Computational Statistics and Data Analysis 103 (2016) 263–283

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Heteroscedasticity testing for regression models: A dimension reduction-based model adaptive approach

Xuehu Zhu [a,d], Fei Chen [c], Xu Guo [e,f], Lixing Zhu [b,d,*]

[a] School of Mathematics and Statistics, Xi'an Jiaotong University, China
[b] School of Mathematics and Statistics, Suzhou University of Science and Technology, China
[c] School of Statistics and Mathematics, Yunnan University of Finance and Economics, China
[d] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[e] School of Statistics, Beijing Normal University, China
[f] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, China

ARTICLE INFO

Article history:
Received 19 June 2015
Received in revised form 5 January 2016
Accepted 24 April 2016
Available online 20 May 2016

Keywords:
Heteroscedasticity testing
Model-adaption
Sufficient dimension reduction

ABSTRACT

Heteroscedasticity testing is of importance in regression analysis. Existing local smoothing tests suffer severely from curse of dimensionality even when the number of covariates is moderate because of use of nonparametric estimation. A dimension reduction-based model adaptive test is proposed which behaves like a local smoothing test as if the number of covariates was equal to the number of their linear combinations in the mean regression function, in particular, equal to 1 when the mean function contains a single index. The test statistic is asymptotically normal under the null hypothesis such that critical values are easily determined. The finite sample performances of the test are examined by simulations and a real data analysis.

1. Introduction

As heteroscedasticity structure would make a regression analysis more different than that under homoscedasticity structure, a heteroscedasticity check is required to accompany before stepping to any further analysis since ignoring the presence of heteroscedasticity may result in inaccurate inferences, say, inefficient or even inconsistent estimates. Consider a regression model with the nonparametric variance model:

$$Var(Y|X) = Var(\varepsilon|X), \tag{1}$$

where Y is the response variable with the vector of covariates $X \in \mathbb{R}^p$ and the error term ε satisfies $E(\varepsilon|X) = 0$. Heteroscedasticity testing for the regression model (1) has received much attention in the literature. Cook and Weisberg (1983) and Tsai (1986) proposed respectively two score tests for a parametric structure variance function under linear regression models and first-order autoregressive models. Simonoff and Tsai (1994) further developed a modified score test under linear models. Zhu et al. (2001) suggested a test that is based on squared residual-marked empirical process. Liero (2003) advised a consistent test for heteroscedasticity in nonparametric regression models, which is based on the L^2-distance between the underlying and hypothetical variance functions. This test is analogous to the one proposed by Dette and Munk (1998). Dette (2002), Zheng (2009) and Zhu et al. (2015a), extended the idea of Zheng (1996), which was primitively used for testing mean

* Corresponding author at: Department of Mathematics, Hong Kong Baptist University, Hong Kong.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

http://dx.doi.org/10.1016/j.csda.2016.04.009
0167-9473/

regressions, to heteroscedasticity check under several different regression models. Further, Lin and Qu (2012) extended the idea of Dette (2002) to semi-parametric regressions. Moreover, Dette et al. (2007) studied a more general problem of testing the parametric form of the conditional variance under nonparametric regression models.

The hypotheses of interest are:

$$H_0 : \exists\, \sigma^2 > 0 \quad \text{s.t. } P\{Var(\varepsilon|X) = \sigma^2\} = 1$$
$$\text{against}$$
$$H_1 : P\{Var(\varepsilon|X) = \sigma^2\} < 1, \quad \forall\, \sigma^2. \tag{2}$$

To motivate the test statistic construction, we comment on Zhu et al. (2001)'s test and Zheng (2009)'s test as the representatives of global smoothing tests and local smoothing tests, respectively. Thanks to the fact that under the null hypothesis,

$$E(\varepsilon^2 - \sigma^2|X) = 0 \Leftrightarrow E\left\{(\varepsilon^2 - \sigma^2)I(X \le t)\right\} = 0 \quad \text{for all } t \in \mathbb{R}^p,$$

Zhu et al. (2001) then developed a squared residual-marked empirical process as

$$V_n(x) = n^{-1/2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \{I(x_i \le x) - F_n(x)\},$$

where $\hat{\varepsilon}_i^2 = \{y_i - \hat{g}(x_i)\}^2$ with $\hat{g}(\cdot)$ being an estimate of the regression mean function. A quadratic functional form such as the Crämer–von Mises type test can be constructed. But, there exist two obvious disadvantages of this global smoothing test though it works well even when the local alternative hypotheses converge to the null hypothesis at a rate of $O(1/\sqrt{n})$. First, the data sparseness in high-dimensional space means that this global smoothing test suffers from the dimensionality problem, even for large sample sizes. Second, it may be invalid in numerical studies of finite samples when the dimension of X is high. This is because the residual-marked empirical process for over heteroscedasticity involves nonparametric estimation of the mean function g and thus, the curse of dimensionality severely affects the estimation efficiency. Zheng (2009)'s test is based on a consistent estimate of $E\{E^2(\varepsilon^2 - \sigma^2|X)f(X)\}$ in the following form:

$$\tilde{S}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \ne i, j=1}^{n} \tilde{K}_h\left(x_i - x_j\right)(\hat{\varepsilon}_i^2 - \hat{\sigma}^2)(\hat{\varepsilon}_j^2 - \hat{\sigma}^2),$$

where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, $\tilde{K}_h(\cdot) = \tilde{K}(\cdot/h)/h^p$ with a p-dimensional multivariate kernel function $\tilde{K}(\cdot)$ and h is a bandwidth. As a local smoothing-based test, Zheng (2009)'s test can work in the scenario where the local alternative models converge to the hypothetical model at the rate of $O(n^{-1/2}h^{-p/4})$, where p denotes the dimension of the covariate X. Note that the bandwidth h converges to zero at a certain rate. Thus, $O(n^{-1/2}h^{-p/4})$ can be very slow when the dimension p is large. Local smoothing tests severely suffer from the curse of dimensionality. To illustrate those disadvantages, Fig. 1 in Section 4 depicts the empirical powers of Zheng (2009)'s test and Zhu et al. (2001)'s test across 2000 replications with the sample size of $n = 400$ against the dimension $p = 2, 4, 6, 8, 10, 12$ for a model. This figure clearly suggests a very significant and negative impact from the dimension for the power performance of Zheng (2009)'s test and Zhu et al. (2001)'s test: when p is getting larger, the empirical power is getting down to a very low level around 0.1 no matter the mean regression function $g(\cdot)$ is fully nonparametric or semiparametric with $\beta^\top X$ in the lieu of X. The details are presented in Section 4.

Therefore, how to handle the serious dimensionality problem is of great importance. The goal of the present paper is to propose a new test that has a dimension reduction nature.

If the mean regression model has some dimension structure, this structure information may be useful to construct efficient heteroscedasticity test statistics. Motivated by this observation, we consider a general regression model in the following form:

$$Y = g(B_1^\top X) + \delta(B_2^\top X)e, \tag{3}$$

where $\varepsilon = \delta(B_2^\top X)e$, B_1 is a $p \times q_1$ matrix with q_1 orthonormal columns and q_1 is a known number satisfying $1 \le q_1 \le p$, B_2 is a $p \times q_2$ matrix with q_2 orthonormal columns, q_2 is an unknown number satisfying $1 \le q_2 \le p$, e is independent of X with $E(e|X) = 0$ and the functions g and δ are unknown. This model is semiparametric in the mean regression function. We assume that under the null hypothesis, the function $\delta(\cdot)$ is a constant. It is worth noting that because the functions g and δ are unknown, the following model with nonparametric variance function $\delta(\cdot)$ can also be reformulated in this form:

$$\begin{aligned} Y &= g(B_1^\top X) + \delta(X)e = g(B_1^\top X) + \delta(B_2 B_2^\top X)e \\ &\equiv g(B_1^\top X) + \tilde{\delta}(B_2^\top X)e, \end{aligned}$$

where B_2 is any orthogonal $p \times p$ matrix. That is, $q_2 = p$. In other words, any nonparametric variance model (1), up to the mean function, can be reformulated as a special multi-index model with $q_2 = p$. This model covers many popularly used models in the literature, including the single-index models, the multi-index models and the partially linear single index models. When the model (3) is a single index model or partially linear single index model, the corresponding number of the index becomes one or two, respectively.