$ and 6. Specifically $b = \frac{i}{4},\ i = 1,\ 2,\ 3,\ 4$ was chosen.

For each above combination, first uniform pseudo-random numbers are generated by GGUBS in IMSL package. Using Box-Millur transformation method, we genereted the normal random errors for sapmles of each size.

For samples of each size, the proportions of times T_B, T_N, T_m, and Z_m on or above their approximate upper α-level critical value, 0.05, were recorded. In this case, the approximate upper α-level critical values were empirically found by simulation with the null distribution of these statistics. Because it is generally it is well known that the normal approximation does not alwayes work well. In doing this, we used 10000 trials to get more satisfactory critical values for these statistics in each case.

Table 4.1 Powers Against The Alternative f_1 when $\alpha = 0.05$

Sample Size	Type	b			
		0.25	0.50	0.75	1.00
20	T_N	0.074	0.168	0.435	0.750
	T_B	0.165	0.533	0.850	0.979
	T_m	0.168	0.426	0.738	0.933
	Z_m	0.168	0.533	0.838	0.979
40	T_N	0.091	0.271	0.772	0.958
	T_B	0.314	0.822	0.995	1.000
	T_m	0.270	0.741	0.978	1.000
	Z_m	0.306	0.823	0.995	1.000

In summary, simulation results are likely to support our asymptotic analysis for the most part. As we have seen in the examinatoin of Table 4.1 and Table 4.2, Z_m has excellent power against alternatives with lower frequency even though m is not so large, compared with the Buckley test. But Z_m may have some difficulties in detecting somewhat higher frequency alternatives. So, Z_m is to be preferred in situation where lower frequency alternatives to the null are likely to be considered. But provided some larger m, its power can be improved against alternatives with higher frequency.

Table 4.2 Powers Against The Alternative f_2 when $\alpha = 0.05$ with $j = 3$

Sample Size	Smoothing Parameter	Type	γ 0.5	1.0	1.5	2.0
$n = 20$		T_N	0.045	0.106	0.410	0.849
		T_B	0.063	0.123	0.365	0.835
	$m = 9$	T_m	0.148	0.555	0.930	1.000
		Z_m	0.168	0.637	0.964	1.000
$n = 30$		T_N	0.056	0.120	0.501	0.937
		T_B	0.072	0.187	0.659	0.981
	$m = 9$	T_m	0.240	0.778	0.989	1.000
		Z_m	0.260	0.834	0.993	1.000

REFERENCE

1 Buckley, M.J. (1991), Detecting a Smoothing Signal : Optimizing of Cusum Based Procedures, *Biometrika*, 78, 253-262.

2 Durbin, J. and Watson, G.S. (1971), Testing for Serial Correlation in least squares regresion III, *Biometrika*. 58, 1-19.

3 Eubank, R.L. and Hart, J.D. (1992), Testing Goodness-of Fit in Regression via Order Selection Criteria, *The Annals of Statistics*, 20, 1412-1425.

4 Eubank, R.L. and Hart, J.D. (1993), Commonality of Cusum, von Neumann and Smoothing-Based Goodness-of-Fit Tests, *Biometrika*, 80, 89-98.

5 Hirostu (1986), Cumulative Chi-Squared Statistic as a Tool for Testing Goodness-of-Fit, *Biometrika*, 73, 165-173.

6 Munson, P.J. and Jernigan, R.W. (1989), A Cubic Spline Extension of The Durbin-Watson Test, *Biometrika*, 76, 39-47.

7 Nair, V.N. (1986), On Testing Against Ordered Alternatives in Analysis of Variance Model, *Biometrika*, 73, 493-499.

8 von Neumann, J. (1941), Distrbution of the Ratio of the Mean Squared Successive Difference to the Variance, *The Annals of Mathematical Statistics*, 12, 367-395.

9 von Neumann, J., Kent, R.H., Bellinson, H.R. and Hart, B.I. (1941), The Mean Square Successive Difference, *Annals of the Institute of Statistical Mathematics*, 12, 153-162.