Parker Test for Heteroskedasticity Based on Sample Fitted Values

Jingming Jiang[1*], Guangming Deng[1,2]

[1]College of Science, Guilin University of Technology, Guilin, China
[2]Institute of Applied Statistics, Guilin, Guangxi, China
Email: *824512842@qq.com, dgm@glut.edu.cn

Abstract

To address the drawbacks of the traditional Parker test in multivariate linear models: the process is cumbersome and computationally intensive, we propose a new heteroscedasticity test. A new heteroskedasticity test is proposed using the fitted values of the samples as new explanatory variables, reconstructing the regression model, and giving a new heteroskedasticity test based on the significance test of the coefficients, It is also compared with the existing Parker test which is improved using the principal component idea. Numerical simulations and empirical analyses show that the improved Parker test with the fitted values of the samples proposed in this paper is superior.

Keywords

1. Introduction

A basic assumption of classical linear regression analysis is that the random error terms of the model μ_i are homoskedastic, *i.e.* they have the same variance σ^2. However, studies have shown that heteroskedasticity is an almost universal phenomenon when regression analysis is performed with cross-sectional or time-series data. Therefore, the study of heteroskedasticity in econometric modelling has become a hot issue for many scholars. There are many different tests for heteroskedasticity. For example, the graphical test, the Parker test, the Spearman's rank correlation test, the Glejser test, the White test, the G-Q test and so on [1]-[7], Bai Xuemei [8] proposed various methods to test heteroscedasticity, including Parker's test model and its existing shortcomings, Liu Ming and Huang Hengjun [9] proposed to use sample fitting value $\hat{y}_i$ as the standard

for sorting sample points when conducting G-Q test on the linear regression model with multiple explanatory variables, and improved the method of sorting sample points several times by using the order of each explanatory variable, thus improving the work efficiency.

Among them, the traditional Parker test can not only test the existence of heteroskedasticity in the one-dimensional linear regression model, but also write the specific expression of heteroskedasticity. However, for multiple linear regression models, the traditional Parker test does not have a specific equation to test, and can only test each explanatory variable one by one, which is a tedious process and computationally intensive, and can lead to multiple heteroskedasticity models. Tan Xin [10] *et al*. improved on this problem by using the idea of principal components to build a heteroskedasticity model with all sample principal components instead of a single explanatory variable, changing from multiple equations to one equation, thus greatly simplifying the traditional Parker test method for testing heteroskedasticity in multiple linear regression models. However, there are certain problems with this improved method. First, the significant coefficient of the principal component of the improved Parker test with the principal component idea does not mean that the heteroskedasticity is related to the corresponding explanatory variable, so the regression coefficients of the principal component cannot be compared with the regression coefficients of the corresponding independent variable in the traditional test. Secondly, the Parker test based on the idea of principal component also needs to calculate the principal component of the sample to test the heteroscedasticity. Therefore, on the premise of no loss of sample information, this paper uses the sample fitting value as a new explanatory variable to establish a regression model, carries out the significance test of the coefficient, and gives a new heteroscedasticity test method, which simplifies the steps of heteroscedasticity test for the multiple linear regression model. It is also compared with Tan Xin's improved Parker's test. To ensure the completeness of the study, a brief introduction to the definition of heteroskedasticity and the traditional Parker test is given below.

2. Heteroskedasticity Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \mu_i, (i = 1, 2, 3, \cdots, n, \, k \geq 1) \tag{1}$$

The linear regression model is a univariate regression model when $k = 1$ and multiple regression model when $k \geq 2$, the number of k explanatory variables is the number of explanatory variables, y called the explanatory variable (dependent variable) and $x_1, x_2, \cdots, x_k$ called the explanatory variable (independent variable). In the regression model (1), if there is $\text{var}(\mu_i) = \sigma^2$ homoscedasticity for all $i(i = 1, 2, 3, \cdots, n)$, regardless of the value taken x, then the μ_i have homoscedasticity, and when homoscedasticity is not satisfied, but the other basic assumptions are satisfied, *i.e.* when it is $\text{var}(\mu_i) = \sigma_i^2$ not equal to a constant, the variance of the random error term is no longer a constant, but is different from each other, then the error term is said μ_i to have heteroscedasticity. This

model is called a linear regression model with heteroskedasticity, or a heteroskedasticity model for short.

In particular, heteroskedasticity is usually unavoidable when discussing cross-sectional data. For example, when discussing a linear regression model of firm profits with a number of explanatory variables, it is clear that the profits of large-scale firms have greater volatility than those of small-scale firms. Another example is that when examining a linear regression relationship between household income (the explanatory variable) and savings (the dependent variable), since high-income households have a larger surplus in addition to necessary household expenditures and a larger discretionary component, the variability in the amount of savings is also greater, which is the variation in the amount of savings is greater.

Heteroskedasticity in the model generally arises from four sources: omission of some important explanatory variables from the model; poor model setup; measurement error arising from the sample data; and variation in error over time.

3. The Traditional Parker Test

The Park test was proposed by Park in 1966. The Park test is based on a residual diagram that suggests σ_i^2 a function of the explanatory variables x_i, and then formulates the diagram as a function of $\sigma_i^2 = \sigma^2 x_i^{\beta} \mathrm{e}^{\mu_i}$, taking the logarithm to obtain $\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln x_i + \mu_i$, as σ_i^2 it is unknown, Parker proposed that σ_i^2 be represented by the residual squared of e_i^2. The following regression is performed:

$$\ln e_i^2 = \ln \sigma^2 + \beta \ln x_i + \mu_i$$

If β the above equation is statistically significant, then the data is heteroscedastic; if β it is not statistically significant, then there is no heteroscedasticity.

Specific steps of the traditional Parker test.

Step 1: Estimate the original regression using ordinary least squares to derive the square of e_i^2 the sample residuals.

Step 2: Regress the e_i^2 log of the explanatory variables associated with the heteroskedasticity on the basis of:

$$\ln e_i^2 = \ln \sigma^2 + \beta \ln x_i + \mu_i$$

Step 3: Perform a statistical test on the above equation and reject the null hypothesis of homoscedasticity if β it is statistically significant, or accept the null hypothesis of homoscedasticity if β it is not statistically significant.

4. Parker's Test Improved by Principal Components Thinking

Principal component analysis of explanatory variable x_1, x_2, x_3, all the principal components obtained were logarithmically processed with e_i^2, and the following heteroscedasticity model was established according to the new data:

$$\ln e_i^2 = \ln \sigma^2 + b_1 \ln |z_{i1}| + b_2 \ln |z_{i2}| + b_3 \ln |z_{i3}| + \mu_i \qquad (2)$$

where z_{i1}, z_{i2}, z_{i3} denotes the principal components generated based on the explanatory variables x_{i1}, x_{i2}, x_{i3}.

The least squares method was used to estimate the coefficients for model (2). The significance of the coefficients b_1, b_2, b_3 in the model is tested using p-values. Comparing the p values obtained with $\alpha = 0.05$ the coefficients in the model, the presence of significant coefficients in the b_1, b_2, b_3 model indicates heteroskedasticity, while conversely, the assumption of homoskedasticity is satisfied.

5. An Improved Parker Test Based on Sample Fitted Values

The above-mentioned traditional Parker test method is a complex and cumbersome process of construct in k a structural form of heteroskedasticity to test for the presence of heteroskedasticity in the original model. Based on this, we propose to use the sample fitting value $\hat{y}$ as the new explanatory variable to establish the regression equation with the residual logarithm, and compare the two methods with the Parker test improved by the principle component idea to compare the effectiveness and simplicity of the two methods.

Specific steps to improve the Parker test.

Step 1: OLS estimation of Equation (1) to obtain the sample fit $\hat{Y}_i$ and residuals.

$$\hat{Y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \ (i = 1, 2, 3, \cdots, n, k \geq 1)$$

Step 2: Establish a e_i^2 regression with the logarithm of the fitted value $\hat{Y}_i$ as the explanatory variable and the logarithm of the squared residuals as the explained variable.

$$\ln e_i^2 = \ln \sigma^2 + \beta \ln \hat{Y}_i + \mu_i$$

Step 3: Perform a statistical test on the above equation, if β statistically significant, reject the null hypothesis of same variance; if β statistically insignificant, accept the null hypothesis of homoscedasticity.

6. Random Simulation

The study of heteroskedasticity test analysis through random simulation can demonstrate the usefulness and validity of both the Parker test with sample principal components as explanatory variables and the Parker test with sample fitted values as explanatory variables.

1) Generation of simulation data

To generate the random simulation data, three sets of sample variables x_1, x_2, x_3 all with mean 0, variance 1 and standard normal distribution, were set with a sample size of 400. Considering that the correlation between the three variables might affect the test results, an additional covariance (variance 1) was set between the two variables, Pearson's correlation coefficient was 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The parameters $b_0 = 5$, $b_1 = 8$, $b_2 = 3$, $b_3 = 2$ were set

and the explanatory variables y_i were generated as follows:

$$y_i = 5 + 8x_1 + 3x_2 + 2x_3 + \mu_i, (i = 1, 2, 3, \cdots, 400) \tag{3}$$

where μ_i is the 400 normal random terms from the mean of 0, set in the form of: $\mu = x_1 \xi$

where ξ are mutually independent random variables that follow a standard normal distribution.

Obviously μ this form of random term is highly susceptible to heteroskedasticity, and the heteroskedasticity is related to the explanatory variables x_1.

p_1 denotes the value b_1 of p the coefficient test, p_2 denotes the p value of the coefficient tes b_2, and p_3 denotes the value b_3 of p the coefficient test. Since the Parker test with the fitted value as the new explanatory variable has only one p value, the results of the test are written centered as shown in Table 1.

The following conclusions were drawn from Table 1.

Table 1. Significance tests for coefficients in the model (one simulation) Table

The correlation Coefficient between the explanatory variables r	P-value names for coefficient tests	value for Parker's test with sample Principal components as explanatory variables	Values for Parker's test with sample fit values as explanatory variables	whether the findings of the two sets of tests are the same
0.1	P1	2.03e−05		
	P2	1.28e−05	1.94e−06	Same
	P3	0.0525		
0.2	P1	4.48e−10		
	P2	0.82232	5.088e−09	Same
	P3	0.00584		
0.3	P1	1.19e−10		
	P2	0.869265	1.059e−10	Same
	P3	0.000508		
0.4	P1	4.76e−09		
	P2	0.000176	1.48e−15	Same
	P3	0.396450		
0.5	P1	3.21e−14		
	P2	0.00315	8.08e−15	Same
	P3	0.71990		
0.6	P1	7.29e−16		
	P2	1.92e−05	6e−12	Same
	P3	0.21		
0.7	P1	2e−16		
	P2	0.4502	2.2e−16	Same
	P3	0.0439		
0.8	P1	2e−16		
	P2	0.2852	2e−16	Same
	P3	0.0233		
0.9	P1	2e−16		
	P2	0.8128	2e−16	Same
	P3	0.0957		

a) From the Parker test modified with the principal component idea, it is clear from the results that not only are the coefficients p_1 significant, but some of the p_2, p_3 coefficients are also significant, contradicting our assumption that heteroscedasticity is x_1 correlated. This is because the Parker method, modified with the principal component idea, contains all the information, and the principal component z_1 contains the x_1, x_2, x_3 information, so the significant coefficient does not mean x_1 that the coefficient is significant.

b) Comparing the two methods, the significance results are the same. The improved Parker method using the fitted values of the samples is all significant, indicating the hypothesis that the heteroskedasticity is x_1 correlated, leading to the overall significance of this multiple linear regression model. In contrast to the improved Parker test with the principal component idea, the results of the new and improved method are clear at a glance, there are no conditions that violate the hypothesis, and only one p value is needed to obtain the presence of heteroskedasticity in the multiple linear regression model.

A further 10,000 randomised simulation experiments were conducted at each correlation coefficient occasion and the simulated data were used to test for heteroskedasticity, comparing the Parker test with the principal component as the explanatory variable with the Parker test under the modified approach, the results of which are shown in Table 2.

The following conclusions were drawn from Table 2.

Regardless of the magnitude of the correlation coefficients between the explanatory variables, the number of significance counts for the test with the sample principal component as the explanatory variable is lower than the number of significance counts under the improved method, indicating that the improved Parker test with the principal component idea violates the prior assumption of the existence of heteroskedasticity in the multiple linear regression model,

Table 2. Significance tests for coefficients in the model (10,000 simulations) Table.

The correlation coefficient between the explanatory variables r	Number of significant p-values for Parker's test with sample principal components as explanatory variables	Number of significant p-values for the Parker test with the sample fitted values as explanatory variables	Number of inconsistent findings between the two methods
0.1	9092	10,000	0
0.2	9915	10,000	0
0.3	9001	10,000	0
0.4	9034	10,000	0
0.5	9013	10,000	0
0.6	9230	10,000	0
0.7	9102	10,000	0
0.8	10,000	10,000	0
0.9	9020	10,000	0

further indicating that the improved method is superior to the method with the sample principal component as the explanatory variable. It follows that the new method can replace the method of using the main component as the interpretation variable. The new method is also more concise and easy to calculate.

7. Analysis of Practical Examples

1) Data sources

The regional gross domestic product (y), per capita consumption expenditure (x_1), per capita regional general budget expenditure (x_2), and price index of fixed asset investment (x_3) by region were collected in the statistical yearbook for 31 provinces in 2018, in RMB. First, the following multiple regression model is established:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \mu_i, (i = 1, 2, 3, \cdots, 31) \tag{4}$$

and the OLS regression of (4) is performed to obtain a set of fitted values $\hat{y}$, and the square of the residuals is calculated e_i^2

$$\hat{y} = -9.574\text{e}+09 + 1.493\text{e}+08 x_1 - 8.809\text{e}+03 x_2 + 9.255\text{e}+07 x_3 \tag{5}$$

a) Methodological steps for using sample principal components as new variables:

Step 1: OLS regression of equation (5) to obtain a set of residuals e_i^2

Step 2: Calculate the sample principal components, get the first, second and third principal components z_1, z_2, z_3, the new variable contains all the information of the explained quantity x_1, x_2, x_3, bring 31 groups of observations into the above equation to get $z_1, z_2, z_3, \cdots, z_{31}$

Step 3: Regress the logarithm of $\ln \sigma_i^2$ to z_i, which is

$$\ln e_i^2 = \ln \sigma^2 + \beta_1 \ln z_{i1} + \beta_2 \ln z_{i2} + \beta_3 \ln z_{i3} + \mu_i$$

Step 4: Calculate and obtain the significance test of the regression coefficients. Multiple regression model (4) have Heteroskedasticity.

b) Methodological steps with fitted values $\hat{Y}_i$ as new variables.

Step 1: Perform an OLS regression on the following equation to obtain a set of residuals e_i^2;

Step 2: Calculate the sample fit $\hat{Y}_i$ (as it reflects the overall variation of the data and captures information on variance changes) to obtain a set of $\hat{Y}_i$, $i = 1, 2, 3, \cdots, 31$;

Step 3: $\ln \sigma_i^2$ return the pairs $\hat{Y}_i$;

Step 4: Calculate and obtain the significance test of the regression coefficients.

Table 3. Sample principal components as new variables for Parker test output.

Logarithm of residuals Regression model $\ln \sigma_i^2$ on principal components z_i	R^2	Pr (>|t|)
$\ln e_i^2 = -336.0304 - 976.4507 \ln z_1 + 1.2851 \ln z_2 + 978.9833 \ln z_3$	0.1547	0.1102 0.0484 0.1109

Table 4. output of the Parker test with the sample fitted values as new variables.

Logarithm $\ln \sigma_i^2$ of the residuals Regression model of the fitted values $\hat{Y}_i$ to the sample	R^2	Pr (>|t|)
$\ln e_i^2 = 2.3875 + 1.7562 \hat{y}_i$	0.2432	0.00482

At the significance level $\alpha = 0.05$, the results of the test in **Table 3** show that the values β_2 of p are less than 0.05, indicating that the regression coefficient is significantly non-zero, which is the logarithmic value of the residuals is related to the second principal component, but it is still not possible to find whether the heteroskedasticity is related to x_1, x_2 or x_3.

At the significance level $\alpha = 0.05$, it was learned from **Table 4** tests that the p regression coefficient is less than 0.05, which means that the logarithm of the residuals is significantly non-zero, $\hat{Y}_i$ there is a correlation between the logarithm and the residuals.

The same conclusion was obtained using a modified Parker test, with heteroskedasticity in the multiple regression model.

In summary, it can be seen that the Parker test modified with principal components and the Parker test modified with sample fitted values reach the same conclusion, although both methods only require a heteroskedasticity model to obtain the existence of heteroskedasticity in the original model, the implementation of the Parker test modified with sample fitted values is simpler and faster in the experimental process, omitting the step of calculating the principal components of the sample, and the results are more valid. In contrast, the sample fit also contains all the information on the explanatory variables and the fit is more reflective of the variance trends in the overall data. Therefore, it can be concluded that the method using the sample fitted values as new variables is more effective than the method using the sample principal components. This indicates that the improved method can replace the Parker test with the principal components as the explanatory variables.

8. Conclusions

There are many different methods of testing for heteroskedasticity in regression models, and scholars at home and abroad have proposed many different tests that are more effective than the traditional methods. In the paper, a new test method is proposed on the basis of the traditional Parker test method. When constructing the auxiliary regression model, the regression equation of the logarithm of the residuals of the fitted original regression model against the fitted values calculated by least squares (OLS) is fitted to the original regression model, and the effectiveness and simplicity of the newly proposed method in the paper are proved after simulation and example analysis.

The deficiencies of the article and the prospect of the future:

when constructing the auxiliary regression model, the improved Parker test proposed by us only considers the regression equation of fitting residual loga-

rithm to fitting value, but does not consider the specific form of heteroscedasticity, which needs to be further studied in the future work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Denniscook, R. and Weisberg, S. (1983) Diagnostics for Heteroscedasticity in Regression. *Biometrika*, **70**, 1-10. https://doi.org/10.1093/biomet/70.1.1

[2] Park, R.E. (1966) Estimation with Heteroscedastic Error Terms. *Econometrica*, **34**, 888. https://doi.org/10.2307/1910108

[3] Yin, Y. and Carroll, R.J. (1990) A Diagnostic for Heteroscedasticity Based on the Spearman Rank Correlation. *Statistics & Probability Letters*, **10**, 69-76. https://doi.org/10.1016/0167-7152(90)90114-M

[4] Glejser, H. (1969) A New Test for Heteroscedasticity. *Journl of the American Statistical Association*, **64**, 316-323. https://doi.org/10.1080/01621459.1969.10500976

[5] White, H. (1980) A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heterosedasticity. *Econometrica*, **48**, 817-838. https://doi.org/10.2307/1912934

[6] Goldfeld, S.M. and Quandt, R.E. (1965) Some Tests for Homoscedasticity. *Publications of the American Statistical Association*, **60**, 539-547. https://doi.org/10.1080/01621459.1965.10480811

[7] Zhang, S. and Yu, Q.W. (1984) Econometrics. Shanghai Jiao Tong University Press, Shanghai.

[8] Bai, X.M. (2002) Test Methods and Review of Heterovariance. *Journal of Northeastern University of Finance and Economics*, 26-29.

[9] Liu, M. and Huang, H.J. (2018) Improvement of the Isovariance G-Q Test Method. *Statistics and Information Forum*, **33**, 3-9.

[10] Tan, X. and Deng, G.M. (2019) Improvement of the Iisovariance Parker Inspection Method. *Statistics and Information Forum*, **34**, 10-16.