Ekonometri ve İstatistik Sayı:8 2008 33-44

İSTANBUL ÜNİVERSİTESİ
İKTİSAT FAKÜLTESİ
EKONOMETRİ VE İSTATİSTİK
DERGİSİ

AN ASYMPTOTIC TEST FOR THE DETECTION OF HETEROSKEDASTICITY

Mehmet Yüce *

Abstract

An asymptotic test for heteroskedasticity has been developed. The test does not rely on any assumption about heteroskedasticity, and introduces two alternative statistics based on the same idea. Power of these two alternative test statistics has been measured by Monte Carlo simulations. For large samples they performed fairly well, whereas for sample sizes ≤ 100, their power was influenced by the structure of the heteroskedasticity

Keywords:: Heteroskedasticity, large sample test, regression analysis, violations from the assumptions of classical linear regression model, residual analysis, asymptotic properties, Monte Carlo simulations, the power of the test
Jel Classification: C000, C100, C190

Özet

Bu makalede heteroskedastisiteye (değişen varyans) yönelik asimptotik bir test geliştirilmiştir. Test, herhangi bir heteroskedastisite varsayımına dayanmamaktadır ve aynı düşünceye dayanan iki alternatif istatistik sunmaktadır. Bu iki test istatistiğinin güçleri Monte Carlo simülasyonları ile ölçülmüştür. Büyük örneklemler için oldukça iyi performansları olamsına karşın 100'den küçük örneklem büyüklüğü için testin gücü heteroskedastisitenin yapısından etkilenmektedir.

Anahtar Kelimeler: Heteroskedastisite, büyük örneklem testi, regresyon analizi, klasik regresyon modelinin varsayımlarından sapmalar, kalıntı analizi, asimptotik özellikler, Monte Carlo simülasyonları, testin gücü

* **Adres:** *Statistician, Ph. D student of International Economics Program at Yeditepe University.*
E-Mail: *mehmetyuce1979@yahoo.com*

1. Introduction

There are several tests to detect heteroskedasticity. Some of these tests are based on an assumption on the structure of the heteroskedasticity. Goldfeld-Quandt test and Park test can be given as examples of this type (Goldfeld and Quandt, 1965:540-541 and Park, 1964:888). On the other hand, some of these heteroskedasticity tests require an estimation of an auxiliary regression model. For instance, White test is based on the significance of an auxiliary regression which involves nonconstant disturbance variance (in practice ordinary squared residuals-e_i^2) as the dependent variable and regressors with higher orders and cross-products as explanatory variables (White, 1980:821-827). Another example is Breusch-Pagan-Godfrey test (Breusch and Pagan, 1979:1288-1290). In this test, a model in which ordinary residual squares are taken as dependent variable and some (or all) of the regressors are taken as independent variable is needed to be estimated. Another test that requires estimation of an auxiliary regression is Ramsey test (Ramsey, 1969:252). In this test, a model, in which fitted values ($\hat{Y}_i$) and their higher orders are regressors and ordinary residuals are taken as dependent variable, is used to determine the heteroskedasticity. Moreover, Glejser (Glejser, 1969:316) suggested several auxiliary models addition to Park test.
There are also other tests like modified Levene test (Glass, 1966:188 and Neter et al., 1990:765) which has the same logic that of the Goldfeld-Quandt test, and Spearman correlation test based on the rank correlation between regressors and residuals (Gujarati, 1999:372). There are other tests if repetitive data is available like Bartlett test (Glaser, 1976:488) and Hartley test (see Neter et al., 1990:764).
As it can be seen above, these kinds of tests suggest an assumption on the structure of heteroskedasticity or require a kind of auxiliary regression that depends on explanation of the regressors or estimated observations on ordinary residuals. More clearly, they seek a structure which can be modeled between residuals and regressors. If there is heteroskedasticity which does not depend on the regressors or does not satisfy the assumptions, these tests can fail to catch the existence of heteroskedasticity.
In fact, in the definition of heteroskedasticity, no statistical relationship is assumed for ε_{ij} and X_j. Having an insignificant statistical relationship between the two, does not necessarily mean that the disturbance variances are all the same.
In this paper, a diagnostic test for heteroskedasticity will be introduced. This test does not suppose special assumption for heteroskedasticity and does not require estimation of an auxiliary regression model. The test, here to be developed is asymptotic and depends on the ordinary least square residuals.

2. The Underlying Idea

Suppose we have two populations whose means are equal to each other and let the latter has much greater variance than the former. Consider that we draw an observation from each of these populations independently by random sampling. Here, it is reasonable to assume that the observation which comes from the latter tends to be larger in absolute value than the observation which comes from the former. On the contrary, if we have many populations with same mean and variance and if we draw an observation from each of these populations independently by random sampling, it can be expected that the sample observations have to differ in a certain band. We can extend this idea to regression model which has homoskedastic disturbances. When the model is estimated and the residuals are calculated, we eliminate the

trend and for the rest, the error structure which is represented by residuals remains. Therefore, if homoskedasticity assumption holds, by the analogy given above, the residuals will vary within a band which graphically exhibits an approximate rectangular shape, but if homoskedasticity assumption does not hold, the situation will be different.
The graph of residuals against the independent variable(s) which is ordered in ascending manner, or against the fitted values ($\hat{Y}_i$) can be like one of the graphs depicted in Fig. 1.

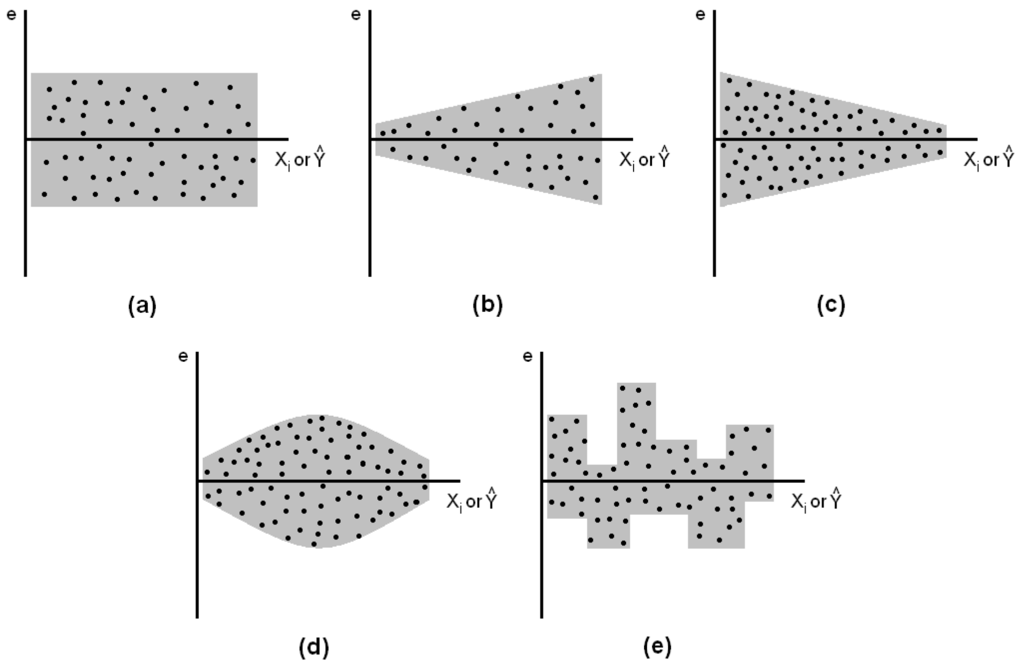

Figure 1: Some graphs of the residuals against independent variable or fitted values

Fig. 1a shows a homoskedastic residual structure. Fig. 1b and Fig. 1c shows outward-opening and inward-opening funnel patterns which are the indicators of heteroskedasticity structure that is assumed in the Goldfeld-Quandt test ($\sigma_i^2 = \sigma^2 X_i^2$). Fig. 1d depicts the elliptic shape heteroskedasticity structure. In this case, Goldfeld-Quandt test and modified Levene test which have the same logic are unable to detect the heteroskedasticity. Fig. 1e shows the irregular type of heteroskedasticity. In this case, variances of some disturbances are different from others but this does not obey to a significant rule. Many tests mentioned above become unable to catch the heteroskedasticity for this case.

3. Theory, Test Statistic and its Alternative

Homoskedasticity assumption for linear model can be shown as follows:

$$Var(\varepsilon_{ij} \mid X_j) = E(\varepsilon_{ij}^2 \mid X_j) = \sigma^2 \text{ (constant)},$$

where ε_{ij} is disturbance term, X_j is the level of independent variable and σ^2 is the disturbance variance. The same assumption can be equivalently expressed using the dependent variable Y_{ij}:

$$Var(Y_{ij} \mid X_j) = \sigma^2 \text{ (constant)}$$

The violation of this assumption-heteroskedasticity can be shown as follows:

$$Var(\varepsilon_{ij} \mid X_j) = \sigma_j^2$$

which means variances of disturbances are not equal to each other for every level of X.
Consider the model

$$Y = X\beta + \varepsilon$$

where Y is the n x 1 vector of observations, X is an n x p matrix containing p nonstochastic vectors, β is the p x 1 parameter vector and ε is the n x 1 stochastic disturbance vector which satisfies the properties

$$E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I \text{ and } \varepsilon \sim N(0, \sigma^2 I).$$

From the previous discussion the homoskedasticity hypothesis can be expressed as

H_0 : $\sigma_i^2 = c$ for $\forall$ i, where c is a constant

And because ε is unobservable for sample we can write

H_0 : $e_i^2 = c$ for $\forall$ i, where c is a constant.

Here, we idealized the homoskedasticity situation by assuming that squared residuals or absolute residuals are equal to the same amount which gives the sense that residuals are derived from a homoskedastic population, or more clearly, from subpopulations whose variances are equal to the same amount (σ^2) in accordance with the previous discussion.
It is obvious that if n is odd it is impossible for $\forall$ i, $e_i = \pm\sqrt{c}$ because of the restriction $\sum e_i = 0$, but it seems plausible, however, when $n \to \infty$, the relative difference between

absolute residuals decays to zero. Here √c analogically represents the disturbance standard deviation.
The test statistic depends on the idea that sum of squares of errors ($\sum e_i^2$) becomes greater than the expression $\left(\sum |e_i|\right)^2 / n$ when the relative difference of squared (or absolute) residuals becomes bigger which indicates the heteroskedasticity.
In an idealized manner, if all $e_i = \pm\sqrt{c}$, the two expressions become equal to each other:

$$\sum e_i^2 = nc = \left(\sum |e_i|\right)^2 / n$$

But if some squared residuals are relatively different than the other ones, $\sum e_i^2$ will be greater than $\left(\sum |e_i|\right)^2 / n$. The statistic of the test relies on the different behaviors of these two expressions on the two cases.

Theory

Proposition 1:

When the number of observations goes to infinity ($n \to \infty$) and under the null hypothesis which implies the homoskedasticity, the expected value of Ω becomes equal to σ^2, the variance of the population which is distributed as $N(\mu, \sigma^2)$, Here the Ω is

$$\frac{(n.MAE)^2 \pi}{\pi - 2 + 2n(n-p)}$$

MAE is the mean absolute error and defined as $MAE = \frac{\sum |e_i|}{n}$. Thus, under the null hypothesis and when $n \to \infty$

$$\lim_{n \to \infty} E\left[\frac{\left(\sum |e_i|\right)^2 \pi}{\pi - 2 + 2n(n-p)}\right] = \sigma^2$$

That means Ω is asymptotically unbiased estimator of σ^2.

Proof:

For the proof of above expression, we will use mean deviation. The amount of dispersion in a population is measured to some extent by totality of deviations from mean. Mean deviation of a population is defined as

$$\delta = \int_{-\infty}^{\infty} |X - \mu| dF$$

Here, F is the cumulative distribution function and μ is the mean of that population.

If X~N(0,1)

$$\delta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |X| e^{\frac{-X^2}{2}} dx = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} X e^{\frac{-X^2}{2}} dx = \sqrt{\frac{2}{\pi}}$$

For X~N(μ, σ^2),

$$\delta = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |X - \mu| e^{\frac{-(X-\mu)^2}{2\sigma^2}} dx = \sigma\sqrt{\frac{2}{\pi}} \qquad (1)$$

Let d be the sample mean deviation

$$d = \frac{1}{n}\sum_{i=1}^{n} \left| X_i - \overline{X} \right|.$$

Then for symmetrical populations, approximately

$$Var(d) \cong \frac{1}{n}\left(\sigma^2 - \delta^2\right)$$

$$Var(d) \cong \frac{\sigma^2}{n}\left(1 - \frac{2}{\pi}\right) \qquad (2) \qquad \text{(Stuart and Ord, 1994:361)}$$

In this normal case, the mean and variance of d can be obtained exactly.

Lemma 1:

Let g be a linear function of random variables. If

$$g(X_1, \ldots\ldots\ldots\ldots X_k) = \sum_{i=1}^{k} a_i X_i$$

then,

$$Var(g) = \sum a_i^2 Var(X_i) + \sum_{i \neq j} a_i a_j Cov(X_i X_j). \qquad (3)$$

Since,

$$E(d) = E(\left| X_i - \overline{X} \right|)$$

and (X_i - $\overline{X}$) has zero mean and exact variance $\frac{\sigma^2(n-1)}{n}$ by (3), (1) gives

$$E(d)=\sqrt{\frac{2\sigma^2(n-1)}{\pi.n}}$$ (4)(see Stuart and Ord, 1994:197 and 361 for the proof)

Lemma 2:

If the regression model is given as simple linear model of the form

$$Y_i=\alpha+\beta X_i+\varepsilon_i$$

and the sample regression of it as

$$Y_i=a+bX_i+e_i,$$

for the residual, it is possible to write

$$e_i=(\alpha-a)+(\beta-b)X_i+\varepsilon_i$$

e_i differ from ε_i by a term depending on the estimation error. When $n\to\infty$

$$\underset{n\to\infty}{p\lim}\, a_n=\alpha \quad \text{and} \quad \underset{n\to\infty}{p\lim}\, b_n=\beta$$

and therefore residuals tend in probability to ε_i if the assumptions of classical linear regression model hold. The empirical distribution of the e_i then tends in probability to the probability distribution of the ε_i; and the empirical moments of the e_i to the theoretical moments of the ε_i (Malinvaud, 1970:88).
By using this result and as $n\to\infty$, by (4), it is possible to write

$$E\left(\sum|e_i|\right)\cong\sqrt{\frac{2\sigma^2(n-p)}{\pi.n}} \tag{5}$$

considering $\sum e_i=0$. Here, p is the number of parameters that are estimated in the sample regression function. As $n\to\infty$, by (2)

$$Var\left(\sum|e_i|\right)\cong\frac{\sigma^2}{n}\left(1-\frac{2}{\pi}\right) \tag{6}$$

From (5) and (6), and by using the relation $Var(X_i)=E(X^2)-[E(X)]^2$

$$E\left[\sum|e_i|\right]^2\cong\frac{\sigma^2(\pi-2+2n(n-p))}{\pi}$$

Thus,

$$\lim_{n\to\infty} E\left(\frac{\left(\sum|e_i|\right)^2 \pi}{\pi - 2 + 2n(n-p)}\right) = \sigma^2$$

we reach the Proposition 1.

3.2. Test Statistic

For the test statistic we will use the result of Proposition 1. As it is known

$$E(MSE) = E\left(\frac{\sum e_i^2}{n-p}\right) = \sigma^2$$

We can compare this with the result of Proposition 1 as n converges to infinity. When the homoskedasticity assumption holds, the two expression Ω and MSE will approximately have the same value, but if ε_i is not homoskedastic, Ω and MSE will differ from each other. MSE becomes greater than Ω. Thus, we can easily construct a test by comparing Ω and MSE.
An approximate test statistic can be obtained by the way which is similar to the case where the equality of σ^2 to a specific value (σ_0^2) is tested. Here the hypotheses are constructed as

$H_0:\ \sigma^2 = \sigma_0^2$ or $H_0:\ \sigma^2 \le \sigma_0^2$

against

$H_1:\ \sigma^2 > \sigma_0^2$

If $\frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} \rangle \chi^2_{n-p,\alpha}$, the null hypothesis is rejected. In our test the specific value (σ_0^2) is Ω. Therefore, under the null hypothesis which implies homoskedasticity,

$$\frac{(n-p)MSE}{\left(\frac{n^2 MAE^2 \pi}{\pi - 2 + 2n(n-p)}\right)} = \frac{\sum e_i^2}{\left(\frac{\left(\sum|e_i|\right)^2 \pi}{\pi - 2 + 2n(n-p)}\right)} \overset{A}{\sim} \chi^2_{n-p}$$

We will denote this statistic as A.

3.3. Alternative of the Test Statistic

In this section another approach will be used to obtain a different statistic for the test.

Proposition 2:

We can test the equality of σ^2 to a specific value (σ_0^2) by the following proposition:
We have

$$E(\hat{\sigma}^2) = \sigma^2 \quad \text{and} \quad Var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p} \quad \text{where } \hat{\sigma}^2 = \text{MSE.}$$

Thus, asymptotically, we can write

$$\frac{MSE - \sigma^2}{\sqrt{\dfrac{2\sigma^4}{n-p}}} \overset{A}{\sim} t_{n-p}$$

For σ_0^2, again we have Ω. Therefore, as n→∞,

$$\frac{MSE - \dfrac{\left(\sum |e_i|\right)^2 \pi}{\pi - 2 + 2n(n-p)}}{\sqrt{\dfrac{2\left[\dfrac{\left(\sum |e_i|\right)^2 \pi}{\pi - 2 + 2n(n-p)}\right]^2}{n-p}}} \overset{A}{\sim} t_{n-p}$$

We will denote this statistic as B. So we obtained two test statistics which have the same logic.

4. Simulation Experiments

In this part, Monte Carlo simulation experiments will be carried out for the two test statistics that were deduced in the previous part. Simulation experiments were made in Matlab 7.0. The dependent variable will be obtained from the expression

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where ε_i is drawn from N(0, 1). The X_i is nonstochastic variable with n elements and contains successive numbers from 1 to n. In this simulation experiment, α and β are chosen to be as α = 0.5 and β = 1.
1000 trials are to be realized for each experiment and the sample size is chosen to be n = 50, n = 100, and n = 250.
Although it was stated before that the test developed here does not depend on any heteroskedasticity assumption, in order to measure the power of the two test statistics we have to express the structure of the heteroskedasticity. The hypothesis of the test experiment is

$$H_0 : \sigma_i^2 = \sigma^2 \qquad \text{against} \qquad H_1 : \sigma_i^2 = \sigma^2 X_i^k \tag{A}$$

Here, k is a parameter which regulates the intensity of heteroskedasticity. The case of k = 2 is depicted in Fig. 1 b and c.
To see the performance of the test, we will calculate the power of the test. That is namely, 1-Type 2 error. The results for statistics A and B for sample sizes 50, 100, and 250 is shown in Table 1 for the significance level α = 0.05.

Sample size	The statistic	A	B	The intensity parameter k
50		0.238	0.152	2
50		0.654	0.530	3
50		0.877	0.798	4
100		0.635	0.494	2
100		0.989	0.960	3
100		1	0.998	4
250		0.997	0.989	2
250		1	1	3
250		1	1	4

Table 1: The power of test statistics A and B for k = 2, 3 and 4. The sample sizes are taken as 50, 100 and 250 where α = 0.05

The results show that the power of these two test statistics becomes higher when the sample size increases. In fact, the two statistics start to behave well when the sample size is 100 and k = 3. After this the two statistics nearly have the same power.
The same simulation can be made for two- and three-independent variable linear regression model. For the two-independent variable model, the parameters are taken as α = 0.5, β_1 = 1 and β_2 = 1.5. X_1 is again a nonstochastic variable with n elements and contains successive numbers from 1 to n. X_2 is the second independent variable which contains random numbers obtained from uniform distribution varying between 1 to n. For the alternative hypothesis we use a generalized form of (A). That is,

$$H_0 : \sigma_i^2 = \sigma^2 \qquad \text{against} \qquad H_1 : \sigma_i^2 = \sigma^2 \prod_{i=1}^{m} X_i^k$$

Here, k is again the intensity parameter of heteroskedasticity and m=2.

The results for statistics A and B for sample sizes 50, 100, and 250 is shown in Table 2 for the significance level $\alpha = 0.05$.

Sample size	The statistic	A	B	The intensity parameter k
50		0.825	0.758	2
50		0.93	0.89	3
50		0.976	0.954	4
100		1	1	2
100		1	1	3
100		1	1	4
250		1	1	2
250		1	1	3
250		1	1	4

Table 2: The power of test statistics A and B for the two-independent variable linear model.

The same experiment can be extended to three-independent variable linear model. In this case, the parameters are taken as $\alpha = 0.5$, $\beta_1 = 1$, $\beta_2 = 1.5$ and $\beta_3 = 2$. The same alternative hypothesis is used where m=3. The results are shown in Table 3.

Sample size	The statistic	A	B	The intensity parameter k
50		0.99	0.98	2
50		0.989	0.979	3
50		0.994	0.99	4
100		0.999	0.997	2
100		1	0.999	3
100		1	1	4
250		1	1	2
250		1	1	3
250		1	1	4

Table 3: The power of test statistics A and B for three-independent variable linear model.

As it is seen from these three experiments, the power of A is slightly better than B. The intensity parameter and the sample size are positively related with the power.

5. Final Remarks

In this paper an asymptotic test for heteroskedasticity has been presented. As it is known, there exists a considerable amount of tests in the literature. The advantage of the test here deduced is that it does not require an estimation of an auxiliary regression or assume any structure of heteroskedasticity. The power of the two statistics depends on the intensity of the heteroskedasticity as it can be seen from the results. The performance of the two alternative statistics becomes well as the sample size increases.

REFERENCES:

Breusch, T. S. and Pagan, A. R., (1979), A simple test for heteroskedasticity and random coefficient variation, *Econometrica*, Vol. **47**, No. 5, 1288-1290.

Gujarati, D. N., (1999), *Temel Ekonometri*, Çevirenler: Ümit Şenesen ve G. G. Şenesen. Literatür Yay. İstanbul, 372.

Glaser, R. E., (1976), Exact critical values for Bartlett's test for homogeneity of variances, *Journal of American Statistical Association*, Vol. **71**, no. 354, 488.

Glass, G. V., (1966), Testing homogeneity of variances, *American Educational Research Journal*, Vol. **3**, No. 3, 188.

Glejser, H., (1969), A new test for heteroskedasticity, *Journal of American Statistical Association*, Vol. **64**, No. 325, 316.

Goldfeld, S. M. and Quandt, R. E., (1965), Some tests for homoskedasticity, *Journal of American Statistical Association*, Vol. **60**, No. 310, 540-541.

Malinvaud, E., (1976), *Statistical Methods for Econometrics*, 2nd edition, North-Holland, Amsterdam, 88.

Neter, J., Kutner, M. H., Nachtsheim, C. J and Wasserman, W., (1996), *Applied linear statistical models*. 4th edition, McGraw-Hill, Boston, 764-765.

Park, R. E., (1964), Estimation with heteroscedastic error terms, *Econometrica*, c.34, **4**, 888.

Ramsey, J. B., (1969), Tests for specification error in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society*, **B31**, 2, 252.

Stuart, A. and Ord, J. K., (1994), *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. 6th edition, John Wiley & Sons Inc. New York, 197 and 361.

White, H., (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, Vol. **48**, No. 4, 821-827.