Testing Method of Heteroscedasticity in Multiple Linear Regression Model

Fang Cai

School of Economics, Guangzhou College of Commerce, Guangzhou 511363, China.

Abstract

The multiple linear regression model is one of the most common models. When the model is applied to economic structure analysis, economic forecasting, policy evaluation and testing, etc., in order to play its role more accurately, it must be ensured that the model does not have heteroscedasticity. Therefore, it is also an important issue to test whether the model has heteroscedasticity. This article mainly sorts out the test methods of multiple linear regression heteroscedasticity and its applicability, in the hope that the multiple linear regression heteroscedasticity test can be selected appropriately.

Keywords

Multiple Linear Regression Model; Heteroscedasticity; Testing Method.

1. Introduction

In order to determine the quantitative form between variables, regression analysis is usually used. In classic linear regression analysis, the least squares method is the most used. To use the least squares method to fit a quantitative relationship close to reality, five assumptions need to be satisfied, including the homoscedasticity assumption. If there is heteroscedasticity, the prediction accuracy of the model will be reduced. Therefore, when fitting the model, it is necessary to detect whether there is heteroscedasticity.

2. The concept of heteroscedasticity

Heteroskedasticity means that the variance (degree of dispersion) of the observed value of the explained variable changes with the change of the explanatory variable. Assuming that there is a one-time quantitative relationship between Y and X, the overall regression function can be assumed to be $Y_i = \alpha + \beta X_i + \mu_i$ (or $E(Y/X_i) = \alpha + \beta X_i$), where $\mu_i$ refers to the random error term of the actual observation value deviating from the expected value. When X takes $X_i$, the Y value of a subject with its expected value $E(Y/X_i)$ of the axis of symmetry normal distribution,
the variance of the distribution of $D(Y/X_i)$. If $X$ takes any value, the distribution variance of $Y$ is the same, it is called homoscedasticity. On the contrary, it is called heteroscedasticity. As shown in Figure 1, the left side is homoscedasticity, and the right side is heteroscedasticity. The variance of $Y$ is equal to the variance of the random error term, so heteroscedasticity can also be expressed as $\text{Var}(\mu_i) = \sigma^2(i = 1,2,3\ldots,n)$, and homoscedasticity is expressed as $\text{Var}(\mu_i) = \sigma^2(i = 1,2,3\ldots,n)$.

Similarly, the heteroscedasticity of the multiple linear regression model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \cdots + \beta_k X_{ki} + \mu_i$ means that the degree of dispersion of $Y$ changes with the change of any one or several $X$.

3. Consequences of heteroscedasticity

3.1. The least squares estimate of parameters is unbiased but not effective

The parameter estimation formula is the estimation equation of the parameter $\hat{\beta}_i$, which is a set of parameters that is solved by the least square method so that the sum of squares of the residual $e_i$ is the smallest. If all the premises are met, the parameter estimation formula calculated by the least square method is an unbiased and effective estimation value, that is, the calculated expected value of the parameter is equal to the true parameter value and is the one with the smallest variance among all the calculation methods. If there is heteroscedasticity, it is no longer valid.

3.2. Hypothesis testing loses meaning

Because when the significance hypothesis test is performed on the parameters, the constructed statistics all use the variance estimates of the parameters. Therefore, when the variance estimates of the parameters are biased (large or small), use this value to perform statistical tests, and the conclusions obtained are unreliable, and the test loses its meaning.

3.3. Model prediction failure

Since the calculation formula of $Y$ prediction interval contains the variance estimation value of the parameter, when there is heteroscedasticity, the accuracy of the $Y$ prediction value will decrease, and the prediction function of the model will be invalid.

4. Multiple linear regression test method for heteroscedasticity

4.1. Goldfeld-Quandt test and its improvement

When the Goldfeld-Quandt test method is used to test the heteroscedasticity of the multiple linear regression model, the multiple regression model is converted to multiple univariate linear regression models, and then Goldfeld-Quandt is used to test each independent variable for heteroscedasticity. If all univariate regression models do not have heteroscedasticity, it can be approximated that the multiple linear regression model or data does not have heteroscedasticity. Conversely, if any univariate regression model has heteroscedasticity, it is considered that the original multiple linear regression model or data has heteroscedasticity.

4.1.1. Prerequisites

First, this method is used to test increasing or decreasing variance; second, this test method is only suitable for large samples; Third, in addition to the same variance assumption does not hold, other assumptions are true.

4.1.2. Inspection steps

The first step is to arrange $X_i$ from large to small; the second step is to remove the middle $c$ samples ($c$ is equal to $1/5$ to $1/4$ of the total number of samples $n$); the third step is to perform linear regression on the remaining two sets of subsamples with the same sample size, and
calculate the residual sum of squares ESS (which is $\sum e_i^2$), the fourth step is to calculate the statistic $F = \frac{\sum e_{i1}^2}{\sum e_{i2}^2}$, where $\sum e_{i1}^2$ refers to the sum of squared residuals obtained by the regression of the group of samples with the larger X value, $\sum e_{i2}^2$ refers to the residual sum of squares obtained by the regression of the group of samples with the smaller X value. The fifth step is to judge whether there is heteroscedasticity, given the significance level, check the F distribution table to obtain the critical value $F_\alpha = F_\alpha(\frac{n-c}{2}, \frac{n-c}{2} - k)$, where $n$ is the number of samples, $c$ is the number of samples removed in the second step, and $k$ is the number of parameters, if the statistic $F > F_\alpha$ calculated in the fourth step, there is heteroscedasticity, otherwise, it does not exist.

### 4.1.3. Kolmogorov-Smirnov test based on Goldfeld-Quandt

In the heteroscedasticity test of the univariate linear regression model, if there is no increasing or decreasing relationship between the variance of the error term and the independent variable, using the G-Q test method to test the regression model for heteroscedasticity may lead to wrong judgments. Moreover, in the heteroscedasticity test of the multiple linear regression model, the G-Q test is to first convert the multiple linear regression model into multiple one-variable linear regression models, and then according to the results of the G-Q test on multiple one-variable linear regression models, to judge whether there is heteroscedasticity in the multiple linear regression model. Since the multiple linear regression model is not equivalent to the simple addition of multiple univariate linear regression models, the G-Q test results of each univariate linear regression model cannot accurately and effectively judge whether the multiple linear regression model has heteroscedasticity. That is, in the heteroscedasticity test of the multiple linear regression model, the G-Q test is not an accurate and effective test method. Therefore, Xiaojin Zhang, Jianyong Niu, and Shunyong Li (2019) proposed the Kolmogorov-Smirnov test method based on Goldfeld-Quandt [1].

The specific steps are: the first step is to randomly divide the sample into two equal or approximately equal parts, and repeat $m$ times; the second step is to perform linear regression on the two parts of samples obtained by random sampling each time, find the respective residual square sums, and calculate the random sampling statistics $F = \frac{RSS_1}{RSS_2} = \frac{n}{n-2} - k \frac{n}{2} + 1 - k$ to obtain $m$ F statistics, where $n$ is the number of samples, $k$ is the number of parameters; the third step is to compare the empirical distribution formed by the sample observation value $F_1, F_2, F_3, \ldots, F_m$ with the F distribution with degrees of freedom $\frac{n}{2} - k$ and $\frac{n}{2} - k$ (or $\frac{n}{2} - k$ and $\frac{n}{2} + 1 - k$), if there is a significant difference, there is heteroscedasticity, otherwise, there is no heteroscedasticity.

### 4.2. White test and its improvement

#### 4.2.1. Prerequisites

The prerequisites of the White test are looser than the Goldfeld-Quandt test, and only a large sample is required.

#### 4.2.2. Inspection steps

Suppose the model is $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \mu_t$, the test steps are: the first step is to use ordinary least squares to fit the model, find the residual $e_{it}$, and calculate the square of the residual, $e_i^2$; the second step is to construct the auxiliary equation, $e_i^2 = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \alpha_5 X_{2t}^2 + \alpha_6 X_{3t}^2 + \alpha_7 X_{4t}^2 + \alpha_8 X_{2t} X_{3t} + \alpha_9 X_{2t} X_{4t} + \alpha_{10} X_{3t} X_{4t} + \nu_i$, where $\nu_i$ is the random error term; the third step is to calculate the $R^2$ value of the auxiliary equation, and calculate the statistic $nR^2 \sim \chi^2(p)$, where $n$ is the number of samples, and $p$ is the number of coefficients in the auxiliary equation containing the X term; the fourth step is to judge whether
there is heteroscedasticity, given the significance level \( \alpha \), check the critical value \( \chi^2_{\alpha}(p) \) of the \( \chi^2 \) distribution table, if \( nR^2 > \chi^2_{\alpha}(p) \), there is heteroscedasticity, otherwise, there is no heteroscedasticity.

4.2.3. Improvement of White test

The White test does not require prior information, and can not only test whether there is heteroscedasticity, but also which variable is causing the heteroscedasticity. However, it can be seen from the auxiliary equation of White’s test that the more explanatory variables of the original model, the more complicated the auxiliary equation is. When there are too many explanatory variables, this method is not suitable. Therefore, it was proposed to use the fitted value \( \hat{Y} \) of the multiple linear regression model.

The specific test steps are: the first step is to calculate the fitted value \( \hat{Y}_i \) and the residual square \( (e_i^2) \) of the original multiple linear regression model, the second step is to construct the auxiliary equation \( e_i^2 = \hat{\alpha}_1 + \hat{\alpha}_2 \hat{Y}_i + \hat{\alpha}_3 \hat{Y}_i^2 + v_i \), and calculate \( F \sim F(2, n - 3) \), where \( n \) is the number of \( e_i^2 \); the third step is to determine whether there is heteroscedasticity, given the significance level \( \alpha \), if \( F > F_{\alpha}(2, n - 3) \), there is heteroscedasticity, otherwise, there is no heteroscedasticity.

4.3. Glejser test and its improvement

4.3.1. Prerequisites

The Glejser test also has only one condition, that is, the sample is required to be a large sample.

4.3.2. Inspection steps

Suppose the model is \( Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \mu_i \), the first step is to use the sample data to establish a regression model, calculate the residual sequence \( e_i \), and calculate the absolute value of the residual sequence \( |e_i| \); the second step is to use \( |e_i| \) to regress each \( X_i \), since the specific function form is unknown, various function forms are needed to experiment. The normally assumed function form is \( |e_i| = \beta_0 + \beta_1 X_{hi}^h + v_i (h = \pm 1, \pm 1, \pm 2 \ldots) \), where \( v_i \) is the error term; the third step is to determine whether there is heteroscedasticity, and use the t-test method to test the significance of each of the above experimental model parameters \( \beta_j \), if the parameters of a certain model are significantly non-zero, it is considered that there is heteroscedasticity, and then the form of heteroscedasticity can be determined according to \( R^2 \) (select the function form with the largest \( R^2 \) among the functions that pass the t-test), otherwise, it is considered to have the same variance.

4.3.3. Improvement of Glejser test

Since the Glejser test does not require prior information, it can also test which explanatory variable causes which form of heteroscedasticity, which paves the way for the subsequent correction of heteroscedasticity. However, according to the functional forms of the above experiments, it can be found that various functional forms between each explanatory variable and the residual sequence \( |e_i| \) need to be tested. The process is extremely cumbersome, and the workload increases exponentially with the increase of explanatory variables. Therefore, Yi Tang and Changhuan Feng (2018) proposed to use principal component analysis to improve the Glejser test [2].

The idea is to use principal component analysis to analyze the sample explanatory variables, and extract the first principal component as a new variable, which contains as much information as possible about all explanatory variables. Then use this new variable to regress with the absolute value of the residual sequence \( |e_i| \), and test various functional forms between the new variable and \( |e_i| \); finally, use t test and \( R^2 \) to determine whether there is
heteroscedasticity and what is the form of heteroscedasticity. The specific steps are as the second and third steps of the Glejser test above.

4.4.  Park test and its improvement

4.4.1.  Prerequisites

The prerequisite for the Park test is a large sample.

4.4.2.  Inspection steps

Suppose the model is \( Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \mu_i \), the first step is to establish a regression model, calculate the residual \( e_i \), and calculate the residual square \( e_i^2 \); the second step is to construct a heteroscedasticity function form, \( \sigma_i^2 = \sigma^2 X_{ij}^{b_j} e^{v_{ij}}, \) where \( j \) is the number of \( X, b_j \) is an unknown parameter, and \( v_{ij} \) is a random variable with independent homoscedasticity. Write the function form in logarithmic form, \( \ln\sigma_i^2 = \ln \sigma^2 + b_j \ln X_{ij} + v_{ij} \), Then replace \( \sigma_i^2 \) with \( e_i^2 \) to get \( \ln e_i^2 = \ln \sigma^2 + b_j \ln X_{ij} + v_{ij} \); the third step is to calculate \( \ln e_i^2 \) and \( \ln X_{ij} \), and establish a linear regression of \( \ln e_i^2 = \ln \sigma^2 + b_j \ln X_{ij} + v_{ij} \) to calculate \( \ln \sigma^2 \) and \( b_j \); the fourth step is to judge whether there is heteroscedasticity, if \( b_j \) is significantly not 0, there is heteroscedasticity; if \( b_j \) is significantly 0, there is no heteroscedasticity; the fifth step is to perform the above steps for each \( X \), if all \( b_j \) are significantly equal to 0, then the multiple linear regression model does not have heteroscedasticity, if any \( b_j \) is significantly not 0, then the model has heteroscedasticity.

4.4.3.  Improvements to the Park test

Although Park test can not only test the existence of heteroscedasticity, but also detect the specific expression of heteroscedasticity. But it is necessary to do a regression to detect each \( X \), the process is cumbersome, second, if there is a correlation between the explanatory variables, the constructed multiple heteroscedastic models will be inaccurate. Therefore, Xin Tan and Guangming Deng (2019) proposed to use principal component analysis to improve Parker's test [3].

The specific steps are as follows: the first step is to use all explanatory variables \( X \) to perform principal component analysis to obtain \( k \) principal components, where \( k \) is the number of \( X \), and calculate the absolute value of the principal components (Because \( X \) is all positive economic data, but the principal component analysis may be negative, so the absolute value is taken); the second step is to calculate the absolute value of the principal components (Because \( X \) is all positive economic data, but the principal component analysis may be negative, so the absolute value is taken); the third step is to calculate \( \ln\sigma_i^2 \) and \( \ln X_{ij} \), and establish a linear regression of \( \ln\sigma_i^2 = \ln \sigma^2 + b_j \ln X_{ij} + v_{ij} \) to calculate \( \ln \sigma^2 \) and \( b_j \); the fourth step is to judge whether there is heteroscedasticity, if \( b_j \) is significantly not 0, there is heteroscedasticity; if \( b_j \) is significantly 0, there is no heteroscedasticity; the fifth step is to perform the above steps for each \( X \), if all \( b_j \) are significantly equal to 0, then the multiple linear regression model does not have heteroscedasticity, if any \( b_j \) is significantly not 0, then the model has heteroscedasticity.

4.5.  Breusch-Pagan test

4.5.1.  Prerequisites

The sample is required to be a large sample, because only in the case of a large sample can the residual square be able to better estimate the true error.
4.5.2. Steps of Breusch-Pagan inspection

Suppose the model is \( Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \mu_i \), the first step is to use the model to do regression, get the residual \( e_i \), and calculate the residual square \( e_i^2 \); the second step is to establish an auxiliary equation, \( e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \nu_i \), and use this equation for regression; the third step is to determine whether there is heteroscedasticity, if \( \alpha_1, \alpha_2, \alpha_3, \alpha_4 \) are all significantly zero, there is no heteroscedasticity, otherwise, if the significance is not all zero, there is no heteroscedasticity.

5. Summary

Based on the above analysis, the following conclusions are obtained. First, almost all heteroscedasticity testing methods require large samples, because almost all involve the use of residuals instead of random error terms, and only when the sample size is large, the residuals may be random A good estimate of the error term; Second, in addition to the Goldfeld-Quandt test which requires increasing or decreasing variance, other methods basically do not have this requirement; Third, Glejser test and Park test can find which variable causes heteroscedasticity, and can determine the form of heteroscedasticity, but when these two tests are used for heteroscedasticity test of multiple linear regression model, the process is more complicated, and When there is a correlation between explanatory variables, the form of heteroscedasticity may be incorrect. Therefore, the improvement of these two methods by scholars is the improvement of principal component analysis. The improvement of Glejser test is mainly to reduce the number of variables, and The improvement of Park test is mainly to make the explanatory variables independent; Fourth, both the White test and the Breusch-Pagan test can determine which variable causes the heteroscedasticity, but when there are too many explanatory variables, the auxiliary function of the White test is more complicated and not suitable.

References

