Decision Making: Applications in Management and Engineering
Vol. 4, Issue 1, 2021, pp. 1-18.
ISSN: 2560-6018
eISSN: 2620-0104
DOI: https://doi.org/10.31181/dmame2104001d

VIOLATION OF THE ASSUMPTION OF HOMOSCEDASTICITY AND DETECTION OF HETEROSCEDASTICITY

Irena Djalic[1*] and Svetlana Terzic[1]

[1] University of East Sarajevo, Faculty of Transport and Traffic Engineering, Doboj, Republic of Srpska, Bosnia and Herzegovina

Received: 1 September 2020;
Accepted: 20 October 2020;
Available online: 24 October 2020.

Original scientific paper

Abstract: *In this paper, it is assumed that there is a violation of homoskedasticity in a certain classical linear regression model, and we have checked this with certain methods. Model refers to the dependence of savings on income. Proof of the hypothesis was performed by data simulation. The aim of this paper is to develop a methodology for testing a certain model for the presence of heteroskedasticity. We used the graphical method in combination with 4 tests (Goldfeld-Quantum, Glejser, White and Breusch-Pagan). The methodology that was used in this paper showed that the assumption of homoskedasticity was violated and it showed existence of heteroskedasticity.*

Key words: *Economic phenomena; heteroskedasticity; homoskedasticity; random errors.*

1. Introduction

Econometrics is a discipline that determines the connection between economic phenomena and confirms or does not confirm economic theory, starting from mathematical equations and forming econometric models suitable for testing. Regression analysis is one of the most commonly used tool in econometrics to describe the relationships between economic phenomena. One of the classic assumptions of linear regression is homoskedasticity. Homoskedasticity implies that the variance of random error is constant and equal for all observations. When the random errors of the classical linear regression model are not homoskedastic, then they are heteroskedastic (Mladenović & Petrović, 2017).

The main goal of the paper is to show how the linear regression model behaves in conditions of violating the assumption of homoskedasticity and how this violation is detected. The basic contribution of the paper is that in one place it gives a developed method of detecting violating of homoskedasticity, ie the existence of

* Corresponding author.
E-mail addresses: i.naric@yahoo.com (I. Djalic), terzic_svetlana@yahoo.com (S. Terzic)

homoskedasticity in linear regression models. This paper presents a methodology for detecting heteroskedasticity in linear regression models by a combination of a graphical method and four tests.

After the introduction, a review of the literature was performed, after which the basics of heteroskedasticity were presented. In this part of the paper, the Goldfeld-Quantum, Glejser, White and Breusch-Pagan tests are presented. At the end of the paper, concluding remarks were made and recommendations for further research were given.

2. Literature review

Aue et al. (2017) state that heteroskedasticity is a common characteristic of financial time series and most often refers to the process of model development using autoregressive conditional heteroskedastic and generalized autoregressive conditional heteroskedastic processes. Ferman and Pinto (2019) formed a model of inference that works with adjusting differences in differences with several treated and many controlled groups in the presence of heteroskedasticity. Charpentier et al. (2019) developed the Gini-White test, which shows greater strength in solving the problem of heteroskedasticity than the ordinary White test in cases when external observations affect the data. Moussa (2019) analyzes cases in which heteroskedasticity is the result of individual effects or idiosyncratic errors, or both. Linton and Xiao (2019) study the effective estimation of nonparametric regression in the presence of heteroskedasticity and conclude that in many popular nonparametric regression models their method has a lower asymptotic variance than the usual unweighted procedures. A large number of authors pay attention to heteroskedasticity and develop models for solving certain problems (Baum & Schaffer, 2019; Brüggemann et al., 2016; Lütkepohl & Netšunajev, 2017; Cattaneo et al., 2018; Ou et al., 2016; Sato & Matsuda, 2017). Taşpınar et al. (2019) investigate the properties of finite samples of the heteroskedasticity-robust generalized method of moments estimator (RGMME), ie develop a robust spatial econometric model with an unknown form of heteroskedasticity. Crudu et al. (2017) propose a new inference procedures for models of instrumental variables in the presence of many, potentially weak instruments that are robust to the presence of heteroskedasticity. Lütkepohl and Velinov (2016) compare models of long-term restriction that are widely used to identify structural shocks in vector autoregressive (VAR) analysis based on heteroskedasticity. Harris and Kew (2017) test adaptive hypotheses for a fractional differential parameter in a parametric ARFIMA model with unconditional heteroskedasticity of unknown shape. In the case of heteroskedasticity, there are occasionally precise theoretical reasons for assuming that the errors have different variances for different values of the independent variable. Very often, arguments for the presence of heteroskedasticity are so well defined, and sometimes there is a vague suspicion that the assumption of homoskedasticity is too strong (Barreto & Howland, 2006). It is important to note that heteroskedasticity is a common occurrence in spatial samples due to the nature of collection of data. Obvious sources of heteroskedasticity are associated with different dimensions for different regions in the study area, unequal concentrations of population and economic activity in rural and urban areas (Arbia, 2006). Baum and Schaffer (2019) provide advice and guidance to researchers who wish to use tests to check heteroskedasticity.

3. Methodology

The simplest form of linear regression, which shows a linear relationship between two phenomena, is a simple linear regression:

$$Y = \alpha + \beta X + \varepsilon \tag{1}$$

ε is a random error that we make during linear regression, and α and β are unknown parameters. To estimate the unknown parameters, we use a sample. For fixed n values of the independent variable X the values of the variable Y are determined. In this way, n pairs (X_1, Y_1), (X_2, Y_2), ... , (X_n, Y_n) are obtained, which forming the model of the simple linear regression sample:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1,2, ..., n \tag{2}$$

The assumption of homoskedasticity for the random variable ε_i is:

$$v(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 = const., \quad \text{for each i = 1,2, ..., n} \tag{3}$$

When this assumption is violated, that is, when the random errors of the classical linear regression model do not satisfy this characteristic, then they are heteroskedastic.

If the assumption of homoskedasticity (Jovičić, 2011):

$$v(\varepsilon_i) = E\left[\varepsilon_i - E(\varepsilon_i)\right]^2 = E(\varepsilon_i^2) = \sigma^2, \qquad \text{for each } i \tag{4}$$

is not met, but the variances are different and valid:

$$v(\varepsilon_i) = \sigma_i^2 \qquad i = 1, \ldots, n, \tag{5}$$

respectively (Mladenović & Petrović, 2017),

$$v(\varepsilon_1) = \sigma_1^2, v(\varepsilon_2) = \sigma_2^2, ..., v(\varepsilon_n) = \sigma_n^2, \qquad \sigma_1^2 \neq \sigma_2^2 \neq ... \neq \sigma_n^2 \tag{6}$$

it can be said that the errors are heteroskedastic or there is heteroskedasticity in the model.

Figure 1 presents a model where heteroskedasticity of the error is assumed. The growth of savings with increasing income is shown, where the variance of savings is smaller with different income levels. The variance is not constant, but increases with the growth of income, which corresponds to real economic relations (Mladenović & Petrović, 2017).

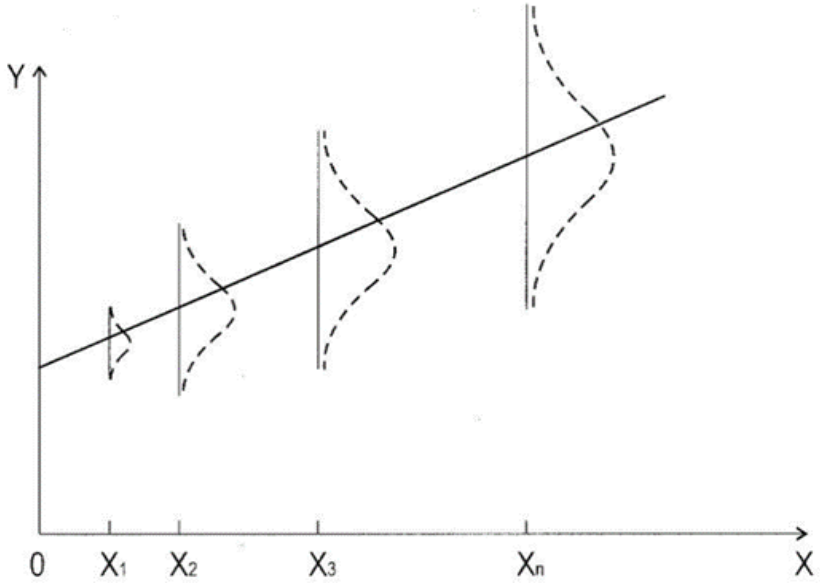

Figure 1. Heteroskedastic errors

Source: Mladenović & Petrović (2017)

Heteroskedasticity can also be caused by errors of specification . For example, by omitting an important regressor whose influence will be covered by the error, a different variation of the error for different observations can be obtained. Similarly, the wrong functional form of the model can lead to heteroskedasticity of the error. As data collection techniques are advancing, which implies the provision of representative samples for statistical processing, so do errors and thus their dispersions are decreasing. And this may be another reason for the occurrence of heteroskedasticity.

3.1. Consequences of heteroskedasticity

The presence of heteroskedasticity in the model of dependence of savings on income can be represented on the basis of the following point scatter diagram (Figure 2):

Violation of the assumption of homoscedasticity and detection of heteroscedasticity

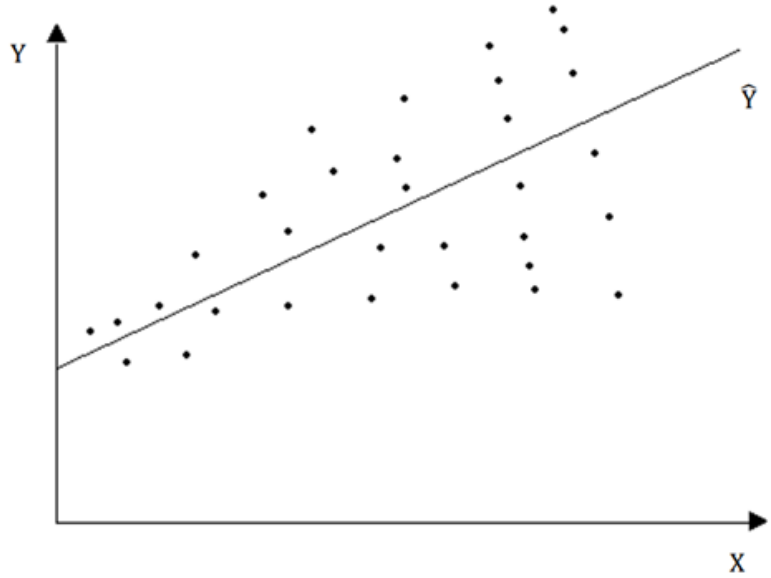

Figure 2. Diagram of distribution of points (Mladenović & Petrović, 2017)

Estimates of unknown parameters using the ordinary least squares method are determined from the condition that the residual sum of squares, $\sum e_i^2$, is minimal. In that case, all squares of the residuals have the same weight, ie they give the same information when forming the necessary estimates. This condition is not precise enough for the sample presented in Figure 2. Data that are far from the sampling regression line provide less useful information about its position than those that are closer to it. Higher residual values in absolute terms correspond to more distant data. These residues dominate in the total residual sum of squares. Therefore, it is realistic to expect that the application of ordinary least squares method does not provide estimates with desirable statistical properties.

Suppose that in the model (Mladenović & Petrović, 2017):

$$Y_i = \beta_0 + \beta X_i + \varepsilon_i, \tag{7}$$

there is heteroskedasticity:

$$v(\varepsilon_i) = \sigma_i^2, \quad i = 1, 2, \ldots, n \tag{8}$$

The estimate b of the parameter β, obtained using the ordinary least squares method, is unbiased, because the corresponding proof does not use the assumption of the stability of the variance of the random error.

To determine the variance of the estimate b we start from the expression:

$$b - \beta = \sum_{i=1}^{n} w_i \varepsilon_i \ , \tag{9}$$

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2} \ , \tag{10}$$

based on which the variance is:

$$\begin{aligned} v(b) &= E\left[b - E(b)\right]^2 \\ &= E(b-\beta)^2 \\ &= E\left[\sum_{i=1}^{n} w_i \varepsilon_i\right]^2 \\ &= E\left[w_1^2\varepsilon_1^2 + w_2^2\varepsilon_2^2 + \ldots + w_n^2\varepsilon_n^2 + 2w_1w_2\,\varepsilon_1\varepsilon_2 + 2w_1w_3\varepsilon_1\varepsilon_3 + \ldots\right] \\ &= w_1^2 E\left(\varepsilon_1^2\right) + w_2^2 E\left(\varepsilon_2^2\right) + \ldots + w_n^2 E\left(\varepsilon_n^2\right) + 2w_1w_2 E\left(\varepsilon_1\varepsilon_2\right) + 2w_1w_3 E\left(\varepsilon_1\varepsilon_3\right) + \ldots \end{aligned} \tag{11}$$

In the Eq. (11), all elements of the form $E\left(\varepsilon_i\varepsilon_j\right)$, $i \neq j$ are equal to zero.

The expression for the variance of the estimate *b* is:

$$v(b) = w_1^2\,E\left(\varepsilon_1^2\right) + w_2^2 E\left(\varepsilon_2^2\right) + \ldots + w_n^2 E\left(\varepsilon_n^2\right)$$

$$= \sigma_1^2 w_1^2 + \sigma_2^2 w_2^2 + \ldots + \sigma_n^2 w_n^2 .$$

$$= \sum_{i=1}^{n} w_i^2 \sigma_i^2$$

$$= \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \sigma_i^2 = \frac{\sum_{i=1}^{n} x_i^2 \sigma_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \tag{12}$$

The variance of the estimate *b,* in a simple linear regression model, is given by the following expression:

$$v(b) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \tag{13}$$

When the existence of heteroskedasticity is neglected, the estimate of the variance of the estimate *b* is obtained by the following formula:

$$s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \frac{1}{\sum_{i=1}^{n} x_i^2} \tag{14}$$

When the variance of the random error grows in parallel with the explanatory variable then the estimate s_b^2 underestimates the actual variance of the estimate *b*. This arises because the estimate of the random error variance, s^2, underestimates the actual random error variance of the initial model.

Thus, the properties of the estimates of parameter obtained by applying the ordinary least squares method in the presence of heteroskedasticity are:

1. The ratings are unbiased,
2. Estimates do not have minimal variance, that is, they are ineffective.

3. The assessment of the variance of a random error underestimates, in most cases, the actual variance. Therefore, the estimate of the variance of the estimate of slope, , also underestimates the variance .
4. Confidence intervals and tests based on the assessment of the variance of a random error are unreliable.

3.2. Testing of heteroskedasticity

The true nature of heteroskedasticity is usually unknown, so the choice of the appropriate test depends on the nature of the data. But as the amount of error variation around the mean value typically depends on the height of the independent variables, all tests rely on examining whether the error variance is some function of the regressor. Certain methods for testing the existence of heteroskedasticity are presented below.

3.2.1. Graphic method

One of the simplest methods for examining the existence of heteroskedasticity consists in visually viewing the residuals of the estimated model. It is common to form point scattering diagram of residual e_i or their absolute value, $|e_i|$, and independent variable x_i. Since the variance of a random error $E\left(\varepsilon_i^2\right)$, there is an opinion that on the point scattering diagram of residual values should be replaced by their square, e_i^2 .

Based on the point scatter diagrams, we can conclude about whether heteroskedasticity exists, and if so, in what form it occurs, ie how the variance of random error is generated. Figure 3 presents graphs of some of the possible point scatter diagrams (Mladenović, 2011).

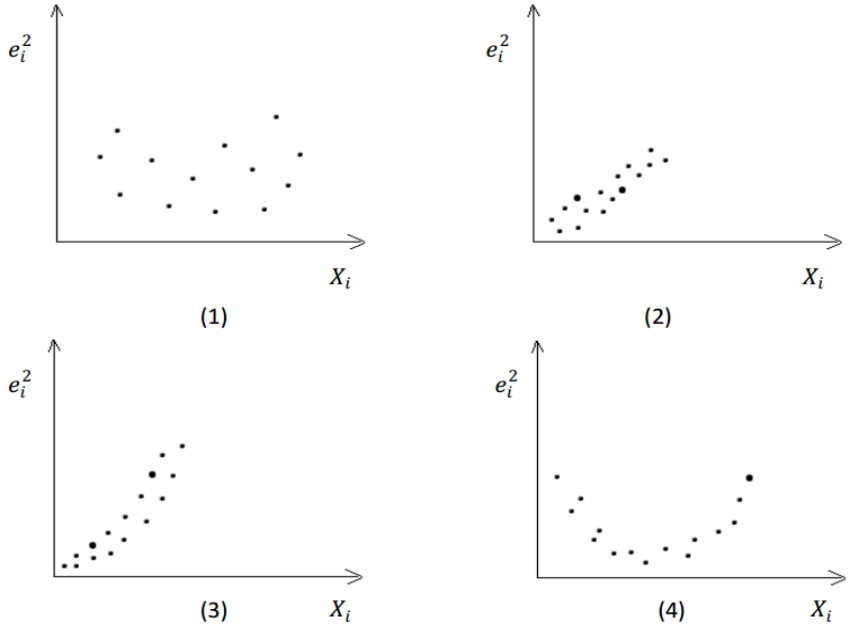

Figure 3. Point scatter diagrams (Mladenović, 2011)

The first graph corresponds to a model in which there is no systematic dependence between the variances of random errors and the independent variable x_i. In such a model, random errors can be considered homoskedastic. Other graphs show the regularity in the position of the points on the scatter diagram, suggesting possible heteroskedasticity. The second graph indicates a linear dependence, while the third and fourth graphs represent the dependence expressed in square form, in the sense that the variance of the random error is correlated with x_i^2.

Graphic methods are only a means of preliminary analysis. In order to get a more precise answer to the question of whether heteroskedasticity is present or not, it is necessary to use appropriate tests.

3.2.2. Goldfeld-Quandt test

One of the earliest, which is very simple and often used is the Goldfeld-Quandt test (Kalina & Peštová, 2017). This test tests the null hypothesis of random error constancy versus alternative that the variance of a random error is a linear function of the independent variable. It is assumed that the random error is non-autocorrelated and with normal distribution. The test procedure is as follows (Mladenović & Petrović, 2017):

- Observations from the sample are arranged according to the increasing values of the independent variable.
- From the set of *n* observations, *c* central observations are omitted, so that further analysis is based on two sets of observations: the first $\frac{n-c}{2}$ and the last $\frac{n-c}{2}$ observations where is necessary to ensure that $\frac{n-c}{2} > k$, and *k* is the number of evaluated parameters.
- We individually evaluate the two regressions based on the first $\frac{n-c}{2}$ and the last $\frac{n-c}{2}$ observations. The obtained sums of the residual squares are denoted by $\sum e_1^2$ and $\sum e_2^2 \sum e_1^2 \sum e_1^2$ ($\sum e_1^2$ corresponds to the regression with the lower, and $\sum e_2^2 \sum e_1^2 \sum e_1^2$ to the regression with the higher values of the independent variable).

The homoskedasticity of a random error implies the same degree of variation in two subsets of observations, which is manifested by approximately the same values of the variable sums $\sum e_1^2$ and $\sum e_2^2 \sum e_1^2 \sum e_1^2$. In this case, the quotient of these two sums is close to the value of 1. On the contrary, the existence of heteroskedasticity results in a higher value of the residual sum $\sum e_2^2 \sum e_1^2 \sum e_1^2$. The purpose of the test is to check whether $\frac{\sum e_2^2}{\sum e_1^2}$ is statistically significantly different from 1. Assuming that the null hypothesis of constant variance is correct, the following holds:

$$\frac{\sum e_1^2}{\sigma^2} : x^2_{\frac{n-c-2k}{2}} \tag{15}$$

$$\frac{\sum e_2^2}{\sigma^2} : x^2_{\frac{n-c-2k}{2}} \tag{16}$$

where the *k* is the number of parameters for evaluation in the known model. It follows that the observed relationship:

$$\frac{\sum e_2^2}{\sum e_1^2}$$

has an *F* – distribution with $\frac{n-c-2k}{2}$ and $\frac{n-c-2k}{2}$ degrees of freedom.

Therefore, the Goldfeld-Quandt test statistic is in a form:

$$F = \frac{\sum e_2^2}{\sum e_1^2} \tag{17}$$

If the calculated value of F – statistics is higher than the corresponding critical value at a given level of significance, we conclude that there is heteroskedasticity in the model.

3.2.3. Glejser test

The application of this test does not require a priori knowledge of the nature of heteroskedasticity, but it is reached during the testing. The test procedure is as follows (Im, 2000):

- The initial regression $Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i$ is estimated by the method of ordinary least squares and the residuals e_i are calculated .
- The next regression is estimated:

$$|e_i| = \delta_0 + \delta_1 X_i^h + \text{error} \tag{18}$$

The values 1, −1, and 1/2 are usually assigned to the parameter h so that regressions are evaluated:

$$|e_i| = \delta_0 + \delta_1 X_i + \text{error} \tag{19}$$

$$|e_i| = \delta_0 + \delta_1 / X_i + \text{error} \tag{20}$$

$$|e_i| = \delta_0 + \delta_1 \sqrt{X_i} + \text{error} \tag{21}$$

- The statistical significance of the evaluation of the parameter δ_1 is tested using the *t*-test.
- The coefficients of determination obtained for different values of the parameter *h* are compared.

The statistical significance of the estimate δ_1 leads to the conclusion that there is heteroskedasticity. The very character of heteroskedasticity is determined according to the regression with the highest value of the coefficient of determination.

3.2.4. White test

The test is based on the comparison of the variance of the estimators obtained by the method of ordinary least squares in the conditions of homoskedasticity and heteroskedasticity. If the null hypothesis is correct, the two estimated variances would differ only due to fluctuations in the sample. The null hypothesis about the homoskedasticity of a random error is tested against the widely placed alternative hypothesis that the variance of a random error depends on the explained variables, their squares and intermediates, ie. the variation of the residuals under the combined action of the regressors is examined.

The White test consists of the following steps (White, 1980):

Step 1: The model $Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i$ should be estimated to obtain a series of residuals e_i ie their squared values.

Step 2: Evaluate the auxiliary regression in which the squares of the residuals of the function of all regressors of the model, their squares and intermediate products, ie apply the method of ordinary least squares on $e_i^2 = \delta_0 + \delta_1 Z_{i1} + \delta_2 Z_{i2} + \ldots + \delta_p Z_{ip} + error$, i = 1, 2, ..., n, where for simple regression $p = 2, Z_{i1} = X_i$ *and* $Z_{i2} = X_i^2$ so the test is based on analysis of model $e_i^2 = \delta_0 + \delta_1 X_i + \delta_2 X_i^2 + error$, i = 1, 2, ..., n. The significant influence of the independent variables X_i and X_i^2 at e_i^2 results in a high value of the coefficient of determination R^2.

The significant influence of the independent variables X_{i1} and X_{i2}, the specification is $p = 5$, $Z_{i1} = X_{i1}$, $Z_{i2} = X_{i2}$, $Z_{i3} = X_{i1} X_{i2}$, $Z_{i4} = X_{i1}^2$ and $Z_{i5} = X_{i2}^2$. Due to the possible large loss of degrees of freedom, it is possible to use instead of individual values of the regressors, their linear combination: Y_i, Y_i^2 .

Step 3: Based on the value of the coefficient of determination from the auxiliary regression, R_W^2, form the White test nR_W^2, where n is the sample volume. Asymptotically, under the null hypothesis of homoskedasticity, the test statistic nR_W^2 leads to χ^2 distribution with the number of degrees of freedom equal to the number of regressors in the auxiliary regression: $nR_W^2 \sim X_p^2$.

Step 4: If the calculated value of the test statistics is greater than the tabular value, ie if the coefficient of determination in the auxiliary function of the residual square is high enough, the homoskedasticity hypothesis is rejected.

The White test is not sensitive to the deviation of errors from normal and it is simpler, so it is more often used to test the existence of heteroskedasticity. In the case that there are multiple regressors, the introduction of squares and all intermediates in the auxiliary regression can mean a large loss in the number of degrees of freedom. That is why the White test is often performed without intermediates.

3.2.5. Breusch-Pagan test

This test is based on the idea that the estimates of the regression coefficients obtained by the least squares method should not differ significantly from the maximally plausible estimates, if the homoskedasticity hypothesis is true (Halunga et al., 2017). The null hypothesis about the homoskedasticity of random error is tested against the broadly set alternative hypothesis about the influence of a number of

Violation of the assumption of homoscedasticity and detection of heteroscedasticity

factors on the variance of random error. For simplicity, assume that test examines the influence of the explanatory variable X_i in simple regression. The testing procedure is as follows (Mladenović & Nojković, 2017):

Residuals e_i are formed from the regression Y_i at a constant and X_i.

The average value of the sum of the squares of the residual is determined: $sp^2 = \frac{\sum e^2}{n}$, and then forms a new variable: $G_i = \frac{e_1^2}{sp^2}$, $i = 1, 2, \ldots, n$.

From regression G_i at X_i the explained sum of the squares of the dependent variable is obtained $\left(\sum \hat{g}_i^2\right)$.

The Relationship $\frac{\sum \hat{g}_i^2}{2}$ has X^2 distribution with one degree of freedom.

The heteroskedasticity hypothesis will be accepted when the value of the calculated ratio $\frac{\sum \hat{g}_i^2}{2}$ is greater than the critical value of X^2 distribution with one degree of freedom.

4. Application of the model: Data simulation

Table 1 shows the data so as to simulate the next deviation $\sigma_i^2 = 0.01X_i^2$. The population straight line is $Y = 2 + 3X$, where Y is savings and X is income. In line Y_i, $i = 1, \ldots, 30$, there are values Y to which errors ε_i have been added.

Table 1. Display of simulated data

No.	X_i	Y	ε_i	Y_i
1	10	32	-0.13677	31.86323
2	10	32	1.045263	33.04526
3	10	32	0.324248	32.32425
4	10	32	-1.80589	30.19411
5	10	32	0.568473	32.56847
6	10	32	-0.17024	31.82976
7	10	32	0.676169	32.67617
8	10	32	-0.57257	31.42743
9	10	32	-1.53944	30.46056
10	10	32	-0.38377	31.61623
11	20	62	-4.85783	57.14217
12	20	62	-1.66701	60.33299
13	20	62	9.513881	71.51388
14	20	62	0.817791	62.81779
15	20	62	-11.1762	50.82381
16	20	62	-6.47024	55.52976

17	20	62	9.51661	71.51661
18	20	62	2.045394	64.04539
19	20	62	5.286107	67.28611
20	20	62	7.451416	69.45142
21	30	92	19.59112	111.5911
22	30	92	33.2486	125.2486
23	30	92	-23.2211	68.7789
24	30	92	-28.8606	63.13944
25	30	92	38.95497	130.955
26	30	92	-1.97921	90.02079
27	30	92	-36.9439	55.05615
28	30	92	12.09004	104.09
29	30	92	41.06767	133.0677
30	30	92	-38.0374	53.96263

Based on the values Y_i and X_i from Table 1 evaluate the linear regression model is evaluated:

$$Y_i = \alpha + \beta X_i + \varepsilon_{i,}$$
$$\varepsilon_i : N(0, 0.01X_i^2),$$
$$i = 1, 2, \ldots, 30$$

After evaluation, the following results were obtained (Table 2):

Table 2. Coefficients

Model	Estimated value	Standard error	p - value
$\widehat{\boldsymbol{\alpha}}$	1.022	8.880	0.909
$\widehat{\boldsymbol{\beta}}$	3.090	0.411	0.000

After the obtained coefficients, the analysis of model variance was performed (Table 3):

Table 3. Analysis of variance

	Sum of quares	No. of degrees freedom	Average value of Sum	p - value
REGRESSIONAL	19090.319	1	19090.319	0.000
RESIDUAL	9462.096	28	337.932	
TOTAL	28552.414	29		

The coefficient of determination was determined, R^2 = 0.669.
Figure 4 graphically shows the simulation model.

Violation of the assumption of homoscedasticity and detection of heteroscedasticity

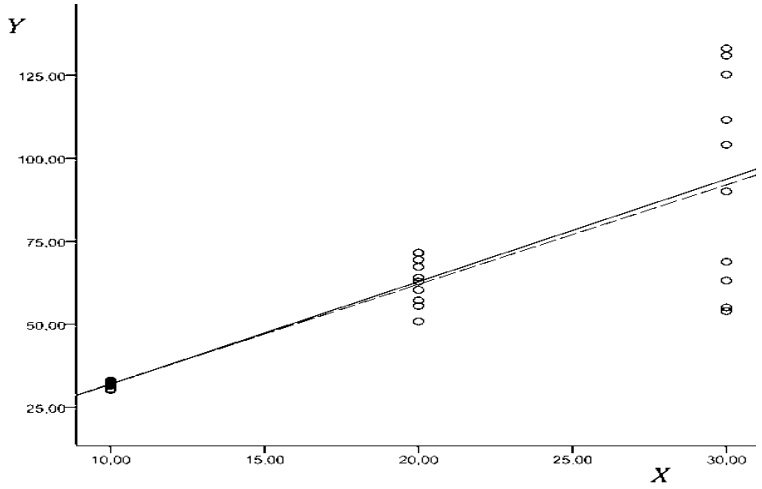

Figure 4. Graphic representation of the population model

Figure 4 shows the population line $Y = 2 + 3X$ by interrupted line, while the sample line $\hat{Y} = 1.022 + 3.090X$ is shown by full line. The graph clearly shows that the scatterings are higher for higher values of the independent variable X and that sample line $\hat{Y}$ slightly deviates from the line Y. After the graphical representation of the model, it can be assumed that certain deviations exist, so we will test the heteroskedasticity with the previously described tests.

4.1. Graphic method

Figure 5 in graph (a) clearly shows the relationship between the residuals and the independent variable X (the larger X, the larger residuals), while in diagram (b) the dependence of the squared residuals with respect to X can be seen (the dependence in the square form).

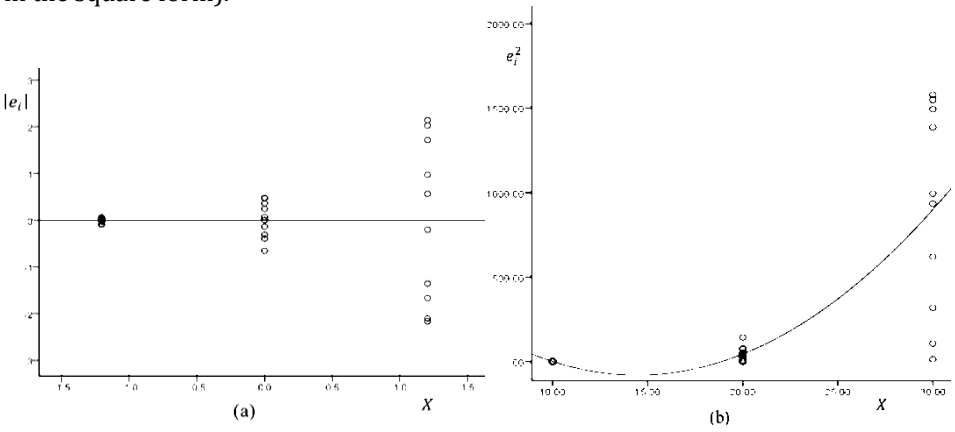

Figure 5. Diagrams of Residual Scattering

4.2. Goldfeld-Quandt test

After the order of observations in ascending order of magnitude X two models (for the first 15 and last 15 observations) of linear regression $Y_i = \alpha + \beta X_i + \varepsilon_i$ are evaluated.

The first 15 observations:

$$\hat{Y}_i = 3.075 + 2.873 X_i \qquad R^2 = 0.92 .$$

(3.324) (0.235)

The last 15 observations:

$$\hat{Y}_i = 9.516 + 2.803 X_i \qquad R^2 = 0.222 .$$

(39.369) (1.454)

The residual sum for the first 15 observations is 18.411, and for the last 15 observations it is 704,495. Based on these residuals, the value of the test statistic is:

$$F = \frac{704.495}{18.411} = 38.26 .$$

As the critical value of the F - distribution with 13 and 13 degrees of freedom and a significance level of 0.05 is 2.58, this test shows that heteroskedasticity is present (the value of the test statistics is higher than the critical value).

4.3. Glejser test

Three linear regression models are being tested:

Model 1: $|e_i| = \alpha + \beta X_i + error$

Model 2: $|e_i| = \alpha - \beta / X_i + error$

Model 3: $|e_i| = \alpha - \beta \sqrt{X_i} + error$

The results are shown in Table 5.

Table 5. The results of the Glejser test

	estimated PARAMETERS	ESTIMATED VALUES	standard error	p -value	R^2
MODEL 1	$\hat{\alpha}$	-15.419	4.169	0.001	0.631
	$\hat{\beta}$	1.335	0.193	0.000	
MODEL 2	$\hat{\alpha}$	31.510	4.500	0.000	0.467
	$\hat{\beta}$	-331.137	66.813	0.000	
MODEL 3	$\hat{\alpha}$	-37.469	7.804	0.000	0.593
	$\hat{\beta}$	11.151	1.745	0.000	

The estimated parameters that stand next to the regressors are statistically significant. All parameters are suitable for testing the hypothesis of heteroskedasticity, and based on the coefficient of determination, the first is preferred because it is the largest. This test also shows the presence of heteroskedasticity.

4.4. White test

Auxiliary linear regression was estimated:

$$e_i^2 = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

and the values are shown in the following Table 4:

Table 4. Coefficients

Model	estimated value	standard error	p - value
$\widehat{\boldsymbol{\beta}}_0$	766.923	484.031	0.125
$\widehat{\boldsymbol{\beta}}_1$	-117.142	54.964	0.042
$\widehat{\boldsymbol{\beta}}_2$	4.053	1.360	0.006

where the coefficient of determination is $R_W^2 = 0.607$. It can be observed that the coefficients along with X_i and X_i^2 are statistically significant while the constant is not. White's test statistic is $nR_W^2 = 30 \cdot 0.607 = 18.21$ which is greater than the tabular value of the χ^2 istribution with two degrees of freedom, 5.991. It is the same conclusion as before, that heteroskedasticity is present.

4.5. Breusch-Pagan test

Based on the linear regression equation $Y_i = 1.022 + 3.090 X_i$ the estimated value of the error variance is obtained:

$$\hat{\sigma}^2 = \frac{9462.10}{30} = 315.403$$

The new regression equation:

$$\hat{p}^2 = -1.852 + 0.143 X_i$$

(0.609) (0.028)

where is:

$$p_i = \frac{e_i^2}{\hat{\sigma}^2} = \frac{e_i^2}{315.403}$$

Test statistics is:

$$\frac{\sum \hat{g}_i^2}{2} = \frac{40.660}{2} = 20.33$$

The critical value of the χ^2 distribution with one degree of freedom and a significance level of 0.05, is 3.841, so it is also concluded that heteroskedasticity is present.

4.6. Discussion

After testing, it is clear that all four tests show the presence of heteroskedasticity in a given model. The Goldfeld-Quandt test shows that the F – distribution is equal to 2.58 and it is higher than the corresponding critical value at a given level of significance (0.05). Based on this we can conclude that heteroskedasticity is present in the model. In the Glejser test the parameter δ_1 is tested and the coefficients of determination obtained for different values of the parameter *h* are compared. In this model (Table 5) all parameters are suitable for testing the hypothesis of heteroskedasticity, and based on the coefficient of determination, the first is preferred because it is the largest. This test also shows the presence of heteroskedasticity. White test shows that the calculated value (18.21) of the test statistics is greater than the tabular value and we can conclude heteroskedasticity is present. In the Breusch-Pagan test the value of the calculated ratio is 20.33 and it is greater than the critical value of X^2 distribution that is 3.841 with one degree of freedom, and we also can conclude that heteroskedasticity is present.

5. Conclusion

One of the classic assumptions of linear regression is homoskedasticity, and when it is disturbed, heteroskedasticity occurs. Graphical methods and heteroskedasticity tests are used to detect heteroskedasticity, although it is not possible to say with certainty which test is the best. In this paper, we explained and applied the graphical method and four tests (Goldfeld-Quantum, Glejser, White and Breusch-Pagan test). Through a review of the literature, it can be seen that many authors have addressed this issue and used various tests to detect heteroskedasticity.

The tests were applied by data simulation. It can be seen that the graphical method and all four applied tests confirm the presence of heteroskedasticity, so we can conclude that all four tests showed a good result and that it can be confirmed the assumption of the existence of heteroskedasticity in the model.

Future researchers are left with the question of solving heteroskedasticity, ie the question of removing heteroskedasticity from the model. When eliminating heteroskedasticity, care must be taken which method can be used depending on the form σ_i^2 .

Author Contributions: Each author has participated and contributed sufficiently to take public responsibility for appropriate portions of the content.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Arbia, G. (2006). Spatial econometrics: statistical foundations and applications to regional convergence. Springer Science & Business Media.

Aue, A., Horváth, L., & F. Pellatt, D. (2017). Functional generalized autoregressive conditional heteroskedasticity. Journal of Time Series Analysis, 38(1), 3-21.

Barreto, H., & Howland, F. (2006). Introductory econometrics: using Monte Carlo simulation with Microsoft excel. Cambridge University Press.

Baum, C., & Schaffer, M. (2019). IVREG2H: Stata module to perform instrumental variables estimation using heteroskedasticity-based instruments.

Brüggemann, R., Jentsch, C., & Trenkler, C. (2016). Inference in VARs with conditional heteroskedasticity of unknown form. Journal of econometrics, 191(1), 69-85.

Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018). Inference in Linear Regression Models with Many Covariates and Heteroskedasticity Supplemental Appendix.

Charpentier, A., Ka, N., Mussard, S., & Ndiaye, O. H. (2019). Gini Regressions and Heteroskedasticity. Econometrics, 7(1), 4.

Crudu, F., Mellace, G., & Sándor, Z. (2017). Inference in instrumental variables models with heteroskedasticity and many instruments. Manuscript, University of Siena.

Ferman, B., & Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. Review of Economics and Statistics, 101(3), 452-467.

Halunga, A. G., Orme, C. D., & Yamagata, T. (2017). A heteroskedasticity robust Breusch–Pagan test for Contemporaneous correlation in dynamic panel data models. Journal of econometrics, 198(2), 209-230.

Harris, D., & Kew, H. (2017). Adaptive long memory testing under heteroskedasticity. Econometric Theory, 33(3), 755-778.

Im, K. S. (2000). Robustifying Glejser test of heteroskedasticity. Journal of Econometrics, 97(1), 179-188.

Jovičić, M. (2011). Ekonometrijski metodi i modeli. Centar za izdavačku delatnost. Ekonomski fakultet. Beograd.

Kalina, J., & Peštová, B. (2017). Exact Inference in Robust Econometrics under Heteroscedasticity. 11th International Days of Statistics and Economics MSED 2017.[Proceedings.] Slaný: Melandrium, 636-645.

Linton, O., & Xiao, Z. (2019). Efficient estimation of nonparametric regression in the presence of dynamic heteroskedasticity. Journal of Econometrics, 213(2), 608-631.

Lütkepohl, H., & Netšunajev, A. (2017). Structural vector autoregressions with heteroskedasticity: A review of different volatility models. Econometrics and statistics, 1, 2-18.

Lütkepohl, H., & Velinov, A. (2016). Structural Vector Autoregressions: Checking Identifying Long-Run Restrictions via Heteroskedasticity. Journal of Economic Surveys, 30(2), 377-392.

Mladenović, Z. (2011). Uvod u ekonometriju. Centar za izdavačku delatnost. Ekonomski fakultet. Beograd.

Mladenović, Z. & Nojković, A., (2017). Zbirka rešenih zadataka iz ekonometrije. Centar za izdavačku delatnost. Ekonomski fakultet. Beograd.

Mladenović, Z. & Petrović, P., (2017). Uvod u ekonometriju. Centar za izdavačku delatnost. Ekonomski fakultet. Beograd.

Moussa, R. K. (2019). Heteroskedasticity in One-Way Error Component Probit Models. Econometrics, 7(3), 35.

Ou, Z., Tempelman, R. J., Steibel, J. P., Ernst, C. W., Bates, R. O., & Bello, N. M. (2016). Genomic prediction accounting for residual heteroskedasticity. G3: Genes, Genomes, Genetics, 6(1), 1-13.

Sato, T., & Matsuda, Y. (2017). Spatial autoregressive conditional heteroskedasticity models. Journal of the Japan Statistical Society, 47(2), 221-236.

Taşpınar, S., Doğan, O., & Bera, A. K. (2019). Heteroskedasticity-consistent covariance matrix estimators for spatial autoregressive models. Spatial Economic Analysis, 14(2), 241-268.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817-838.