Assessing the Adequacy of Variance Function in Heteroscedastic Regression Models

Lan Wang*

School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455, U.S.A.

and

Xiao-Hua Zhou†

HSR&D, VA Puget Sound Health Care System, 1100 Olive Way, 1400, Seattle, WA 98101, U.S.A.

and

Department of Biostatistics, University of Washington, F600, HSB, Box # 357232, Seattle, WA 98198, U.S.A.

SUMMARY. Heteroscedastic data arise in many applications. In heteroscedastic regression analysis, the variance is often modeled as a parametric function of the covariates or the regression mean. We propose a kernel-smoothing type nonparametric test for checking the adequacy of a given parametric variance structure. The test does not need to specify a parametric distribution for the random errors. It is shown that the test statistic has an asymptotical normal distribution under the null hypothesis and is powerful against a large

* *email:* lan@stat.umn.edu
† *email:* azhou@u.washington.edu

class of alternatives. We suggest a simple bootstrap algorithm to approximate the distribution of the test statistic in finite sample size. Numerical simulations demonstrate the satisfactory performance of the proposed test. We also illustrate the application by the analysis of a radioimmunoassay data set.

KEY WORDS: goodness-of-fit test, heteroscedastic errors, kernel smoothing, pseudo-likelihood, variance function

1. Introduction

The problem of modeling heteroscedasticity frequently appears in data analysis. It is well known that modeling variance function is essential for efficient estimation of the regression mean and is sometimes of independent interest to the researchers. Moreover, whether the heteroscedasticity is appropriately taken into account can influence the estimation of other important quantities, such as confidence intervals, prediction intervals, and test statistics for regression coefficients. For instance, the quality of estimation in the analysis of assay data has been found to highly depend on the variance structure (Davidian, Carroll and Smith, 1988). See also Ruppert et al. (1997), Zhou, Stroupe and Tierney (2001) for examples in other fields.

In Section 5 of this paper, we analyze a data set from Carroll and Ruppert (1982, Section 2.8), which consists of 108 measurements from a calibration experiment of an assay for estimating the concentration of an enzyme esterase. The response variable Y is the radioimmunoassay (RIA) counts, and the covariate x is the concentration of esterase. A scatter plot of this data set is given in the top panel of Figure 1.

[Figure 1 about here.]

The heteroscedasticity exhibited in this data set is severe. Larger variance is associated with larger response. This naturally suggests the researchers to consider the variance as a function of the mean, for example, a power-of-the-mean variance function. But would this provide an adequate fit to the data? To answer this question, we need to develop an effective goodness-of-fit test for checking the adequacy of the proposed variance function.

Rigorous procedures for testing the goodness-of-fit of a given variance function are very lacking. Although many tests have been proposed for checking whether a variance function is constant or not, such as Breusch and Pagan (1980), White (1980), Cook and Weisberg (1983), Müller and Zhao (1995), Diblasi and Bowman (1997), Cai, Hurvich and Tsai (1998), these tests do not tell whether a specific variance function adequately describes the variability in the data. Classical tests, such as the Wald test, the likelihood ratio test and the score test, may be constructed for this purpose but they require the specification of a specific alternative model. Although the classical tests are powerful against the specified alternative, they may completely lose the power if the true alternative is not in the specified direction, see the numerical results in Section 4. Recently, Bedrick (2000) and Arbogast and Bedrick (2004) proposed new procedures to check the adequacy of the variance function in a log-linear model. Their methods allow for a large class of smooth alternatives, but they assume normal random errors and have not investigated general heteroscedastic regression models.

In this paper, we propose a kernel-smoothing type nonparametric test for assessing the goodness-of-fit of a variance function in a general heteroscedastic regression model. The proposed method does not require to specify a parametric distribution for the random errors and is designed to be powerful against different alternatives. It generalizes the smoothing test of Zheng (1996) for checking the lack-of-fit of the mean function. In Section 2 we introduce the test statistic and present its asymptotic properties. We discuss in Section 3 a simple bootstrap algorithm to obtain the critical values. We report numerical simulations in Section 4 and analyze the Esterase data in

Section 5. In Section 6 we generalize the test to the unknown mean function case. Section 7 summarizes the paper. The proofs are given in an appendix.

2. The Testing Procedure

2.1 *Hypothesis of Interest*

Let Y be a response variable, X be an $l \times 1$ vector of covariates and Z be a $q \times 1$ vector of explanatory variables which may contain part or all components of X. A general heteroscedastic regression model based on independent triplets of observations (X_i, Y_i, Z_i), $i = 1, \ldots, n$, can be written as

$$Y_i = f(X_i, \beta) + \epsilon_i, \quad \sigma_i^2 = g(Z_i, \beta, \theta), \tag{1}$$

where f is the conditional mean function, σ_i^2 denotes the conditional variance function $Var(Y_i|Z_i)$, the function g depends on the parameters β and θ where the components in θ are distinct from those in β, and the ϵ_i's are independent random errors with mean zero. This formulation includes the popular log-linear model and the power-of-the-mean model, where the former model corresponds to $f(X_i, \beta) = X_i'\beta$ and $g(Z_i, \beta, \theta) = exp(Z_i'\beta)$, and the latter model has $g(Z_i, \beta, \theta) = \theta_1(f(X_i, \beta))^{\theta_2}$.

We are interested in testing whether the variance function in (1) adequately describes the variability in the data. The null hypothesis is

$$H_0 : \quad \sigma_i^2 = g(Z_i, \beta, \theta), \quad \text{for some } \beta,\ \theta.$$

For example, to check the fit of a log-linear structure for the variance function, H_0 would state that g is an exponential function. The alternative space consists of all twice continuously differentiable functions other than exponential function.

For the transparency of explaining the main ideas, we assume that the mean function f has a known parametric form in the main body of the paper (relaxation of this assumption is discussed in Section 6). Knowledge of the mean function may come from our understanding of the random mechanism which generates the data, the underlying scientific theory or results from previous or similar studies. We suggest to carry out a goodness-of-fit test for the mean function (the modern smoothing test allows for testing the fit of the mean function without a parametric form for the variance function, see for example Zheng, 1996) at the first stage and to proceed with a test for the adequacy of the variance function only when the first test does not yield a significant result. In other words, attentions should be first given to the lower-order moment model and then to the higher-order moment model.

2.2 *The Test Statistic*

The test is motivated by the fact $E[r_i E(r_i|Z_i)p(Z_i)] = E[(E(r_i|Z_i))^2 p(Z_i)]$ is zero under H_0 but is strictly positive for any alternative, where $r_i = \epsilon_i^2 - g(Z_i, \beta, \theta)$, and $p(\cdot)$ is the density function of Z_i.

The test statistic is constructed as an estimator of $E[r_i E(r_i|Z_i)p(Z_i)]$. First, consider only the outer-layer expectation and replace it by the sample mean to obtain $n^{-1}\sum_{i=1}^{n} r_i E(r_i|Z_i)p(Z_i)$. Then, we estimate the product $E(r_i|Z_i)p(Z_i)$ nonparametrically by

$$\frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{h^q} K\left(\frac{Z_i - Z_j}{h}\right) r_j,$$

where $K(\cdot)$ is a kernel function, h is a smoothing parameter which depends on n and converges to 0 at an appropriate rate, and q represents the dimension of Z_i. It is often assumed that $K(u)$ is a nonnegative, bounded, continuous,

symmetric function and $\int K(u)du = 1$. This estimator is called a "leave-one-out" kernel estimator because the i-th observation is left out. Since the r_i's are not observable, they are replaced by:

$$\widehat{r}_i = (Y_i - f(X_i, \widehat{\beta}))^2 - g(Z_i, \widehat{\beta}, \widehat{\theta}), \quad i = 1, \ldots, n, \tag{2}$$

where $(\widehat{\beta}, \widehat{\theta})$ is any $\sqrt{n}$-consistent estimator of (β, θ), for example, the pseudo-likelihood estimator discussed in Section 3.1. The $\widehat{r}_i$'s are correlated due to the unknown estimation of the parameters but we expect them to approximately fluctuate around zero under H_0. A scatter plot of $\widehat{r}_i$ versus Z_i (of course, if Z_i is univariate) would be a useful graphical display to check the validity of the postulated variance structure.

Assembling the above estimators together, we obtain a nonparametric estimator of $E[r_i E(r_i|Z_i)p(Z_i)]$, which is given by

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h^q} K\left(\frac{Z_i - Z_j}{h}\right) \widehat{r}_i \widehat{r}_j. \tag{3}$$

Since a large value of T_n indicates deviations from the null hypothesis, T_n will be used as our test statistic. The statistic T_n is a smoothing-based nonparametric estimator of a population moment condition which is zero if and only if the null hypothesis is true, the test therefore belongs to the class of so-called "moments tests" which includes many popular testing procedures as special cases, such as the Lagrange multiplier test and the information matrix test. In particular, our test is a generalization of a test proposed by Zheng (1996) for testing the goodness-of-fit of the mean regression function.

Under the null hypothesis, T_n can be approximated by

$$T_n' = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h^q} K\left(\frac{Z_i - Z_j}{h}\right) r_i r_j. \tag{4}$$

Note that $T_n^{'}$ has the same form as T_n except that the $\widehat{r}_i$'s are replaced by the independent random variables r_i's. In fact, if $h \to 0$ and $nh^q \to \infty$ as $n \to \infty$, then under certain smoothness and moment conditions similarly as those in Zheng (1996),

$$nh^{q/2}(T_n - T_n^{'}) \to 0, \tag{5}$$

in probability under H_0. The statistic T_n' has the form of a degenerate second-order U-statistic and the theory developed in Hall (1984) can be applied to derive its asymptotic normality. Under H_0, we can show that as $n \to \infty$, $h \to 0$ and $nh^q \to \infty$,

$$nh^{q/2}T_n^{'} \to N(0, \tau^2) \tag{6}$$

in distribution, where $N(a, b)$ denotes the normal distribution with mean a and variance b and

$$\tau^2 = 2\int K^2(u)du \int [\xi^4(z, \beta, \theta) - g^2(z, \beta, \theta)]^2 p^2(z)dz, \tag{7}$$

with $\xi^4(z, \beta, \theta) = E(\epsilon_i^4 | Z_i = z)$. Because of (5), the asymptotic normal distribution given in (6) is also the limiting distribution of $nh^{q/2}T_n$. Thus to test for the adequacy of a given variance structure, a level α asymptotic test rejects the null hypothesis if $nh^{q/2}T_n/\tau > \Phi^{-1}(1-\alpha)$, where $\Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the standard normal distribution.

2.3 *Asymptotic Power Properties*

The nonparametric test T_n has the property of being consistent for any alternative that is twice continuously differentiable. This omnibus property can be established by showing that for any such alternative $nh^{q/2}T_n \to \infty$

in probability as $n \to \infty$. It is worth pointing out that classical parametric tests are only consistent against certain specific alternatives.

Furthermore, the power property is often analyzed for a sequence of local alternatives of the form $\sigma_i^2 = g(Z_i, \beta, \theta) + c_n \Delta(Z_i)$, where c_n is a sequence of numbers converging to zero, $\Delta(Z_i)$ is a function that does not belong to the parametric class under the null hypothesis. Of interest is the rate of c_n which makes the test have a nontrivial power between zero and one. For parametric tests, the rate in general is $n^{-1/2}$; for smoothing-based nonparametric tests, this rate is usually slower than $n^{-1/2}$. We can show that for $c_n = O(n^{-1/2}h^{-q/4})$, $nh^{q/2}T_n$ has an asymptotic normal distribution with a nonzero mean and the same asymptotic variance τ^2 as given in (7). Note that this rate can be made as close to the parametric rate $n^{-1/2}$ as possible if we let h converge to zero slowly.

3. Practical Implementation

3.1 *Pseudo-likelihood Estimation*

The implementation of the test requires estimation of the regression model under the null hypothesis. The book of Carroll and Ruppert (1988) provides a comprehensive review of several common methods for fitting heteroscedastic regression models, among which the pseudo-likelihood method has especially been proven to be simple and effective.

Briefly speaking, the pseudo-likelihood procedure involves iterative steps. Given β^*, a current estimator of β, the estimator of θ is defined to be the value which maximizes

$$-\sum_{i=1}^{n} ln(g(Z_i, \beta^*, \theta)) - \frac{1}{2}\sum_{i=1}^{n} \frac{(Y_i - f(X_i, \beta^*))^2}{g(Z_i, \beta^*, \theta)}. \tag{8}$$

Although (8) has the form of a normal likelihood, the pseudo-likelihood

makes no assumption about the distribution of the underlying data. Call the pseudo-likelihood estimator of θ obtained at this step θ^*, the estimator of β is then updated using the generalized least squares method, which is equivalent to solving the equation

$$\sum_{i=1}^{n} \frac{\partial f(X_i, \beta)}{\partial \beta} \frac{Y_i - f(X_i, \beta)}{g(Z_i, \beta, \theta^*)} = 0. \tag{9}$$

Given a starting value of β, the above process can be repeated until convergence. The pseudo-likelihood estimator is $\sqrt{n}$-consistent and asymptotically normal under very general conditions.

3.2 *A Bootstrap Algorithm*

It is well known that for nonparametric smoothing tests, the bootstrap method often provides more accurate approximation to the distribution of the test statistic than the asymptotic normal theory does when the sample size is small or moderate, see for example Härdle and Mammen (1993). We suggest below a simple bootstrap algorithm for the fixed design case. The same algorithm can be slightly modified and applied to the random design as well. The bootstrap algorithm consists of the following five steps:

1. For a given random sample of observations, obtain the quasi-likelihood estimator $(\widehat{\beta}, \widehat{\theta})$ of (β, θ) under the null hypothesis.

2. Define $\widehat{\epsilon}_i = [Y_i - f(X_i, \widehat{\beta})]/\sqrt{g(Z_i, \widehat{\beta}, \widehat{\theta})}$, $i = 1, \ldots, n$. Center and standardize $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$ such that they have mean zero and variance one.

3. Obtain a bootstrap sample from the standardized variables obtained in Step 2, call them $\widehat{\epsilon}_1^*, \ldots, \widehat{\epsilon}_n^*$, and define $Y_i^* = f(X_i, \widehat{\beta}) + \sqrt{g(Z_i, \widehat{\beta}, \widehat{\theta})}\widehat{\epsilon}_i^*$, $i = 1, \ldots, n$.

4. For the bootstrap sample (X_i, Y_i^*, Z_i), $i = 1, \ldots, n$, calculate the pseudo-likelihood estimator $(\widehat{\beta}^*, \widehat{\theta}^*)$ under the null hypothesis, let $\widehat{r}_i^* = (Y_i^* - f(X_i, \widehat{\beta}^*))^2 - g(Z_i, \widehat{\beta}^*, \widehat{\theta}^*)$. The bootstrap version of the test statistic is

$$T_n^* = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h^q} K\left(\frac{Z_i - Z_j}{h}\right) \widehat{r}_i^* \widehat{r}_j^*. \tag{10}$$

5. Repeat steps 3 and 4 a large number of times. For a specified nominal level of the test, the critical value is then determined as the appropriate quantile of the bootstrap distribution of the test statistic.

4. Numerical Simulations

We investigate the performance of the proposed test in finite sample sizes. The test is calculated with 400 simulation runs and nominal level 0.05. The simulated level thus has a Monte Carlo error of $\sqrt{0.05 * 0.95/400} \approx 1\%$. We use 200 bootstrap samples per run to obtain the critical value. The random data are generated using the statistical software *R*. In the two simulation examples below, we evaluate the goodness-of-fit of the log-linear variance model and the power-of-the-mean model, respectively. To investigate the influence of the smoothing parameter, we report the simulation results for different choices of h, which reflect different degrees of smoothness.

Simulation study 1: log-linear variance function. For this model, we compare the nonparametric test T_n with the classical Wald test. To test for the log-linear variance structure $\sigma_i^2 = exp(\theta_0 + \theta_1 x_i)$, the Wald test fits a more general variance model $\sigma_i^2 = exp(\theta_0 + \theta_1 x_i + \theta_2 x_i^2)$ and evaluates whether the coefficient of the quadratic term θ_2 is zero.

We generate $Y_i = 1 + 2x_i + \sigma_i \epsilon_i$, $i = 1, \ldots, n$, where the x_i's are uniformly distributed on (0,1). The ϵ_i's are taken to be independent standard normal

random variables in order to make fair comparison with the Wald test. Three different functional forms are considered

$$\begin{aligned}
(1) \quad \sigma_i &= exp(-0.5 - 0.25x_i), \\
(2) \quad \sigma_i &= exp(-0.5 - 0.25x_i - 6(x_i - 0.5)^2), \\
(3) \quad \sigma_i &= exp(-0.5 - 0.25x_i - 1.5(sin(2\pi x_i))^2).
\end{aligned} \tag{11}$$

Note that functional form (1) corresponds to the null hypothesis.

Table 1 summarizes the proportion of times the null hypothesis is rejected by the two tests for two different sample sizes $n = 50$, 100 and four different choices of the smoothing parameter h: $0.10, 0.15, 0.20$ and 0.25.

[Table 1 about here.]

It is observed that the T_n test maintains the specified nominal level very well under the null hypothesis while the large-sample Wald test shows a marked tendency to exceed the nominal level. The performance of the Wald test improves with larger sample size. Our simulations (not reported) give an estimated type I error 0.063 for the Wald test when the sample size increases to 150. For the second functional form of σ_i, the Wald test is more powerful than T_n for sample size $n = 50$ but the power of T_n catches up for $n = 100$. This is not surprising since the second alternative is designed to the advantage of the Wald test. Indeed, the Wald test is most powerful if the true deviation from the log-linear variance structure happens in the log-quadratic direction but it can exhibit inferior power if the deviation happens in other directions. In contrast, the smoothing-based conditional moment test is less powerful than the Wald test when the deviation is in the log-quadratic direction, but

it can be much more powerful than the Wald test for deviations in many other directions. This is demonstrated by the simulation results for alternative (3), where the Wald test has very low power while the T_n test shows very high power.

Simulation study 2: power-of-the-mean variance function. This model assumes $\sigma_i^2 = \theta_1(f(X_i, \beta))^{\theta_2}$. In theory, if one is willing to assume a parametric error distribution, a parametric test such as a likelihood-based test can be constructed. However, this is rarely done in practice because unlike the log-linear variance structure where the log-quadratic variance structure provides a natural extended model, such natural nested structure can not be easily specified for the power-of-the-mean variance model.

We generate $Y_i = 20 + 10x_{1i} + 10x_{2i} + \sigma_i\epsilon_i$, $i = 1, \ldots, n$, where the x_{1i}'s are uniformly distributed on (0,1), and the x_{2i}'s are uniformly distributed on (-1.5,1.5). Three different functional forms are considered for σ_i:

$$\begin{aligned} (1) \quad \sigma_i &= 0.05\mu_i^{0.25}, \\ (2) \quad \sigma_i &= 0.05(\mu_i^{0.25} + e^{0.08\mu_i}), \\ (3) \quad \sigma_i &= 0.05(\mu_i^{0.25} + 5x_{2i}^2), \end{aligned} \tag{12}$$

where $\mu_i = 20 + 10x_{1i} + 10x_{2i}$ is the mean for the ith observation. We also consider three different error distributions for the ϵ_i: standard normal, t-distribution with four degrees of freedom, and lognormal. For comparison purpose, the random errors from the t-distribution or lognormal distribution are standardized to have mean zero and variance one.

For two different sample sizes $n = 50, 100$ and four different bandwidths $h = 0.10, 0.15, 0.20$ and 0.25, the proportion of times the nonparametric

test rejects the null hypothesis for various scenarios is summarized in Table 2. The simulation results indicate that the observed level is quite close to the specified nominal level 0.05 for different choices of error distributions, bandwidths and sample sizes. The power performance is also satisfactory. The power is higher for normal errors than for the heavier-tailed errors and increases with the sample size.

[Table 2 about here.]

5. Applications to Esterase Count Data

For the Esterase count data discussed in the introduction, Carrol and Ruppert suggested to fit a linear mean regression function. The local linear smoother imposed on the scatter plot in the top panel of Figure 1 indicates the overall fit of the linear mean function. We further check the validity of this assumption using the test of Zheng (1996). A plot of the p-value of Zheng's test versus the smoothing parameter h is exhibited as the solid line in the bottom panel of Figure 1. Such a plot is often referred to as a *smoothing trace* of the test, see for example King, Hart and Wehrly (1991) and Young and Bowman (1995). The high p-values, for all choices of h, support the linear mean function assumption.

For most of the immunoassays data analysis in the literature, the variance is assumed to be proportional to the mean, which leads to the following regression model for the esterase data

$$Y_i = \beta_0 + \beta_1 x_i + \sigma(\beta_0 + \beta_1 x_i)^{\theta} \epsilon_i, \quad i = 1, \ldots, 108, \tag{13}$$

where the ϵ_i's are independent random errors with mean 0 and variance 1. To test for the adequacy of the power-of-the-mean variance structure,

the nonparametric test T_n gives p-values much higher than 0.05 for a wide range values of h, see the dashed line in the bottom panel of Figure 1. The smoothing trace suggests that the T_n test provides no evidence against the power-of-the-mean variance structure. The pseudo-likelihood method gives estimators for model (13): $\widehat{\beta_0} = -37.42$ with an estimated standard error 12.11, $\widehat{\beta_1} = 18.16$ with an estimated standard error 0.95, $\widehat{\theta} = 1.03$ with an estimated standard error 0.10, and the scale parameter σ is estimated to be 0.24.

Since the response is RIA count, one might anticipate Poisson variation, i.e., the power-of-the-mean model with $\theta_2 = 0.5$. The above estimated model indicates that the esterase data are more heteroscedastic than what Poisson variation would suggest. Merely for comparison purposes, we check the validity of the later model using the T_n test and obtain significant p-values for a wide range of h. The smoothing trace for testing this hypothesis is plotted as the dotted line in the bottom panel of Figure 1. Thus, Possion variation does not seem to be reasonable for the esterase data.

6. Unknown Mean Regression Function

Although in practice a parametric variance function is usually paired with a parametric mean function, it is helpful to relax the assumption that the mean function is completely known when checking the adequacy of the variance function. For this purpose, we consider the following general heteroscedastic regression model:

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

where X is an l-dimensional vector of covariates and the mean function $m(x)$ is only assumed to be smooth, the ϵ_i's are independent with mean zero and variance one. We are interested in testing $H_0 : \sigma^2(x) = g(x, \theta)$ for some θ, i.e., whether the variance function $\sigma^2(x)$ can be modeled parametrically. Of particular interest is to test whether the variance function is constant.

The main advantage of considering model (14) is to avoid the likely adverse effects of a misspecified parametric mean model on checking the lack-of-fit of the variance model. On the other hand, one might be interested in comparing the result of the test in this section with that of the test in Section 2. Any discrepancy suggests possible misspecification of the parametric mean model. The practical implementation of the test proposed in this section requires nonparametric smoothing and may be hindered by the "curse of dimensionality", a problem frequently encountered in high-dimensional data.

Let $\widehat{m}(x)$ be a kernel-smoothing estimator of $m(x)$. Hall and Carroll (1989) verified that the parameters in the parametric variance function can be consistently estimated with $\sqrt{n}$-rate if $m(x)$ is Lipschitz smooth of order $1/2$ or more. Denote $\widehat{r}_i = (Y_i - \widehat{m}(x_i))^2 - g(x_i, \widehat{\theta})$, where $\widehat{\theta}$ is an estimator of θ. Then the $\widehat{r}_i$ estimates $r_i = (Y_i - m(x_i))^2 - g(x_i, \theta)$, which has mean zero under the null hypothesis. Define the test statistic similarly as before

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h^l} K\left(\frac{X_i - X_j}{h}\right) \widehat{r}_i \widehat{r}_j. \tag{15}$$

A somewhat more involved proof (sketched in the appendix) shows that under H_0, as $n \to \infty$, $h \to 0$ and $nh^l \to \infty$,

$$nh^{l/2} T_n \to N(0, \xi^2) \tag{16}$$

in distribution, where $\xi^2 = 2 \int K^2(u)du \int g^4(x, \theta)(E(\epsilon_i^4|x) - 1)^2 p^2(x)dx$.

We explore the finite sample property of the proposed test through a small Monte Carlo study, where the goal is to test whether the variance is homoscedastic, i.e., whether g is a constant function. The random data are generated from $Y_i = 0.5 + 3(x_i - 0.5)^2 + 0.25\epsilon_i,\ i = 1, \ldots, n$, where x_i is uniformly distributed on (0,1) and the ϵ_i's are independent standard normal random variables. We compare the test of this section (denoted by T_{n1}) with the test in Section 2.2 that assumes a quadratic mean function (denoted by T_{n2}) and the test in Section 2.2 with a linear mean function (denoted by T_{n3}). Thus T_{n2} represents the case in which a correct mean model is used and T_{n3} uses an incorrectly specified mean model. For T_{n1}, a bootstrap procedure similar to that in Section 3.2 is used, where $f(X_i, \widehat{\beta})$ is replaced by a nonparametric estimator using kernel smoothing with the optimal plug-in bandwidth. For three different sample sizes $n = 50, 100$ and 150, and four different bandwidths $h = 0.10, 0.15, 0.20$ and 0.25, the estimated levels of the three tests are displayed in Table 3. It is not surprising that T_{n3} appears to be very liberal since the mean function is incorrectly specified. It is also observed that compared with T_{n2} where the mean function is correctly specified, it takes much larger sample size for T_{n1} to work properly. Thus the test with an unknown mean function is not as efficient as the test with a correctly specified parametric mean function, but a test with an incorrectly specified parametric mean function may seriously impair the test for the variance function.

[Table 3 about here.]

7. Summary

We have developed a nonparametric test for assessing the adequacy of an assumed variance structure in heteroscedastic regression models. The emphasis of this paper is on the case where the mean function has a known parametric form. This is motivated by the fact that in practice when a parametric form is assumed for a higher moment (the variance), a parametric form is almost always assumed for the lower moment (the mean). We have also discussed a generalization where the mean function is only assumed to be smooth.

Acknowledgements

We would like to thank the AE, a referee and the Co-editor, whose comments led to significant improvement of the paper. Professor Xiao-Hua Andrew Zhou is presently a Core Investigator and Director of the Biostatistics Unit at the Northwest HSR&D Center for Excellence within the VA Pugest Sound Health Care System. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. The research reported here was supported in part by department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, ECI-03-206 and in part by AHRQ grant R01HS013105.

References

[1] Arbogast, P. G. and Bedrick, E. J. (2004). Model-checking techniques for linear models with parametric variance functions. *Technometrics*. **46,** 404–410.

[2] Bedrick, E. J. (2000). Checking for lack of fit in linear models with parametric variance functions *Technometrics.* **42,** 227–236.

[3] Breusch, T. S. and Pagan, A. R. (1980). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47,** 1287–1294.

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

[5] Carroll, R. and Ruppert, D (1988). Transformation and weighting in regression. New York: Chapman & Hall.

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

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

[8] Davidian, M. Carroll, R. and Smith, W. (1988). Variance functions and the minimum detectable concentration in assays. *Biometrika.* **75,** 549–556.

[9] Hall, P. (1984), Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis,* **14,** 1–16.

[10] Hall, P. and Carroll, R. J. (1989), Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society, Series B,* **51,** 3–14.

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

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

[13] Müller, H. G. and Zhao, P. L. (1995). On a semiparametric variance model and a test for heteroscedasticity. *Ann. Statist.* **23,** 946–967.

[14] Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997), "Local polynomial variance-function estimation", *Technometrics*, **39,** 262-273.

[15] Stute, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *The Annals of Probability.* **12,** 361–379.

[16] White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica.* **48,** 817–838.

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

[18] Zheng, J. X. (1996), A consistent test of functional form via nonparametric estimation techinques. *Journal of Econometrics,* **75,** 263–289.

[19] Zhou, X.H. , Stroupe, K.T. and Tierney, W.M. (2001). Regression analysis of health care charges with heteroscedasticity. *JRSS, Ser. C.* **50(3)**, 303-312.

Appendix A

Sketch of Proofs

Proof of (5). Since $\widehat{r}_i = (Y_i - f(X_i, \widehat{\beta}))^2 - g(Z_i, \widehat{\beta}, \widehat{\theta})$, $r_i = (Y_i - f(X_i, \beta))^2 - g(Z_i, \beta, \theta)$, we have $\widehat{r}_i = r_i + 2\epsilon_i(f(X_i, \beta) - f(X_i, \widehat{\beta})) + (f(X_i, \beta) - f(X_i, \widehat{\beta}))^2 + (g(Z_i, \beta, \theta) - g(Z_i, \widehat{\beta}, \widehat{\theta}))$. As a result, T_n can be decomposed as a sum of ten terms: $T_n = T_n' + \sum_{i=1}^{9} Q_i$, where

$$
\begin{aligned}
Q_1 &= \frac{4}{n(n-1)h^q} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K\left(\frac{Z_i - Z_j}{h}\right) r_i \epsilon_j (f(X_j, \beta) - f(X_j, \widehat{\beta})), \\
Q_2 &= \frac{2}{n(n-1)h^q} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K\left(\frac{Z_i - Z_j}{h}\right) r_i (f(X_j, \beta) - f(X_j, \widehat{\beta}))^2,
\end{aligned}
$$

and Q_i, $i = 3, \ldots, 9$, are similarly defined. Let $\frac{\partial f(X_j, \beta)}{\partial \beta}$ be the $m \times 1$ vector with the ith element $\frac{\partial f(X_j, \beta)}{\partial \beta_i}$, and $\frac{\partial f(X_j, \beta)}{\partial \beta'}$ be the transpose of this vector. Let $\frac{\partial^2 f(X_j, \beta)}{\partial \beta \partial \beta'}$ be an $m \times m$ matrix with the (i, k)th element $\frac{\partial^2 f(X_j, \beta)}{\partial \beta_i \partial \beta_k}$, then we have

$$
\begin{aligned}
Q_1 &= \frac{4}{n(n-1)h^q} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K\left(\frac{Z_i - Z_j}{h}\right) r_i \epsilon_j \frac{\partial f(X_j, \beta)}{\partial \beta'} (\beta - \widehat{\beta}) \\
&\quad + (\beta - \widehat{\beta})' \frac{4}{n(n-1)h^q} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K\left(\frac{Z_i - Z_j}{h}\right) r_i \epsilon_j \frac{\partial^2 f(X_j, \overline{\beta})}{\partial \beta \partial \beta'} (\beta - \widehat{\beta}) \\
&= Q_{11}(\beta - \widehat{\beta}) + (\beta - \widehat{\beta})' Q_{12} (\beta - \widehat{\beta}),
\end{aligned}
$$

where the definition of Q_{11} and Q_{12} should be clear from the context, $\overline{\beta}$ depends on X_j and lies between β and $\widehat{\beta}$. Note that the r_i's are independent

with mean 0, thus $E(Q_{11}) = 0$ and

$$\begin{aligned} & E(Q_{11}^2|X,Z) \\ = & \frac{16}{n^2(n-1)^2h^{2q}} \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1,j_1\neq i_1}^{n}\sum_{j_2=1,j_2\neq i_2}^{n} K\left(\frac{Z_{i_1}-Z_{j_1}}{h}\right) K\left(\frac{Z_{i_2}-Z_{j_2}}{h}\right) \\ & E(r_{i_1}r_{i_2}\epsilon_{j_1}\epsilon_{j_2})\frac{\partial f(X_{j_1},\beta)}{\partial\beta'}\frac{\partial f(X_{j_2},\beta)}{\partial\beta'}. \end{aligned}$$

In order for the expectation to be nonzero, we must have $i_1 = i_2$ and $j_1 = j_2$ or $i_1 = j_2$ and $i_2 = j_1$, we have $E(Q_{11}^2|X,Z) = O(n^{-4}h^{-2q})O(n^2) = O(n^{-2}h^{-2q})$. Since the quasi-likelihood estimator $\widehat{\beta}$ is $\sqrt{n}$-consistent for β, we have $nh^{q/2}Q_{11}(\beta-\widehat{\beta}) = O(nh^{q/2})O_p(n^{-1}h^{-q})O_p(n^{-1/2}) = O_p(n^{-1/2}h^{-q/2}) = o_p(1)$. Similarly, $Q_{12} = O_p(1)$ and $nh^{q/2}(\beta-\widehat{\beta})'Q_{12}(\beta-\widehat{\beta}) = O_p(h^{q/2}) = o_p(1)$. Therefore $nh^{q/2}Q_1 = o_p(1)$. Similarly, we can show $nh^{q/2}Q_i = o_p(1)$, $i = 2,\ldots,9$. □

Proof of (6). From (5), $nh^{q/2}T_n$ and $nh^{q/2}T_n'$ have the same asymptotic distribution. Since the r_i's are independent with mean 0, $nh^{q/2}T_n'$ is a second-order degenerate U-statistic. Its asymptotic normality can be established by checking the condition of Theorem 1 of Hall (1984). □

Proof of (16). For $\widehat{r}_i = (Y_i - \widehat{m}(X_i))^2 - g(X_i,\widehat{\theta})$, where $\widehat{m}(X_i) = [(n-1)h^l]^{-1}\sum_{k\neq i} Y_k K((X_k-X_i)/h)/\widehat{p}(X_i)$ and $\widehat{p}(X_i) = [(n-1)h^l]^{-1}\sum_{k\neq i} K((X_k-X_i)/h)$, and $r_i = (Y_i-m(X_i))^2 - g(X_i,\theta)$, we have $\widehat{r}_i = r_i + 2\sigma(X_i)\epsilon_i(m(X_i)-\widehat{m}(x_i)) + (m(X_i)-\widehat{m}(X_i))^2 + [g(X_i,\theta)-g(X_i,\widehat{\theta})]$. Similarly as in the proof of (5), T_n can be decomposed as a sum of ten terms: $T_n = T_n' + \sum_{i=1}^{9} Q_i$, where $T_n' = [n(n-1)h^l]^{-1}\sum_{i=1}^{n}\sum_{j\neq i} K\left(\frac{X_i-X_j}{h}\right) r_i r_j$, and

$$Q_1 = \frac{4}{n(n-1)h^l}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_i-X_j}{h}\right) r_i\sigma(X_j)\epsilon_j(m(X_j)-\widehat{m}(X_j)),$$

$$Q_2 = \frac{2}{n(n-1)h^l}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) r_i (m(X_j) - \widehat{m}(X_j))^2,$$

and Q_i, $i = 3, \ldots, 9$, are similarly defined. To show $nh^{l/2}Q_1 = o_p(1)$, we make use of the following fact:

$$\begin{aligned} m(X_j) - \widehat{m}(X_j) &= \frac{\widehat{s}(X_j) - s(X_j)}{p(X_j)} - \frac{(\widehat{s}(X_j) - s(X_j))(\widehat{p}(X_j) - p(X_j))}{p(X_j)\widehat{p}(X_j)} \\ &\quad - \frac{s(X_j)(\widehat{p}(X_j) - p(X_j))}{p^2(X_j)} + \frac{s(X_j)(\widehat{p}(X_j) - p(X_j))^2}{p^2(X_j)\widehat{p}^2(X_j)}, \end{aligned}$$

where $s(X_j) = m(X_j)p(X_j)$ and $\widehat{s}(X_j) = \widehat{m}(X_j)\widehat{p}(X_j)$. Based on the above decomposition, $nh^{l/2}Q_1$ can be written as $nh^{l/2}Q_1 = Q_{11} + Q_{12} + Q_{13} + Q_{14}$. For instance,

$$Q_{11} = \frac{4}{n(n-1)h^l}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) r_i \sigma(X_j)\epsilon_j \frac{\widehat{s}(X_j) - s(X_j)}{p(X_j)}.$$

Since

$$\begin{aligned} &\widehat{s}(X_j) - s(X_j) \\ = &\frac{1}{(n-1)h^l}\sum_{k\neq j} K\left(\frac{X_i - X_j}{h}\right)(m(X_k) - m(X_j)) \\ &+ \frac{1}{(n-1)h^l}\sum_{k\neq j} K\left(\frac{X_i - X_j}{h}\right)\sigma(X_k)\epsilon_k + m(X_j)(\widehat{p}(X_j) - p(X_j)), \end{aligned}$$

Q_{11} can be further written as $Q_{11} = Q_{11A} + Q_{11B} + Q_{11C}$. By directly checking mean and variance, we can show $Q_{11A} = o_p(1)$, $Q_{11B} = o_p(1)$. And we can show $Q_{11C} = o_p(1)$ by employing a result of Stute (1984): $\sup_x |\widehat{p}(x) - p(x)| = (n^{-1}h^{-l}(lnh^{-l}))^{1/2}$ almost surely. This proves that $Q_{11} = o_p(1)$. Similarly, we can show $Q_{1i} = o_p(1)$, for $i = 2, 3, 4$, which yields $Q_1 = o_p(1)$. We prove $nh^{l/2}(T_n - T_n') = o_p(1)$ by showing $Q_i = o_p(1)$, for $i = 2, \ldots, 9$ using the same technique. The asymptotic normality is proved by applying the result of Hall (1984) on T_n'. □

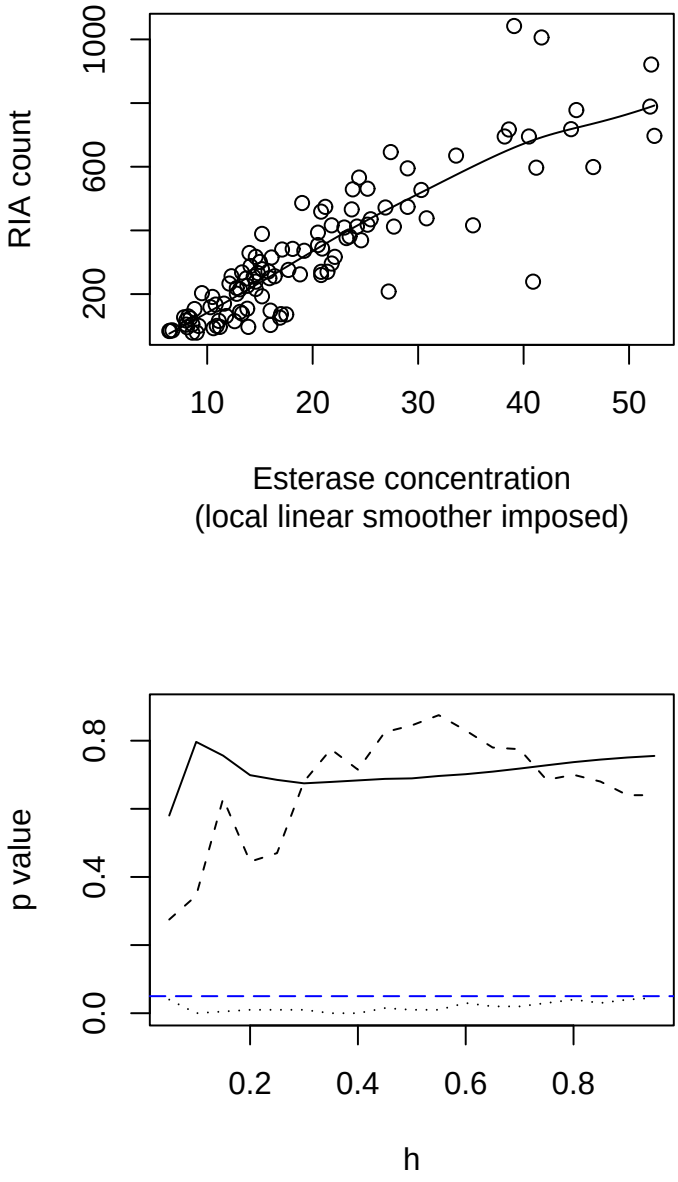

Figure 1. Analysis of Esterase data. The top graph is a scatter plot; the bottom graph contains smoothing traces for three different hypotheses: The solid line is for testing the linearity of the mean function; the dashed line is for testing the power-of-the-mean variance structure, the dotted line is for testing Poisson variation, and the horizontal dashed line has intercept 0.05.

Table 1

Estimated powers of the T_n test and the Wald test for the three functional forms of $\sigma(x_i)$ specified in (11) *and two different sample sizes $n = 50, 100$. The nominal level is 0.05.*

$\sigma(x_i)$	h	$n = 50$		$n = 100$	
		T_n test	Wald test	T_n test	Wald test
(1)	0.10	0.048	0.103	0.055	0.070
	0.15	0.050		0.053	
	0.20	0.048		0.053	
	0.25	0.050		0.053	
(2)	0.10	0.633	0.995	0.943	1.000
	0.15	0.735		0.970	
	0.20	0.780		0.983	
	0.25	0.810		0.988	
(3)	0.10	0.658	0.140	0.973	0.165
	0.15	0.690		0.980	
	0.20	0.648		0.963	
	0.25	0.513		0.903	

Table 2

Estimated powers of the T_n test for the three functional forms of $\sigma(x_i)$ specified in (12)*, three different error distributions and two different sample sizes $n = 50, 100$. The nominal level is 0.05.*

		$n = 50$			$n = 100$		
$\sigma(x_i)$	h	normal	t_4	lognormal	normal	t_4	lognormal
(1)	0.10	0.053	0.058	0.045	0.063	0.063	0.045
	0.15	0.043	0.050	0.050	0.043	0.058	0.025
	0.20	0.038	0.050	0.040	0.040	0.050	0.048
	0.25	0.053	0.043	0.045	0.030	0.040	0.045
(2)	0.10	0.455	0.305	0.180	0.773	0.473	0.238
	0.15	0.533	0.350	0.183	0.848	0.550	0.230
	0.20	0.598	0.383	0.190	0.885	0.608	0.268
	0.25	0.665	0.418	0.193	0.933	0.645	0.248
(3)	0.10	0.583	0.370	0.260	0.875	0.610	0.313
	0.15	0.708	0.468	0.323	0.945	0.728	0.398
	0.20	0.750	0.505	0.343	0.968	0.790	0.443
	0.25	0.765	0.545	0.335	0.990	0.792	0.463

Table 3

Estimated powers of three tests for testing homoscedasticity when the mean function is quadratic. T_{n1} assumes unknown mean function and estimates it nonparametrically; T_{n2} assumes a quadratic mean function and T_{n3} assumes a linear mean function. The nominal level is 0.05.

sample size	h	test		
		T_{n1}	T_{n2}	T_{n3}
50	0.10	0.123	0.065	0.175
	0.15	0.088	0.063	0.110
	0.20	0.080	0.060	0.068
	0.25	0.080	0.043	0.030
100	0.10	0.080	0.073	0.530
	0.15	0.075	0.065	0.478
	0.20	0.075	0.050	0.330
	0.25	0.075	0.070	0.228
150	0.10	0.055	0.048	0.813
	0.15	0.058	0.048	0.780
	0.20	0.050	0.045	0.635
	0.25	0.050	0.048	0.430