*The Canadian Journal of Statistics*
Vol. 31, No. ?, 2003, Pages ???-???
*La revue canadienne de statistique*

# A Fully Nonparametric Diagnostic Test For Homogeneity of Variances

Lan WANG and Xiao-Hua ZHOU

*Key words and phrases:* Constant variance; health care costs; heteroscedastic errors; nearest-neighbour windows; nonparametric regression; residual.
*MSC 2000*: Primary 62G10; secondary 62G20, 62J02.

*Abstract:* For the regression model $Y_i = m(x_i) + \sigma(x_i)\epsilon_i$, $i = 1, \ldots, m$, we propose a new nonparametric diagnostic test for checking the constancy of the conditional variance function $\sigma^2(x)$. The proposed test does not assume a known parametric form for the conditional mean function $m(x)$. The test statistic is inspired by recent asymptotic theory in analysis of variance when the number of factor levels is large. Simulation studies demonstrate the good finite-sample properties of the test. We apply the proposed test to a study on the effect of drug utilization on health care costs.

**Title in French: we can supply this**

*Résumé :* For the regression model $Y_i = m(x_i) + \sigma(x_i)\epsilon_i$, $i = 1, \ldots, m$, we propose a new nonparametric diagnostic test for checking the constancy of the conditional variance function $\sigma^2(x)$. The proposed test does not assume a known parametric form for the conditional mean function $m(x)$. The test statistic is inspired by recent asymptotic theory in analysis of variance when the number of factor levels is large. Simulation studies demonstrate the test has good finite-sample properties. We apply the proposed test to a study on the effect of drug utilization on health care costs.

## 1. INTRODUCTION

Homoscedasticity or constant variance is a standard assumption in regression models. Ignoring heteroscedasticity can lead to inefficient estimation or incorrect inference (Chapter 14, Ruppert, Wand and Carroll, 2003). Heteroscedasticity in linear models can result in substantial inefficiency in ordinary least squares (OLS) estimation (Greens, 2000). In many applications, however, it is not uncommon to find the assumption of homoscedasticity violated. Examples include calibration experiments of the physical and biological sciences, radioimmunoassay, econometrics, pharmacokinetic modelling (Davidian and Carroll, 1987) and prospective payment modelling (Maciejewski, 2004). It is therefore important to be able to assess the adequacy of this assumption.

Graphical diagnostic procedures, such as plotting the residuals versus the fitted value (or the covariate), often provide useful visual aid (e.g., Cook and Weisberg, 1982, §2.3). Formal tests are needed to evaluate whether the patterns observed are due to random fluctuations.

Consider the following nonparametric regression model:

$$Y_{mi} = g(x_{mi}) + \sigma(x_{mi})\epsilon_{mi},\ i = 1, ..., m, \tag{1}$$

where $Y_{mi}$ is the response, $g(x)$ is an unknown regression function, $x_{m1}, \ldots, x_{mm}$ are fixed design points on [0,1], $\sigma^2(x)$ is the variance function, and $\epsilon_{mi}$'s form a triangular array of row-wise independent random variables with mean 0 and variance 1. For simplicity in notation, the sample size $m$ in the subscript will be omitted whenever feasible. Of interest is to test if this regression model is homoscedastic. The null hypothesis thus is:

$$H_0 : \sigma^2(x) = \sigma^2 \text{ for all } x, \tag{2}$$

for some unknown positive constant $\sigma^2$. We will allow for a general nonparametric alternative which only assumes $\sigma^2(x)$ to be any nonconstant smooth function. We will propose a fully nonparametric approach for testing (2). The test will not require direct estimation of $g(x)$, which can be any Lipschitz continuous function. It also allows flexible distributions for the $\epsilon_{mi}$'s.

When the functional form of $\sigma^2(x)$ is restricted to some parametric class, various parametric and semiparametric tests were proposed (Bickel, 1978, Cook and Weisberg, 1983, Davidian and Carroll, 1987, Carroll and Ruppert, 1988, among others). These tests generally assume a known parametric form for $g(x)$. Normality assumption is often imposed, too. For the general situation when $\sigma^2(x)$ belongs to an infinite dimensional space of smooth functions, Eubank and Thomas (1993) propose a test but their method assumes normality and requires the choice of some weight function. The approach of Müller and Zhao (1995) requires that the relation between $g(x)$ and $\sigma^2(x)$ follows a generalized linear model. Zheng (1996) provides a nonparametric Lagrange Multiplier (LM) test using kernel estimation of the score function. Diblasi and Bowman (1997) construct a test based on nonparametric smoothing of the residuals on a suitably transformed scale but they haven't derived the asymptotic distribution of the test, instead they approximate the critical value numerically. The method of Dette and Munk (1998) is based on an estimator for the best $L^2$ approximation of the variance function by a constant. When the regression function is estimated via wavelet thresholding methods and the error variances depend on the observed covariates through a parametric relationship of some known form, a score test is given by Cai, Hurvich and Tsai (1998).

In Section 2, we introduce the statistic for testing (2), present the main asymptotic results under the null hypothesis and local alternatives and discuss generalizations to other settings. Results from simulation studies to investigate the finite sample behaviour of the test statistic are reported in Section 3. In Section 4, we illustrate our method on a real data set to study health care cost. Technical details are given in an appendix.

## 2. TEST STATISTIC

### *2.1 Notations*

Let r(x) denote a Lipschitz continuous positive density function defined on [0,1]. Assume the design points $x_1, x_2, ..., x_m$ satisfy a standard assumption for fixed-design nonparametric regression model:

$$\int_0^{x_i} r(x)dx = \frac{i}{m}, \quad i = 1, ..., m. \tag{3}$$

Define $R_j^2 = \frac{1}{2}(Y_{j+1} - Y_j)^2$, $j = 1, ..., n$, where $n = m - 1$. If $\sigma^2(x)$ and $m(x)$ defined in (1) are Lipschitz continuous, we can easily show $E(R_j^2) = \sigma^2(x_j) + O(n^{-1})$. This provides a simple asymptotically unbiased local estimator for the variance function. The main advantage of this estimator is that it does not require to estimate the regression function $m(x)$.

Let $\widehat{\sigma}^2$ be an estimate of the variance under the null hypothesis, which can be taken as, for example, Rice's (1984) estimator $\widehat{\sigma}^2 = \frac{1}{2(m-1)} \sum_{j=1}^{m-1} (Y_{j+1} - Y_j)^2$, then

$$B_j = R_j^2 - \widehat{\sigma}^2, \quad j = 1, \ldots, n,$$

provide an asymptotically centred version of the sequence $R_j^2$.

### *2.2 ANOVA with large number of factor levels*

To introduce our test statistic, we first diverge to briefly review some related results in analysis of variance (ANOVA) when the number of cells becomes large as the sample size becomes large. We consider a balanced one-way ANOVA of $n$ cells and $k_n$ observations $V_{i1}, \ldots, V_{ik_n}$ in cell $i$, $i = 1, \ldots, n$. To test the null hypothesis of no cell effects, the following F-type statistic can be used:

$$F_n = \frac{MST}{MSE}$$

where

$$MST \;=\; \frac{k_n}{n-1}\sum_{i=1}^{n}(\overline{V}_{i\cdot} - \overline{V}_{\cdot\cdot})^2, \;\; MSE \;=\; \frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}(V_{ij} - \overline{V}_{i\cdot})^2, \tag{4}$$

$\overline{V}_{i\cdot} = k_n^{-1}\sum_{j=1}^{k_n} V_{ij}$, $\overline{V}_{\cdot\cdot} = (nk_n)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k_n} V_{ij}$, and $N = nk_n$ is the total number of observations. Classical large sample results for the F-test assume that the number of observations per cell goes to infinity but the number of cells is fixed. Recently, a different type of asymptotic framework has been studied: the number of cells goes to infinity, the number of observations per cell can either be fixed or goes to infinity at a slower rate, see Akritas and Papadatos (2004), Wang and Akritas (2003) and the reference therein. In this new framework, Akritas and Papadatos (2004) shows the asymptotic normality of $\sqrt{n}(F_n - 1)$ as $n \to \infty$ under the assumption of no cell effects. They assume $V_{ij}, i = 1, .., n, \; j = 1, ..., k_n$, are independent random variables, but allow them to be nonnormal and have heteroscedastic errors. Since $MSE$ converges to a constant, by Slutsky's theorem, the problem reduces to studying the asymptotic distribution of $\sqrt{n}(MST - MSE)$.

### *2.3 An artificial ANOVA*

We describe here, for the nonparametric regression model (1), how a hypothetical one-way layout is formed from $(x_j, B_j)$, $j = 1, \ldots, n$. In Section 2.4 below, we state a test statistic computed from this artificial ANOVA.

In the artificial ANOVA, a cell is constructed using the nearest neighbourhood method. More specifically, a symmetrized window $W_i$ is created around each covariate value $x_i, i = 1, \ldots, n$, by including the $k_n$ nearest covariate values. In what follows, the windows $W_i$ will also be understood as sets containing the indices $j$ of the covariate values that belong to the window, that is

$$W_i = \left\{ j : \; I\left( |\widehat{F}_x(x_j) - \widehat{F}_x(x_i)| \le \frac{k_n - 1}{2n} \right) \right\},$$

where $\widehat{F}_x(t) = n^{-1}\sum_{i=1}^{n} I(x_i \le t)$, and $I(A)$ is an indicator function for event $A$.

Using these nearest neighbourhood windows, we construct an artificial balanced one-way ANOVA with $n$ categories, the responses in the $i$-th category are those $B_i$-values associated with the covariate values in $W_i$. To separate the hypothetical observations from the original observations, let $V_{il}$, $l = 1, \ldots, k_n$, denote the $k_n$ responses in the $i$-th category of the above hypothetical ANOVA and let

$$\mathbf{V} = \mathbf{V}(\mathbf{X}, (B_j, j = 1, ..., n)') = (V_{11}, \ldots, V_{1k_n}, \ldots, V_{n1}, \ldots, V_{nk_n})' \tag{5}$$

be the vector of the $N \times 1$ vector of observations in the hypothetical one-way layout. We also use (5) to denote the operator which creates the above vector by augmenting $B_j$, $j = 1, ..., n$, according to the covariate $\mathbf{X} = (x_1, \ldots, x_n)'$.

### *2.4 The test statistic*

The new test we propose is based on the simple idea that under the assumption of homoscedasticity, $B_j$ should fluctuate around around zero. A test statistic for testing the the approximate constancy of $B_j$ can be naturally constructed by applying the test statistic in Section 2.2 to the hypothetical one-way layout formed in Section 2.3. Although $B_j$'s are not independent, we expect that the dependence is asymptotically negligible.

Consider the following test statistic

$$T = T(\mathbf{X}, (B_j, j = 1, ...n)') \equiv MST - MSE \tag{6}$$

where $MST$ and $MSE$ are defined in (4) and are calculated from the hypothetical ANOVA. $T(\mathbf{X}, (B_j, j = 1, ...n)')$ can be written as a quadratic form $\mathbf{V}'\mathbf{A}\mathbf{V}$, where

$$\mathbf{A} = \frac{nk_n - 1}{n(n-1)k_n(k_n - 1)} \bigoplus_{i=1}^{n} \mathbf{J}_{k_n} - \frac{1}{n(n-1)k_n}\mathbf{J}_{nk_n} - \frac{1}{n(k_n - 1)}\mathbf{I}_{nk_n}.$$

In the above expression, $\mathbf{I}_{k_n}$ is the $k_n$-dimensional identity matrix, $\mathbf{J}_{k_n} = \mathbf{1}_{k_n}\mathbf{1}'_{k_n}$ where $\mathbf{1}_{k_n}$ is the $k_n$-dimensional column vector of 1's, and $\bigoplus$ is the notation for Kronecker (direct) sum.

The test statistic (6) is related to the one used by Wang, Akritas and Van Keilegom (2002) for testing whether the regression function in (1) is constant. Wang, Akritas and Van Keilegom assume independent data. Here the sequence $R_j^2$, $j = 1, \ldots, n$, is 2-dependent. The asymptotic variance of $T$ is rather complicated. We consider the modified test statistic $T^*$,

$$\begin{aligned} T^* &= T^*(\mathbf{X}, (B_j, j = 1, ...n)') \\ &\equiv T(\mathbf{X}, (B_j, j = 1, ...n)') - \frac{2}{n(k_n - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} B_j B_{j+1} I(j, j+1 \in W_i) \\ &= \mathbf{V}'\mathbf{A}\mathbf{V} - \frac{2}{n(k_n - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} B_j B_{j+1} I(j, j+1 \in W_i). \end{aligned} \tag{7}$$

The term subtracted from $T$ is an estimate of the mean value of $T$ under (2), therefore $T^*$ is a centered version of $T$. In the next section, we will present large sample properties of $T^*(\mathbf{X}, (B_j, j = 1, ...n)')$ under both the null and local alternative hypotheses.

## 3. LARGE SAMPLE RESULTS

Assume the design points $x_1, \ldots, x_n$ satisfy (3). Let $E(\epsilon_i^k) = \mu_k(x_i)$, $k =$ 3, 4, 5, 6, be the higher moments of $\epsilon_i$. For some positive constant $L$ and for all $1 \le i, j \le n$, we have

$$\begin{aligned} &|g(x_i) - g(x_j)| \le L|x_i - x_j|, \\ &|\sigma^2(x_i) - \sigma^2(x_j)| \le L|x_i - x_j|, \\ &|\mu_k(x_i) - \mu_k(x_j)| \le L|x_i - x_j|, \quad k = 3, 4, 5. \end{aligned}$$

We start with some results on asymptotic equivalence under the null hypothesis (2). Lemma 1 below shows the test statistic $T^*$ calculated from the hypothetical one-way ANOVA augmented from $(B_j, j = 1, ..., n)'$ is asymptotically equivalent to the one calculated from the hypothetical one-way ANOVA augmented from $(Z_j, j = 1, ..., n)'$, where $Z_j = \sigma^2[\frac{1}{2}(\epsilon_{j+1} - \epsilon_j)^2 - 1]$. We let $\xrightarrow{p}$ denote convergence in probability, and $\xrightarrow{d}$ denote convergence in distribution.

Lemma 1. Under the above assumptions, if $k_n \to \infty$ and $k_n n^{-1} \to 0$, then under $H_0$,

$$\left(\frac{n}{k_n}\right)^{1/2} [T^*(\mathbf{X}, (B_j, j = 1, ..., n)') - T^*(\mathbf{X}, (Z_j, j = 1, ..., n)')] \xrightarrow{p} 0.$$

Lemma 2 below suggests that the above test statistic has the same asymptotical distribution when $\mathbf{A}$ in (7) is replaced by the simpler block diagonal matrix $\mathbf{A}_D$ below.

Lemma 2. Denote the $N \times 1$ vector $\mathbf{V}^* = \mathbf{V}(\mathbf{X}, (Z_j, j = 1, ...n)'))$, where $\mathbf{V}$ is the operator defined in (5), and assume the conditions of Lemma 1 hold. Then, under $H_0$,

$$\left(\frac{n}{k_n}\right)^{1/2} [T^*(\mathbf{X}, (Z_j, j = 1, ..., n)') - T^{**}(\mathbf{X}, (Z_j, j = 1, ..., n)')] \xrightarrow{p} 0,$$

where with the block diagonal matrix $\mathbf{A}_D = \mathrm{diag}\{\mathbf{B}_1, \ldots, \mathbf{B}_n\},\ \mathbf{B}_i = \frac{1}{n(k_n-1)}[\mathbf{J}_{k_n} - \mathbf{I}_{k_n}]$,

$$\begin{aligned} & T^{**}(\mathbf{X}, (Z_j, j = 1, ..., n)') \\ \equiv\ & \mathbf{V}^{*\prime}\mathbf{A}_D\mathbf{V}^* - \frac{2}{n(k_n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} Z_j Z_{j+1} I(j, j+1 \in W_i). \end{aligned}$$

As a direct result of Lemma 1 and Lemma 2, to derive the asymptotic distribution of $T^*$, we may work directly with $T^{**}(\mathbf{X}, (Z_j, j = 1, ..., n)')$. Direct calculation yields

$$T^{**}(\mathbf{X}, (Z_j, j = 1, ...n)') = \frac{1}{n(k_n-1)} \sum_{i=1}^{n} \sum_{|j_1-j_2|>1}^{n} Z_{j_1} Z_{j_2} I(j_1, j_2 \in W_i).$$

The next theorem presents the asymptotic distribution of $T^*$ under the null hypothesis.

Theorem 1. Assume the conditions of Lemma 1. Then, under $H_0$,

$$\left(\frac{n}{k_n}\right)^{1/2} T^*(\mathbf{X}, (B_j, j = 1, ...n)') \xrightarrow{d} N\left(0, \tau^2\right),$$

where $\tau^2 = \frac{4\sigma^8}{3} \int \mu_4^2(x) r(x) dx$.

*Remark 1.* The term $\int \mu_4^2(x) r(x) dx$ in the expression of $\tau^2$ can be consistently estimated by

$$\frac{1}{4(m-3)(\widehat{\sigma}^2)^4} \sum_{j=2}^{m-3} (Y_j - Y_{j-1})^4 (Y_{j+2} - Y_{j+1})^4 - 6m_4 - 9,$$

where $m_4 = \frac{1}{2(m-1)(\widehat{\sigma}^2)^2} \sum_{j=2}^{m} (Y_j - Y_{j-1})^4 - 3$.

*Remark 2.* The asymptotic normality of $T^*$ still holds if the nearest-neighborhood window size $k_n$ is taken to be fixed: $k_n = k$. In this case, $\tau^2$ has a more complex expression,

$$\begin{aligned} \tau^2 \ =\ & a_1(k)\sigma^8 \int \mu_4^2(x) r(x) dx + a_2(k)\sigma^8 \int \mu_4(x) r(x) dx + a_3(k)\sigma^8 \int \mu_3^2(x) r(x) dx \\ & + a_4(k)\sigma^8, \end{aligned}$$

where

$$\begin{aligned} a_1(k) &= \frac{8k^3 - 45k^2 + 79k - 39}{6k(k-1)^2}, & a_2(k) &= \frac{k^2 - 3k + 1}{k(k-1)^2}, \\ a_3(k) &= \frac{2(k-2)^2}{k(k-1)^2}, & a_4(k) &= \frac{k^2 - 3k + 3}{2k(k-1)^2}. \end{aligned}$$

*Remark 3.* It can be shown that the above asymptotic results still hold in a random design setting, where the covariates $x_i$, $i = 1, \ldots, n$, should be understood as a random sample from a distribution with a positive density function $r(x)$.

We next examine the power of our test against local alternatives of the form:

$$\sigma^2(x) = \sigma^2 + (nk_n)^{-1/4} h(x), \tag{8}$$

where $h(x)$ is a Lipschitz continuous nonconstant function. The asymptotic distribution is given in the following theorem.

THEOREM 2. Assume the conditions of Lemma 1, under the local alternative sequence (8), if $n^{-3}k_n^5 = o(1)$, we have

$$\left(\frac{n}{k_n}\right)^{1/2} T^*(\mathbf{X}, (B_j, j = 1, ..., n)') \xrightarrow{d} N\left(\gamma^2, \tau^2\right),$$

where $\tau^2$ is defined in Theorem 1 and

$$\gamma^2 = \int_0^1 h^2(t)r(t)dt - \left(\int_0^1 h(t)r(t)dt\right)^2.$$

*Remark 4.* By appropriately choosing the rate $k_n$, the test can detect local alternative converging to a null at a rate arbitrarily close to the parametric rate $n^{-1/2}$.

## 4. MONTE CARLO SIMULATIONS

In this section, we investigate the finite sample behaviors of the proposed test. The convergence of the test statistic to normal distribution is often quite slow (e.g., Härdle and Mammen, 1993). For finite sample sizes, the critical value of our test statistic is obtained using the bootstrap. More specifically, let $\widehat{\epsilon}_i = Y_i - \widehat{g}(x_i)$, where $\widehat{g}(\cdot)$ represents some estimate of the regression function, for example, by using least squares method or nonparametric smoothing. Let $\epsilon_i^*, i = 1, \ldots, m$, be a bootstrap sample from centered residuals $\widehat{\epsilon}_i, i = 1, \ldots, m$, and let $Y_i^* = \widehat{g}(x_i) + \epsilon_i^*$. For each bootstrap sample $(x_i, Y_i^*)$, $i = 1, \ldots, m$, calculate the test statistic $T^*$. The critical value is then determined from the appropriate quantile of the bootstrap distribution of the test statistic $T^*$. In the Monte Carlo study, we generate 500 simulated data sets for each scenario. For each data set 500 bootstrap samples are drawn. Different values of the sample size $n$, different forms of variance function and different distributions of the error term $\epsilon_i$ are allowed.

For comparison purpose, we include results from applying the parametric test of Cook and Weisberg (1983, abbreviated as CW test), which is a very powerful test when all the parametric assumptions required are satisfied. Their test requires correct specification of the variance function; here we adopt the assumption that $\sigma^2(x) = \sigma_0^2 exp(\lambda x)$ where $\sigma_0^2$ and $\lambda$ are unknown parameters, thus $\lambda = 0$ corresponds to the null hypothesis. In order to make fair comparisons, the regression function $g(\cdot)$ is taken to be linear $g(x) = 1 + x$ in our simulations. The design points are taken to be uniform on the interval [0,1]: $x_i = 0 : 1/(m-1) : 1$. The random data are generated by the software package Matlab 6.1. We also include results from applying the nonparametric test of Dette and Munk (1998, abbreviated as DM test), which is based on an estimator for the best $L^2$-approximation of the variance function by a constant.

First, we investigate level of the tests. Sample size 50 and 70 are considered. We are particularly interested in the situation when $\epsilon_i$'s are not normally distributed, which is assumed for the CW test (although it is possible to modify their test for other type of error distribution). Three different distributions for $\epsilon_i$ are considered: N(0,1), t(8) and t(4), where $N(\mu, \sigma^2)$ represents a normal distribution with mean 0 and variance $\sigma^2$, $t(b)$ represents a t-distribution with $b$ degrees of freedom; the smaller the b, the heavier is the tail of the distribution. The results are summarized in Table 1. In Table 1 and Table 2, our nonparametric test is calculated for several window sizes $(k_n)$. The test is abbreviated as NP$(k_n)$ test.

Table 1: Empirical level of the tests when $m(x) = 1 + x$.

| n | test | error distribution | | |
|---|---|---|---|---|
| | | N(0,1) | t(8) | t(4) |
| 50 | NP(3) | 0.040 | 0.056 | 0.050 |
| | NP(5) | 0.046 | 0.066 | 0.044 |
| | NP(7) | 0.046 | 0.064 | 0.036 |
| | NP(9) | 0.054 | 0.072 | 0.044 |
| | CW | 0.040 | 0.098 | 0.184 |
| | DM | 0.066 | 0.076 | 0.054 |
| 70 | NP(3) | 0.042 | 0.038 | 0.066 |
| | NP(5) | 0.046 | 0.044 | 0.076 |
| | NP(7) | 0.048 | 0.050 | 0.066 |
| | NP(9) | 0.046 | 0.054 | 0.060 |
| | NP(11) | 0.044 | 0.050 | 0.048 |
| | CW | 0.044 | 0.118 | 0.196 |
| | DM | 0.058 | 0.070 | 0.052 |

Table 2: Empirical power of the tests

| n | test | alternative 1 | alternative 2 | alternative 3 |
|---|---|---|---|---|
| 50 | NP(3) | 0.444 | 0.382 | 0.496 |
| | NP(5) | 0.516 | 0.474 | 0.530 |
| | NP(7) | 0.596 | 0.536 | 0.530 |
| | NP(9) | 0.628 | 0.590 | 0.490 |
| | CW | 1.000 | 0.160 | 0.450 |
| | DM | 0.414 | 0.348 | 0.366 |
| 70 | NP(7) | 0.668 | 0.642 | 0.692 |
| | NP(9) | 0.726 | 0.726 | 0.698 |
| | NP(11) | 0.752 | 0.770 | 0.672 |
| | NP(13) | 0.786 | 0.808 | 0.660 |
| | CW | 1.000 | 0.178 | 0.416 |
| | DM | 0.454 | 0.418 | 0.466 |

It is observe from Table 1 that the NP test maintains the specified nominal level very well, so is the DM test. The NP test is not very sensitive to the window size. The CW test, although behaves well when the distribution of $\epsilon(x_i)$ is normal or only slightly heavy tailed, could become very liberal when the tail is heavy, for example, if the true distribution of $\epsilon(x_i)$ is t(4).

To investigate the power of the tests, we consider the following three alternatives:

$$\begin{aligned}
\text{alternative 1:} \quad Y &= 1 + x + 0.5exp(2x)\epsilon; \\
\text{alternative 2:} \quad Y &= 1 + x + 0.5(1 + sin(10x)\epsilon); \\
\text{alternative 3:} \quad Y &= 1 + x + \sigma_3(x)\epsilon,
\end{aligned}$$

where $\epsilon$ has the standard normal distribution, $\sigma_3(x) = 0.5$ if $x < 0.5$, $\sigma_3(x) = 0.5(x - 0.5)^2$ otherwise. Table 2 summarizes the results. It is observed that if all the assumptions of the parametric test are satisfied (linear relationship in the mean function, normal error and correct specification of the form of the variance function), then the CW test is most powerful. This is the price the nonparametric test has to pay in order to be omnibus. However, when the assumptions of parametric test are violated, the nonparametric tests can be more powerful. This small-scale simulation study is certainly not exhaustive. It will be of interest to carry out more extensive simulations in the future.

From the power study, we observe that the influence of the local window size $k_n$ is small when the sample size is moderately large: $n = 70$. When $n$ is relatively small ($n = 50$), the finite sample power is more sensitive to $k_n$. Choosing a bandwidth to maximize the power of smoothing-based test is still an ongoing area of research. In general, this optimal bandwidth is different from the optimal one to estimate the nonparametric curve. Some discussions on this problem are given in §6.4 of Hart (1997). King, Hart and Wehrly (1991), Young and Bowman (1995) suggest calculating the $p$-value for several different choices of the smoothing parameter, and called the plot of $P$-values versus the smoothing parameters a “significant trace”. The modest goal of this paper is to provide a simple nonparametric diagnostic test that can be used to help determine if more sophisticated procedures, such as estimating the variance function, are needed.

## 5. DATA EXAMPLE

We illustrate the application of our bootstrap test on a clinical trial data set to investigate the effects of drug utilization review (DUR) on health care costs (Tierney et al., 1998). DUR involves comparing drug prescribing with accepted standards to identify potential problems. The response variable is the total health care cost of a patient. Health care costs are highly skewed due to high utilization of a few patients. To “normalize” the data, costs are often log-transformed prior to analysis (Manning, 1998; Zhou et al, 2001). However, a complication is that the transformation may normalize the skewed data but may not stabilize the variance (Manning, 1998). Hence, log-transformed data may also have a heteroscedastic variance. From preliminary analysis, we suspected that the heteroscedasticity may depend on the satisfaction level of a patient with his/her pharmacist. A scatter plot of log-transformed health cost versus patient satisfaction level is given in Figure 1. We apply our bootstrap test to this data set, which consists of 160 data points. The response variable is log transformed and the covariate is transformed to the interval [0,1]. The residuals are obtained using kernel smoothing with Epanechnikov kernel $K(x) = \frac{4}{3}(1-x^2)I(|x| \leq 1)$ and bandwidth $h$. The results of our test for different combinations of bandwidth $h$ and window size $k_n$ are summarized in Table 3, which indicates the possible presence of heteroscedasticity In contrast, the CW test gives a p-value of 0.356.

## 6. CONCLUSION

This paper proposed a fully nonparametric diagnostic test for testing the null hypothesis of homoscedasticity or constant variance. The test is motivated by recent development in analysis of variance with large number of factor levels. The test is asymptotically normal under the null hypothesis. It can detect local alternative converging to the null at a rate arbitrarily close to the parametric rate $n^{-1/2}$. The simulation results demonstrate that using critical value obtained from

Figure 1: Scatter Plot of the Health Care Cost Data

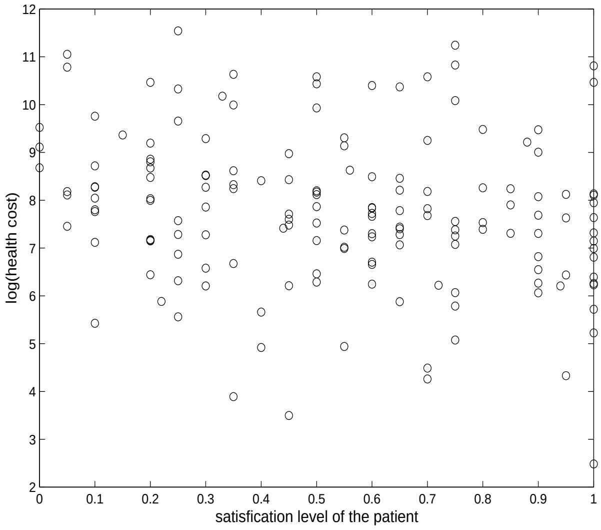

Table 3: Empirical power of the tests

| | $k_n = 3$ | $k_n = 5$ | $k_n = 7$ |
|---|---|---|---|
| h=0.1 | 0.034 | 0.082 | 0.186 |
| h=0.2 | 0.036 | 0.076 | 0.202 |
| h=0.3 | 0.018 | 0.084 | 0.188 |
| h=0.4 | 0.028 | 0.122 | 0.216 |

the bootstrap, the proposed test has satisfactory finite sample performance.

## APPENDIX: DERIVATIONS

We give here an outline of the proofs. More details are given in a technical report available from the authors.

*Proof of Lemma 1.* The proof is done by combining the following two steps:

$$\left(\frac{n}{k_n}\right)^{1/2} [T(\mathbf{X}, (B_j, j = 1, ..., n)') - T(\mathbf{X}, (Z_j, j = 1, ..., n)')] \xrightarrow{p} 0, \tag{9}$$

$$\left(\frac{n}{k_n}\right)^{1/2} \frac{2}{n(k_n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} (B_j B_{j+1} - Z_j Z_{j+1}) I(j, j+1 \in W_i) \xrightarrow{p} 0. \tag{10}$$

We will first prove (10). Notice that

$$B_j = Z_j + \frac{1}{2}(g(x_{j+1}) - g(x_j))^2 + \sigma(g(x_{j+1}) - g(x_j))(\epsilon_{j+1} - \epsilon_j) + (\sigma^2 - \widehat{\sigma}^2),$$

we can write the product $B_j B_{j+1}$ as the sum of 16 terms, one of them is $Z_j Z_{j+1}$. Therefore the left side of (10) can be decomposed into 15 terms:

$$\left(\frac{n}{k_n}\right)^{1/2} \frac{2}{n(k_n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} (B_j B_{j+1} - Z_j Z_{j+1}) I(j, j+1 \in W_i) = \sum_{t=1}^{15} D_t,$$

with

$$\begin{aligned} D_1 &= \left(\frac{n}{k_n}\right)^{1/2} \frac{1}{n(k_n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} Z_j (g(x_{j+2}) - g(x_{j+1}))^2 I(j, j+1 \in W_i), \\ D_2 &= \left(\frac{n}{k_n}\right)^{1/2} \frac{2}{n(k_n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} \sigma Z_j (g(x_{j+2}) - g(x_{j+1}))(\epsilon_{j+2} - \epsilon_{j+1}) I(j, j+1 \in W_i), \end{aligned}$$

$D_i$, $i = 3, \ldots, 15$, are defined similarly. One can easily check that each $D_i$, $i = 1, \ldots, 15$, is $o_p(1)$ by making use of the fact that $g(x_{j+1}) - g(x_j) = O_p(n^{-1})$ uniformly in $j$; for those terms involving random variables $Z_j$ and $\epsilon_j$, we can easily check the mean and variance. Now we check (9). Denote $\xi_j = g(x_{j+1}) - g(x_j)$ and let $\boldsymbol{\xi} = \mathbf{V}(\mathbf{X}, (\xi_j^2, j = 1, ...n)')$, $\boldsymbol{\eta} = \mathbf{V}(\mathbf{X}, (\xi_j(\epsilon_{j+1} - \epsilon_j), j = 1, ...n)')$ where $\mathbf{V}$ is the operator defined in (5), $\mathbf{V}^*$ is defined as before. Then

$$\begin{aligned} & T(\mathbf{X}, (B_j, j = 1, ...n)') - T(\mathbf{X}, (Z_j, j = 1, ...n)') \\ = \; & \frac{1}{4}\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi} + \sigma^2 \boldsymbol{\eta}'\mathbf{A}\boldsymbol{\eta} + (\sigma^2 - \widehat{\sigma}^2)^2 \mathbf{1}_N' \mathbf{A} \mathbf{1}_N + \mathbf{V}^{*\prime}\mathbf{A}\boldsymbol{\xi} + 2\sigma \mathbf{V}^{*\prime}\mathbf{A}\boldsymbol{\eta} \\ & + 2(\sigma^2 - \widehat{\sigma}^2)\mathbf{V}^{*\prime}\mathbf{A}\mathbf{1}_N + \sigma \boldsymbol{\xi}'\mathbf{A}\boldsymbol{\eta} + (\sigma^2 - \widehat{\sigma}^2)\boldsymbol{\xi}'\mathbf{A}\mathbf{1}_N + 2\sigma(\sigma^2 - \widehat{\sigma}^2)\boldsymbol{\eta}'\mathbf{A}\mathbf{1}_N. \end{aligned}$$

We can show that each of the above terms is $o_p(n^{-1/2}k_n^{1/2})$ as in the proof of Theorem 2.3 of Wang, Akritas and Van Keilegom (2002).

*Proof of Lemma 2.* We have

$$\begin{aligned} & \frac{n}{k_n} E(T^* - T^{**})^2 \\ = \; & \frac{n}{k_n} E\left[\mathbf{V}^{*\prime}(\mathbf{A} - \mathbf{A}_D)\mathbf{V}^*\right]^2 \\ = \; & \frac{n}{k_n} \frac{1}{n^2(n-1)^2 k_n^2} \sum_{i_1 \neq i_2} \sum_{i_3 \neq i_4} \sum_{j_1, j_2, j_3, j_4} E(Z_{j_1} Z_{j_2} Z_{j_3} Z_{j_4}) I(j_k \in W_{i_k}, k = 1, \ldots, 4). \end{aligned}$$

In the above sum, the nonzero expectation terms are of the following possible forms: $E(Z_{j_1}^2 Z_{j_2}^2)$ (of order $O(n^2k_n^4)$), $E(Z_j^4)$ (of order $O(nk_n^4)$), $E(Z_{j_1}Z_{j_1+1}Z_{j_2}Z_{j_2+1})$ (of order $O(n^2k_n^4)$), $E(Z_{j_1}^2 Z_{j_2}Z_{j_2+1})$ (of order $O(n^2k_n^4)$), etc. Thus

$$\frac{n}{k_n}E(T^* - T^{**})^2 = \frac{n}{k_n}\frac{1}{n^2(n-1)^2k_n^2}O(n^2k_n^4) = O(k_n n^{-1}) = o(1).$$

*Proof of Theorem 1.* From Lemma 1 and Lemma 2, we only need to show

$$n^{1/2}k_n^{-1/2}T^{**} \xrightarrow{d} N(0,\tau^2).$$

where $T^{**}(\mathbf{X},(Z_j, j=1,...,n)') = \frac{2}{n(k_n-1)}\sum_{i=1}^n \sum_{j_1+1<j_2}^n Z_{j_1}Z_{j_2}I(j_1,j_2 \in W_i)$. It's evident that $E(T^{**}) = 0$ and

$$\begin{aligned}
& Var\left(n^{1/2}k_n^{-1/2}T^{**}\right) \\
= & \frac{4}{n(k_n-1)^2k_n}\sum_{i_1=1}^n\sum_{i_2=1}^n\sum_{j_1+1<j_2}^n\sum_{l_1+1<l_2}^n E(Z_{j_1}Z_{j_2}Z_{l_1}Z_{l_2})I(j_1,j_2 \in W_{i_1}, l_1,l_2 \in W_{i_2}) \\
= & Q_1 + Q_2 + Q_3,
\end{aligned}$$

where $Q_1$ represents the sum of terms when the subscripts $j_1, j_2, l_1, l_2$ form two pairs, i.e., $j_1 = l_1$, $j_2 = l_2$; $Q_2$ is the sum of terms when there in one and only one pair among the subscripts $j_1, j_2, l_1, l_2$, i.e., $j_1 = l_1$, $j_2 \neq l_2$ or $j_1 \neq l_1$, $j_2 = l_2$; $Q_3$ is the sum of terms when there in no pair among the subscripts $j_1, j_2, l_1, l_2$.

$$\begin{aligned}
Q_1 &= \frac{4}{nk_n(k_n-1)^2}\sum_{i_1=1}^n\sum_{i_2=1}^n\sum_{j_1+1<j_2}^n E(Z_{j_1}^2Z_{j_2}^2)I(j_1,j_2 \in W_{i_1}\cap W_{i_2}) \\
&= \frac{\sigma^8}{nk_n(k_n-1)^2}\sum_{i_1=1}^n\sum_{i_2=1}^n\sum_{j_1+1<j_2}^n [\mu_4(x_{j_1})+1]^2 I(j_1,j_2 \in W_{i_1}\cap W_{i_2}) + O(n^{-1}k_n) \\
&= \frac{\sigma^8}{nk_n(k_n-1)^2}\sum_{j=1}^n [\mu_4(x_j)+1]^2[1^2+2^2+\cdots+(k_n-2)^2] + O(n^{-1}k_n) \\
&= \frac{\sigma^8(k_n-2)(2k_n-3)}{6nk_n(k_n-1)}\sum_{j=1}^n [\mu_4(x_j)+1]^2 + O(n^{-1}k_n) \\
&= \frac{\sigma^8(k_n-2)(2k_n-3)}{6k_n(k_n-1)}\int(\mu_4(x)+1)^2 r(x)dx + O(n^{-1}k_n),
\end{aligned}$$

where the second equality makes use of Lemma A.1 of Wang, Akritas and Van Keilegom (2002), the last equality uses (3). Similar calculation yield,

$$\begin{aligned}
Q_2 &= \frac{2\sigma^8(k_n-2)(k_n-3)}{3k_n(k_n-1)}\int(\mu_4^2(x)-1)r(x)dx + O(n^{-1}k_n), \\
Q_3 &= \frac{\sigma^8[(k_n-4)(k_n-3)(2k_n-1)+(k_n-3)(k_n-2)(2k_n-5)]}{12(k_n-1)^2k_n}\int(\mu_4(x)-1)^2 r(x)dx \\
&\quad + \frac{\sigma^8(k_n-2)^2}{2k_n(k_n-1)^2}\int[\mu_4^2(x)-2\mu_4(x)+4\mu_3^2(x)+1]r(x)dx + O(n^{-1}k_n).
\end{aligned}$$

Combining the above, we have $Var\left(n^{1/2}k_n^{-1/2}T^{**}\right) \to \tau^2$ as $k_n \to \infty$. $T^{**} = n^{-1}\sum_{i=1}^n A_i = n^{-1}S_n$, where $A_i = \frac{2}{k_n-1}\sum_{j+1\leq l}^n Z_jZ_lI(j,l \in W_i)$. The asymptotic normality of $(nk_n)^{-1/2}S_n$ can then be proved following the same lines as in the proof of Theorem 2.2 of Wang, Akritas and Van Keilegom (2002), using Markov's blocking technique (see proof of Theorem 27.4 in Billingsley, 1995).

*Proof of Theorem 2.* $B_j = R_j^2 - \widehat{\sigma}^2$, under the local alternative (8), we have $E(R_j^2) = \sigma^2 + (nk_n)^{-1/4}h(x_j) + O(n^{-1})$ and $E(\widehat{\sigma}^2) = \sigma^2 + (nk_n)^{-1/4}\int h(x)r(x)dx + O(n^{-1})$. Denote $B_j^* = B_j - (nk_n)^{-1/4}[h(x_j) - \int h(x)r(x)dx]$. As before, we let $V_{ij}$, $i = 1,\ldots,n$, $j = 1,\ldots,k_n$, be the observations in the hypothetical one-way ANOVA constructed from $B_j$. Further, let $U_{ij}$, $i = 1,\ldots,n$, $j = 1,\ldots,k_n$, denote the observations in the hypothetical one-way ANOVA constructed from $B_j^*$. $\mathbf{V}$, $\mathbf{U}$ denote the $N \times 1$ vector of observations in the two hypothetical one-way ANOVAs, respectively. Then $\mathbf{U} = \mathbf{V} - (nk_n)^{-1/4}\mathbf{S}$, where denoting $S_j = h(x_j) - \int h(x)r(x)dx$, $\mathbf{S}$ is the following $N \times 1$ vector $(S_j, j \in W_1, \ldots, S_j, j \in W_n)'$. The test statistic is calculated from $\mathbf{V}$ and will be called $T_{loc}^*$ here.

$$\begin{aligned}
T_{loc}^* &= \mathbf{V}'\mathbf{A}\mathbf{V} - \frac{2}{n(k_n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n-1} B_jB_{j+1}I(j,j+1 \in W_i) \\
&= \mathbf{U}'\mathbf{A}\mathbf{U} - \frac{2}{n(k_n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n-1} B_j^*B_{j+1}^*I(j,j+1 \in W_i) \\
&\quad +2(nk_n)^{-1/4}\mathbf{S}'\mathbf{A}\mathbf{U} + (nk_n)^{-1/2}\mathbf{S}'\mathbf{A}\mathbf{S} \\
&\quad -\frac{2}{n(k_n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n-1}(nk_n)^{-1/4}(S_jB_{j+1}^* + S_{j+1}B_j^*)I(j,j+1 \in W_i) \\
&\quad -\frac{2}{n(k_n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n-1}(nk_n)^{-1/2}S_jS_{j+1}I(j,j+1 \in W_i). \qquad (11)
\end{aligned}$$

Notice that

$$\begin{aligned}
B_j^* &= \sigma^2\left[\frac{1}{2}(\epsilon_{j+1}-\epsilon_j)^2 - 1\right] + (nk_n)^{-1/4}h(x_j)\left[\frac{1}{2}(\epsilon_{j+1}-\epsilon_j)^2 - 1\right] \\
&\quad +\frac{1}{2}(g(x_{j+1}) - g(x_j))^2 + \sigma(x_j)(g(x_{j+1}) - g(x_j))(\epsilon_{j+1}-\epsilon_j) \\
&\quad -\left(\widehat{\sigma}^2 - \sigma^2 - (nk_n)^{-1/4}\int h(x)r(x)dx\right) + O_p(n^{-1}). \qquad (12)
\end{aligned}$$

following the same lines as the proof of Theorem 1, we can show that

$$n^{1/2}k_n^{-1/2}\left[\mathbf{U}'\mathbf{A}\mathbf{U} - \frac{2}{n(k_n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n-1} B_j^*B_{j+1}^*I(j,j+1 \in W_i)\right] \to N(0,\tau^2).$$

Similarly as in the proof of Theorem 2.3 of Wang, Akritas and Van Keilegom (2002) and making use of expression (12), we can show that

$$\begin{aligned}
\mathbf{S}'\mathbf{A}\mathbf{U} &= O_p(k_nn^{-1/2}) + O_p(n^{-1}k_n^2), \\
\mathbf{S}'\mathbf{A}\mathbf{S} &= k_n\gamma^2 + O_p(k_nn^{-1}) + O_p(n^{-1}k_n^2),
\end{aligned}$$

where $\gamma^2 = \int_0^1 h^2(t)r(t)dt - \left(\int_0^1 h(t)r(t)dt\right)^2$. By checking mean and variance, we can show that the last two terms of (11) are both $o_p(n^{-1/2}k_n^{1/2})$.

## ACKNOWLEDGEMENTS

This research was funded in part by a grant from Agency for Healthcare Research and Quality. The data example was kindly provided by Dr. William Tierney from a study funded by Agency for Health Care Policy and Research. The authors are grateful for insightful comments from the editor, the associate editor and an anonymous referee.

## REFERENCES

M. G. Akritas & N. Papadatos (2004). Heteroscedastic one-way ANOVA and lack-of-fit test. *Journal of the American Statistical Association*, 99, 368-382.

P. Bickel (1978). Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, 6, 266–291.

P. Billingsley (1995). *Probability and Measure.* Third Edition. Wiley-Interscience: New York.

Z. W. Cai, C. M. Hurvich & C. L. Tsai (1998). Score Tests for Heteroscedasticity in Wavelet Regression. *Biometrika*, 85, 229–234.

R. Carroll & D. Ruppert (1988). *Transformation and weighting in regression.* New York: Chapman & Hall.

D. Cook & S. Weisberg (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70, 1–10.

D. Cook & S. Weisberg (1982). *Residuals and Influence in Regression.* Chapman & Hall, New York.

M. Davidian & R. Carroll (1987). Variance Function Estimation. *Journal of American Statistical Association*, 82, 1079–1091.

H. Dette & A. Munk (1998). Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society Series B*, 60, 693–708.

A. Diblasi & A. Bowman (1997). Testing for a constant variance in a linear model. *Statistics & Probability Letters,* **33,** 95–103.

R. L. Eubank & W. Thomas (1993). Detecting heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society Series B*, 55, 145–155.

W. H. Greens (2000). *Econometric Analysis*, 4th Edition, Prentice Hall, Upper Saddle River, New Jersey.

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

J. D. Hart (1997). *Nonparametric regression and lack-of-fit tests*, Springer, New York.

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

M. L. Maciejewski, X. H. Zhou, J. C. Fortney & J. F. Burgess (2004). Alternative Methods for Modelling Heteroscedastic Non-normally Distributed Costs. *Health Economics.* Submitted.

W. G. Manning (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, 17 283-295.

H. G. Müller & P. L. Zhao (1995). On a semiparametric variance model and a test for heteroscedasticity. *The Annals of Statistics*, 23, 946–967.

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

D. Ruppert  M. P. Wand & R. J. Carroll (2003). Semiparametric regression. Cambridge University Press.

W. M. Tierney, M. Overhage, M. Murray, X. H. Zhou, L. Harris& F. Wolinsky (1998). The final report of the computer-based prospective drug utilization review project (1993-1997) *U. S. Agency for Health Care Policy and Research.* Bethesda, MD.

L. Wang, M. G. Akritas & I. Van Keilegom (2002). Nonparametric Goodness-of-fit test for heteroscedastic regression models. *Submitted*

S. G. Young & A. W Bowman (1995). Non-parametric analysis of covariance. *Biometrics*, 51, 920-931.

J. X. Zheng (1996). A consistent nonparametric test of heteroscedasticity. *Preprint,* Department of Economics, University of Texas, Austin.

X. H. Zhou, K. T. Stroupe & W. H. Tierney (2001). Regression analysis of health care charges with heteroscedasticity. *Journal of the Royal Statistical Society Series C*, 50(3), 303-312.

*Received ???*
*Accepted ???*

Lan WANG: lan@stat.umn.edu
*School of Statistics, University of Minnesota*
*Minneapolis, MN 55455, USA*

Xiao-Hua ZHOU: azhou@u.washington.edu
*Department of Biostatistics, School of Public Health & Com Med,*
*University of Washington*
*Seattle, WA 98198, USA*