Journal of Multivariate Analysis 136 (2015) 41–55

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Heteroscedasticity checks for single index models

Xuehu Zhu [a,b], Xu Guo [a,c], Lu Lin [b], Lixing Zhu [a,*]

[a] Hong Kong Baptist University, Hong Kong

[b] Shandong University, Jinan, China

[c] Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

Article history:
Received 9 October 2013
Available online 19 January 2015

AMS subject classification:
62Fxx

Keywords:
Heteroscedasticity check
Single index model
Nonparametric estimation
Dimension reduction

ABSTRACT

To test heteroscedasticity in single index models, in this paper two test statistics are proposed via quadratic conditional moments. Without the use of dimension reduction structure, the first test has the usual convergence rate in nonparametric sense. Under the dimension reduction structure of mean and variance functions, the second one has faster convergence rate to its limit under the null hypothesis, and can detect local alternative hypotheses distinct from the null at a much faster rate than the one the first test can achieve. Numerical studies are also carried out to evaluate the performance of the developed tests. Interestingly, the second one works much better than the first one if the variance function does have a dimension reduction structure. However, it is not robust against the violation of dimension reduction structure, in other words, the power performance of the second test may not be encouraging if without the dimension reduction structure.

1. Introduction

The single index model (SIM) with the scalar outcome variable Y and p-dimensional covariate X is formulated as

$$Y = g(X^\tau \beta) + \varepsilon, \quad E(\varepsilon|X) = 0, \tag{1.1}$$

where $g(\cdot)$ is an unknown smooth function, β is a p-dimensional unknown parameter vector, and ε is the error term whose conditional expectation given X is zero. Here the notation X^τ in (1.1) denotes transposition of X. For identifiability consideration, we assume that the parameter vector β satisfies $\|\beta\| = 1$ and the first component of β is positive, where $\|\cdot\|$ stands for the Euclidean norm. If the link function $g(\cdot)$ is given in advance, the SIM reduces to a generalized linear regression model. Thus, the SIM is comparably flexible in model structure. Further, compared with fully nonparametric regression models, the SIM captures the information of Y through one-dimensional variable $X^\tau\beta$. This feature makes the SIM retain a better interpretability and avoid the curse of dimensionality commonly occurred in nonparametric regression models. Therefore, as a compromise between fully parametric and fully nonparametric regression models, the SIM has drawn much attention due to its wide use in several research fields such as economics and statistics, see [17,12]. The readers can refer to [10] for more detailed information about the SIM. There are two commonly used assumptions on the variance function $E(\varepsilon^2|X)$ in the literature: the first is that the variance function is purely nonparametric, and the second requires that the variance function has a dimension reduction structure like the mean function has. This is often in the case in models with dimension reduction structure such as generalized linear models. In more general semiparametric settings, [2] pointed out

* Corresponding author.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

http://dx.doi.org/10.1016/j.jmva.2015.01.007

that when the central subspace and the central mean subspace have the same dimension, the second case holds naturally. Particularly, there exist a number of proposals dealing with the second case, see for example [16,18,31].

Estimating the mean function in the SIM has been extensively discussed in the literature. For instance, [12] proposed a semi-parametric least squares estimator for general single index structure models; [11] developed an average derivative estimation that can result in an estimator converging to the true value of the index parameter at the rate of $n^{-1/2}$. We will call it $\sqrt{n}$- consistent estimator. [25] proposed an adaptive approach, called minimum average variance estimation (MAVE) which can be used to the SIM with weaker conditions; [4] introduced a method of estimating functions to study the SIM; [18] further proposed a new estimation method based on distance covariance. However, these estimators have adverse consequences for the efficiency and can be even inconsistent in the presence of heteroscedasticity. Thus, heteroscedasticity testing is an important issue for the SIM. There are some relevant proposals in the literature. For instance, [3,21] obtained the score test statistic for parametric structure of the variance function when the mean function is respectively linear regression model and for the first-order autoregressive model. [19] developed a modified score test for linear models. [7,14] proposed tests that are based on a L^2-distance between the underlying and the hypothetical variance function. [30] proposed a marked empirical process based test of the squared residuals for nonparametric model. [5] extended the test statistic introduced by [28], which was originally for goodness of fit of the mean regression, to detect a possible heteroscedasticity. [8] suggested a test for parametric variance function in nonparametric regression model. [29] also considered the heteroscedasticity checking in nonlinear and nonparametric regression models. [15] extended the idea of [5] to semiparametric regressions. Other references include [26,6]. A relevant literature is [20] for testing the SIM structure.

When both the mean and variance functions have dimension reduction structures, it is often the case that both share an identical index such as generalized linear model and its extensions. In this case, we should take this useful structure into consideration when we construct a test statistic.

In this paper, we suggest two test statistics according to the model structures. The first one is a kernel-smoothing type nonparametric test that is against fully nonparametric heteroscedasticity. In fact, this test statistic can be used to check whether there is heteroscedasticity or not without assuming any specific formula of the variance function under the alternative hypothesis. However, when the dimension of covariates is large, it may suffer from the curse of dimensionality in nonparametric estimation. When both mean and variance functions have the common dimension reduction structure with the same index, a test incorporated with the dimension reduction structure is suggested such that the test can completely avoid the curse of dimensionality. The interesting features are as follows. When the dimension reduction structure holds, the second test has the faster convergence rate and can detect the local alternative hypotheses distinct from the null hypothesis at a faster rate than the one the first test can achieve. However, the second test could also perform badly when the dimension reduction structure does not hold. That is, it is not robust against nonparametric model structure. Thus, it is an interesting topic on how to construct a test that enjoys the advantages of two tests.

The rest of this paper is organized as follows. In Section 2, the kernel-smoothing based test statistic is constructed, the asymptotic null distribution and the power performance under the global alternative and local alternative hypotheses are investigated. In Section 3, another test for the models with a dimension reduction structure is suggested. In Section 4, the performances of our tests are examined by numerical simulations and the analyses for the Boston Housing Data and the Car Data. As the test statistics involve estimating β, a brief review and the proofs of the theorems are provided in the Appendix.

2. Testing heteroscedasticity with fully nonparametric variance function

For the SIM (1.1), the hypotheses of interesting are as follows:

$$H_0 : Var(\varepsilon|X = x) \equiv \sigma^2 \quad \text{v.s.} \quad H_1 : Var(\varepsilon|X = x) \not\equiv \sigma^2, \tag{2.1}$$

where $\sigma > 0$ is an unknown constant. Denote by $\{X_i, Y_i\}_{i=1}^n$ an $i.i.d$ sample from (X, Y) and let $\varepsilon_i = Y_i - g(X_i^\tau \beta)$. Note that under H_0, we have $E(\varepsilon_i^2|X_i) = Var(\varepsilon|X_i) = \sigma^2$. Then we obtain the following moment condition: under H_0,

$$S_1 = E\left[(\varepsilon_i^2 - \sigma^2)E((\varepsilon_i^2 - \sigma^2)|X_i)p(X_i)\right] = E\left[E^2((\varepsilon_i^2 - \sigma^2)|X_i)p(X_i)\right] = 0.$$

where $p(\cdot)$ denotes the density function of X which is supposed to be continuous. Under the alternative hypothesis H_1, $Var(\varepsilon|X = x) = \sigma^2(x) \not\equiv \sigma^2$, we have that

$$\begin{aligned} S_1 &= E\left[(\varepsilon_i^2 - \sigma^2)E((\varepsilon_i^2 - \sigma^2)|X_i)p(X_i)\right] = E\left[E^2((\varepsilon_i^2 - \sigma^2)|X_i)p(X_i)\right] \\ &= E\left[(Var(\varepsilon_i|X_i) - Var(\varepsilon_i))^2 p(X_i)\right] > 0. \end{aligned}$$

These observations motivate us to construct a test statistic based on a certain estimator of S_1. We first estimate $\Delta_1(x_i) = E((\varepsilon_i^2 - \sigma^2)|X_i = x_i)p(x_i)$ by

$$\hat{\Delta}_1(x_i) = \frac{1}{(n-1)h^p} \sum_{j\neq i, j=1}^{n} K\left(\frac{X_j - x_i}{h}\right)(\hat{\varepsilon}_j^2 - \hat{\sigma}^2),$$

where $\hat{\varepsilon}_i^2 = (Y_i - \hat{g}(X_i^\tau \hat{\beta}))^2$, $K(\cdot)$ is a multivariate kernel function, h is a bandwidth and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2$. Here $\hat{g}(\cdot)$ and $\hat{\beta}$ are the estimators of $g(\cdot)$ and β respectively. Because they can be obtained by familiar methods, the estimation formulas will be given respectively in the following computation steps and the Appendix.