Journal of the Korean
Data & Information Science Society
1998, Vol. 9, No. 2, 357 ~ 363

Bootstrap Lack of Fit Test based on the Linear Smoothers

Daehak Kim [1]

Abstract

In this paper we propose a nonparametric lack of fit test based on the bootstrap method for testing the null parametric linear model by using linear smoothers. Most of existing nonparametric test statistics are based on the residuals. Our test is based on the centered bootstrap residuals. Power performance of proposed bootstrap lack of fit test is investigated via Monte carlo simulation.

Key Words and Phrases: nonparametric, smoothing parameter, kernel estimator, smoothing, lack of fit, linear smoother, bootstrap

1. Introduction

Let $Y_1, Y_2, \cdots, Y_n$ be observations on the unknown regression function $r(\cdot)$ with a model

$$Y_i = r(x_i) + \epsilon_i, \ i = 1, \cdots, n \tag{1}$$

where ϵ_i are independent random variables having unknown common distribution F with mean 0 and finite variance σ^2. Without loss of generality we assume the design points x_i's are fixed with $0 \leq x_1 \leq \cdots \leq x_n \leq 1$.

For our purposes the principal aim in analyzing the data $(x_1, Y_1), \cdots, (x_n, Y_n)$ is to learn about the relationship between x and Y as it is expressed through the regression function $r(\cdot)$. For a long time parametric lack of fit test has been used to test the postulated null model fits well or not. Eubank and Spiegelman(1990) pointed out the parametric tests are inconsistent against many other alternatives. In order to overcome this difficulties many nonparametric test have been suggested by Cox, et. al.(1988), Eubank and Spiegelman(1990), Azzalini et. al.(1989) and Hărdle and Mammen(1993).

[1]Associate Professor, Department of Statistical information, Catholic University of Taegu-Hyosung, Kyungsan, Kyungbuk, 712-701, Korea

We focus attention on the linear smoothers based on fixed smoothing parameters. By a linear smoother we mean one that is linear in either $Y_1, Y_2, \cdots, Y_n$ or a set of residuals $e_1, e_2, \cdots, e_n$. If applied to residuals, a linear smoother has the form

$$\hat{g}(x:h) = \sum_{i=1}^{n} w_i(x:h)Y_i, \tag{2}$$

where the weights $w_i(x:h)$ are constants that do not depend on the data $Y_1, Y_2, \cdots, Y_n$ or any unknown parameters and h denote the smoothing parameters. Kernel estimator, local polynomial, smoothing splines are all linear in the Y_i's as long as their smoothing parameters are fixed rather than data driven.

Our interests is in testing the null hypothesis that $r(\cdot)$ is in some parametric class of functions S_θ against the general alternative that $r(\cdot)$ is not in S_θ. The basic idea behind lack of fit test is that one computes a smoother and compares it with a curve that is expected under the null hypothesis. If the smoother differs sufficiently from the expected curve, then there is evidence that the null hypothesis is false. As argued in Härdle and Mammen(1993), however convergence to the asymptotic distribution is quite slow so that it is more appropriate not to use asymptotic critical values.

In this paper, we propose a nonparametric lack of fit test based on bootstrap method. A short review for lack of fit test is introduced in section 2. In section 3 we study the Monte carlo simulation of proposed test.

2. Bootstrap Lack of Fit Test

2.1 Lack of fit test

For parametric model S_θ, which would be either linear or nonlinear in the unknown parameters we wish to test the null hypothesis

$$H_0 : r(\cdot) \in S_\theta = \{r(\cdot : \theta) : \theta \in \Theta\}. \tag{3}$$

where Θ is some subset of p dimensional Euclidean space with p finite and for each $\theta \in \Theta$, $r(\cdot : h)$ is a function with domain [0,1]. Let $\hat{\theta}$ be consistent estimator of θ assuming that the null hypothesis is true. Define residuals $e_1, \cdots, e_n$ by

$$e_i = Y_i - r(x_i : \hat{\theta}), i = 1, \cdots, n \tag{4}$$

If the null hypothesis is true these residuals should behave more or less like a batch of zero mean, uncorrelated random variables. Hence when H_0 is true, a linear smooth $\hat{g}(x:h)$ in (2) will tend to be relatively flat and centered about 0. A useful diagnostic is to plot the estimate $\hat{g}(\cdot : h)$ and see how much it differs from the zero

function. Often a pattern will emerge in the smooth that was not evident in a plot of residuals.

An obvious way of testing H_0 is to use a test statistic of the form $T = \frac{\|\hat{g}(\cdot:h)\|^2}{\hat{\sigma}^2}$ where $\|g\|$ is a quantity that measures the size of the function g and $\hat{\sigma}^2$ is model free estimator of the error variance σ^2. Examples of $\|g\|$ are $\{\int_0^1 g^2(x)f(x)dx\}^{1/2}$, $\int_0^1 |g(x)|dx$, $\sup_{0\leq x\leq 1}|g(x)|$ where f is the design density. A convenient approximation to $\int_0^1 g^2(x)f(x)dx$ is $\frac{1}{n}\sum_{i=1}^n \hat{g}^2(x_i:h)$ which leads to the lack of fit statistic

$$R_n = \frac{n^{-1}\sum_{i=1}^n \hat{g}^2(x_i:h)}{\hat{\sigma}^2} \tag{5}$$

A sensible test would reject H_0 for large values of R_n.

2.2 Testing the Fit of Linear Model

We now consider nonparametric lack of fit test using linear smoothers to test the fit of a linear model, in which case the null hypothesis

$$H_0 : r(x) = \sum_{j=1}^{p} \theta_j r_j(x). \tag{6}$$

Define $\mathbf{e}$ to be the column vector of residuals and suppose that our test statistics is of the form (5). Then $\hat{\sigma}^2$ can be written as $\hat{\sigma}^2 = \mathbf{e}'C\mathbf{e}$ for some matrix C not depending on the data and $\hat{g}(x_i : h)$ is of the form (2). The vector of smoothed residuals is denoted by $\hat{\mathbf{g}}$ and is expressible as

$$\hat{\mathbf{g}} = W\mathbf{e} = W(I_n - R(R'R)^{-1}R')\mathrm{Y}, \tag{7}$$

where W is $n \times n$ matrix with ijth element $w_j(x_i)$ and R is $n \times p$ design matrix

$$R = \begin{pmatrix} r_1(x_1) & r_2(x_1) & \cdots & r_p(x_1) \\ r_1(x_2) & r_2(x_2) & \cdots & r_p(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ r_1(x_n) & r_2(x_n) & \cdots & r_p(x_n) \end{pmatrix}$$

and I_n is identity matrix. The statistic $R_n$has the form

$$R_n = \frac{n^{-1}\hat{\mathbf{g}}'\hat{\mathbf{g}}}{\mathbf{e}'C\mathbf{e}} \tag{8}$$

The distribution of R_n under the null hypothesis can be approximated by the ratios of certain quadratic forms.

Now we consider a particular linear model and a particular smoother for specific asymptotic analysis of R_n. Let $r_1(x)$ be known twice continuously differentiable function on [0,1] such that $\int_0^1 r_1(x)dx = 0$ and consider testing the null hypothesis

$$H_0 : r(x) = \theta_0 + \theta_1 r_1(x),\ 0 \leq x \leq 1. \tag{9}$$

We will test H_0 using statistics R_n of the form (5) based on the Priestly-Chao(1972) type kernel smoother.

For $x_i = (i - 1/2)/n, i = 1, 2, \cdots, n$ and let $e_1, e_2, \cdots, e_n$ be the residuals from the least squares fit $\hat{\theta}_0 + \hat{\theta}_1 r_1(x)$ and define the smoother

$$\hat{g}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} e_i K(\frac{x - x_i}{h}), \tag{10}$$

where the kernel K has support (-1,1). we consider the test statistic of the following form

$$R_{n,h} = \frac{n^{-1} \sum_{i=[nh]+1}^{n-[nh]} \hat{g}_h^2(x_i)}{\hat{\sigma}^2} \tag{11}$$

The sum in $R_{n,h}$ is restricted to avoid the complification of boundary effects as Rice(1984) pointed out. The variance estimator $\hat{\sigma}^2$ is any estimator that is consistent for σ^2 under H_0. Asymptotic normality of $R_{n,h}$ is shown by King(1988) under some appropriate conditions.

2.3 Bootstrap lack of fit test

Usually asymptotic results are used for some nonparametric test. By the generic property of nonparametric approach, it is impossible to get an exact critical value for a nonparametric test and therefore only asymptotic results are available by a asymptotic distribution of test statistics. As argued by Härdle and Mammen(1993), however the convergence to the asymptotic distribution is quite slow so that it is most appropriate not to use the asymptotic critical values. This is the motivation for the proposed bootstrap lack of fit test.

The bootstrap, recently developed statistical method, can be used to approximate the sampling distribution of some statistical quantity of interest. The approximate of sampling distribution, called bootstrap distribution is usually approximated by Monte Carlo method using the bootstrap samples. Bootstrap method would be particularly useful when the distribution of statistical quantity of interest is too complicated or can not be estimated directly.

In order to get the desired approximate distribution of $R_{n,h}$ based on bootstrap method, we consider the following procedures. Let $e_1^*, e_2^*, \cdots, e_n^*$ be i.i.d. samples from the estimated and centered residuals under the null hypothesis. With these residuals, we can construct the bootstrap estimate $R_{n,h}^*$ of $R_{n,h}$.

$$R_{n,h}^* = \frac{n^{-1} \sum_{i=[nh]+1}^{n-[nh]} \hat{g}_h^{*2}(x_i)}{\hat{\sigma}^2} \tag{12}$$

where

$$\hat{g}_h^*(x) = \frac{1}{nh} \sum_{i=1}^{n} e_i^* K(\frac{x - x_i}{h}), \tag{13}$$

is Priestly and Chao type kernel smoother based on resampled bootstrap residuals. ¿From the bootstrap estimate $R^*_{n,h}$ we can get the desired critical values of given significance level by approximating the sampling distribution of $R_{n,h}$ via Monte carlo simulation. The algorithm is as follows.

Bootstrap procedures for the lack of fit test

- **step 1**

  1. For given data $(x_1, Y_1), \cdots, (x_n, Y_n)$ get residuals $e_1, e_2, \cdots, e_n$ from the consistent estimator $\hat{r}(\cdot : h)$ under the null hypothesis
  2. Calculate $R_{n,h}$ with fixed bandwidth h

- **step2**

  1. From the residuals $e_1, e_2, \cdots, e_n$ get the centered residuals $\tilde{e}_1, \tilde{e}_2, \cdots, \tilde{e}_n$ by $\tilde{e}_i = e_i - \bar{e}$ where $\bar{e} = \sum_{i=1}^n e_i$
  2. Get the bootstrap residuals $e_1{}^*, e_2{}^*, \cdots, e_n{}^*$ by resampling $\tilde{e}_1, \tilde{e}_2, \cdots, \tilde{e}_n$
  3. Evaluate the $R^*_{n,h}$ with these $e_1{}^*, e_2{}^*, \cdots, e_n{}^*$

- **step 3**

  1. Repeat step 2 for B times and get the desired critical values c_α by the $100 \times (1-\alpha)\%$ percentile of the B values.
  2. Reject the null hypothesis if $R_{n,h}$ is greater or equal to c_α

3. Monte Carlo Simulation

For Monte carlo simulation we assume that the model (1) holds with $x_i = (i - .5)/n, i = 1, 2, \cdots, n$. In order to get the empirical power of proposed bootstrap lack of fit test we considered the following three different functions

$$\begin{aligned} r_1(x) &= \sin(2\pi x) \\ r_2(x) &= 20[(x/2)^2(1-x/2)^2 - 1/30] \\ r_3(x) &= 20[x^2(1-x)^2 - 1/30] \end{aligned}$$

The function shape is plotted in figure 1. Function $r_1(\cdot)$ and $r_3(\cdot)$ can be identified easily as nonlinear function by eyes. The sample sizes n were taken to be 50 and 100. Errors were generated from the standard normal, t distribution with 10 degrees of freedom and double exponential distribution respectively. We considered the linear

null hypothesis of the type $r(x) = x$ for the simple calculations. Also we used Epanechnikov kernel with initially chosen bandwidth $h = 0.1$. We allowed 1000 replications for each function and B=500 bootstrap replication was considered.

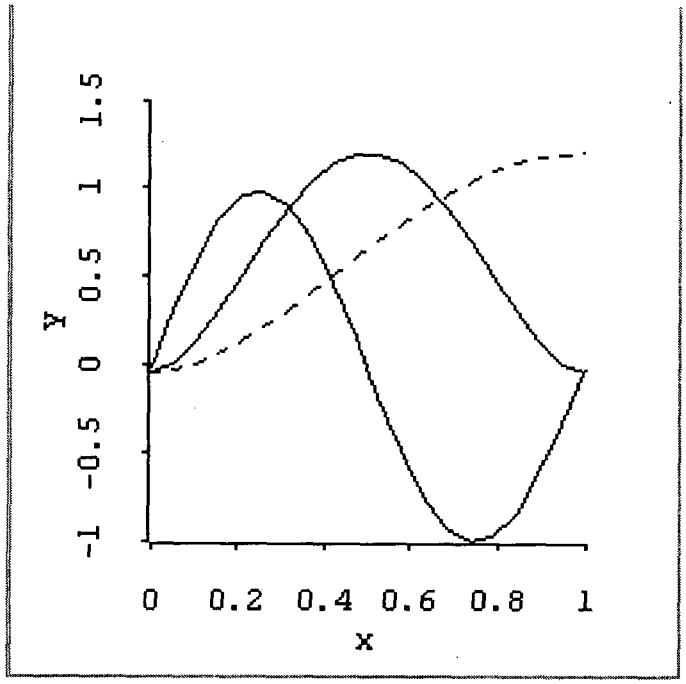

Figure 1. Function plot

Only 5% significance level was considered. All computation was carried out by Workstation SS-10. Random errors were generated from the IMSL fortran subroutines. The results are appeared in Table 1.

Table 1. Empirical power comparison

n	$r(x)$	$N(0,1)$	t_{10}	$d.e.$
50	$r_1(x)$	0.935	0.874	0.887
	$r_2(x)$	0.875	0.683	0.628
	$r_3(x)$	0.918	0.852	0.867
100	$r_1(x)$	0.975	0.926	0.936
	$r_2(x)$	0.845	0.734	0.742
	$r_3(x)$	0.972	0.916	0.935

4. Conclusion

Usually convergence of nonparametric lack of fit test to asymptotic distribution is too low so we couldn't use asymptotic critical values. With the results of simulation we are confident to use the proposed method not asymptotic ones. The empirical power of the $r_2(\cdot)$ looks lower than the other functions because $r_2(\cdot)$ is close to linear function. Bootstrap lack of fit test will also be useful to test the linearity of underlying function when the distribution of error is not normal or the sample size is not to much.

References

1. Azzalini, A., Bowman, A. W. and Härdle, W. (1989). On the use of nonparametric regression for model checking, *Biometrika*, 76, 1-11

2. Cox, D. D., Koh, E., Wahba, G., and Yandell, B. (1988). Testing the (parametric) Null Hypothesis in (Semiparametric) Partial and generalized spline models, *The Annals of Statistics*, 16, 113-119

3. Eubank, R. L., and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques, *Journal of the American Statistical Association*, 85. 387-392

4. Härdle, W. and Mammen, E. (1993). Comparing nonparametric regression versus parametric regression fits, *The Annals of Statistics*, 21, 1926-1947

5. King, E.C. (1988). *A test for the equality of two regression curves based on the kernel smoothers*, Ph.D. dissertation, Department of Statistics, Texas A&M University

6. Priestly, M. B. and Chao, M. T. (1972). Nonparametric function fitting. *Journal of Royal Statistical Society, Ser B,* 34, 385-392

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