The Heteroskedasticity Tests Implementation for Linear Regression Model Using MATLAB

Lyudmyla Malyarets, Katerina Kovaleva, Irina Lebedeva, Ievgeniia Misiura and Oleksandr Dorokhov
Simon Kuznets Kharkiv National University of Economics, Nauky Avenue, 9-A, Kharkiv, Ukraine, 61166
E-mail: aleks.dorokhov@meta.ua
http://www.hneu.edu.ua/

Keywords: regression model, homoskedasticity, testing for heteroskedasticity, software environment MATLAB

Received: September 23, 2017

The article discusses the problem of heteroskedasticity, which can arise in the process of calculating econometric models of large dimension and ways to overcome it. Heteroskedasticity distorts the value of the true standard deviation of the prediction errors. This can be accompanied by both an increase and a decrease in the confidence interval. We gave the principles of implementing the most common tests that are used to detect heteroskedasticity in constructing linear regression models, and compared their sensitivity. One of the achievements of this paper is that real empirical data are used to test for heteroskedasticity. The aim of the article is to propose a MATLAB implementation of many tests used for checking the heteroskedasticity in multifactor regression models. To this purpose we modified few open algorithms of the implementation of known tests on heteroskedasticity. Experimental studies for validation the proposed programs were carried out for various linear regression models. The models used for comparison are models of the Department of Higher Mathematics and Mathematical Methods in Economy of Simon Kuznets Kharkiv National University of Economics and econometric models which were published recently by leading journals.

Pozvetek: Avrorji prispevka se ukvarjajo s problemi ekonometričnih modelov z veliko dimenzijami, kjer je izračun problematičen. Razvijejo metodo v MATLABu za multifaktorske regresijske modele.

1 Introduction

In econometrics, a linear regression model is often used to describe different processes and phenomena. Using matrix notation, the linear model regression can be given as:

$$\mathbf{Y} = \mathbf{XB} + \varepsilon \tag{1}$$

where $\mathbf{Y}$ and ε are $n \times 1$ matrices, $\mathbf{X}$ is $n \times (m+1)$, and $\mathbf{B}$ is $(m+1) \times 1$; n is the number of measurements (sample size); m is the number of independent variables in the regression model.

For the i^{th} row of $\mathbf{X}$ (the i^{th} observation) the linear regression model can be written as follows:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_m x_{im} + \varepsilon_i \tag{2}$$

where y_i are the values of the dependent variable, $y_i \in \mathbf{Y}$; i is the experiment identification number, $i = \overline{1, n}$; x_{ij} are the values of the independent variable $x_j \in \mathbf{X}$ ($j = \overline{1, m}$) in the i^{th} experiment; b_0 is the constant term of the equation; b_j are the regression coefficients, $b_0, b_j \in \mathbf{B}$; ε_i are the residuals (model errors).

An error term is introduced in a regression model because the model does not fully represent the actual relationship between the variables of the model. As a result of this incomplete relationship, there are differences between the observed responses (values of the variable being predicted) in the given dataset and those predicted by a linear function of a set of explanatory variables. The error term is the amount at which the equation may differ from measurements. In other words that is the 'white noise'.

As a rule, the building a linear regression model is done by the method of ordinary least squares (OLS). This method for estimating the unknown parameters is based on the minimization of the sum of the squares of the model errors. The estimators of model parameters determined by OLS are known as best linear unbiased estimators (BLUE). The variances of the model parameters are determined by:

$$S^2_{b_j} = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-m-1} z_{jj} = \sigma_e^2 \cdot z_{jj} \tag{3}$$

where z_{jj} is the diagonal element of matrix $\mathbf{Z} = (\mathbf{X}'\mathbf{X})^{-1}$ which corresponds to the parameter b_j; σ_e is the standard error.

The OLS application requires the realization of a number of conditions [1 – 3]. Only if these conditions are met, the estimates calculated by such a model will be unbiased, efficient and well-off. These conditions are formulated in the form of the Gauss – Markov theorem.

According to this theorem there are four principal assumptions which admit the using of linear regression models for research and prediction. One of them is the homoskedasticity (constant variance) of the errors in relation to any independent variable.

Homoskedasticity makes the assumption that the errors have a constant variance: $\text{var}(\varepsilon) = const$ and independent of causal variables: $\text{cov}(x_j, \varepsilon) = 0$ for all j, $j = \overline{1, m}$. The error ε is a random variable distributed according to the normal law: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ where the mathematical expectation of the error term is zero and the variance is constant. Failure to comply with this requirement leads to bias in the estimates obtained using such a regression model. Thus [4] indicate that estimation uncertainty may increase dramatically in the presence of conditional heteroskedasticity.

The requirement of homoskedasticity also exists in the construction of the econometric model using the maximum likelihood method [5 – 7].

When the scatter of the errors is different, varying depending on the value of one or more of the independent variables, the error terms are heteroskedastic. Namely the distribution law of errors remains normal with a mathematical expectation equal to zero, but the errors of the model are a function of the values of the independent variables: $\varepsilon \sim N(0, f(\mathbf{X}))$, where $f(\mathbf{X})$ is a function that describes the change in the variance of errors as a function of the values of the independent variables.

A similar problem arises during the building of semiparametric [8 – 10] and nonparametric [11, 12] models.

Heteroskedasticity makes difficult to gauge the true standard deviation of the forecast errors. The OLS estimates are no longer BLUE. Thus, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. In particular, heteroscedasticity does not allow us to use equation 3 for the computation of S_{b_j}, since it assumes a uniform dispersion of the errors. Under heteroskedasticity, the sample variance of OLS estimator is

$$\mathbf{Var}(\hat{b_j}) = \sigma_e^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Omega\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \qquad (4)$$

where **Ω** is the covariance matrix, the elements of which are defined as the variance of the model parameters. Under homoskedasticity, **Ω**= I. Equation 4 is correct if there is no autocorrelation.

For these reasons, all the conclusions obtained on the basis of the corresponding t – statistics and F – statistics, as well as interval estimates, will be unreliable.

A unified approach to the estimation of heteroscedasticity is lacking. To solve this problem, a large number of different tests and criteria have been developed: the Spearman rank correlation test, the Park test, the Glaser test, the Goldfeld – Quandt test, the Breusch – Pagan test, the Leven's test, the White test, and so on.

The application of all the above tests is very difficult for the so-called ‘manual’ account, and for a large set of initial data it is completely impossible.

There are a lot of software with which you can identify heteroscedasticity. These are professional packages (SAS, BMDP), universal packages (STADIA, OLIMP, STATGRAPHICS, STATISTICA, SPSS) and specialized packages (DATASCOPE, BIOSTAT, MESOSAUR).

When using economic data researcher can face two main problems. Firstly, all the listed software are quite expensive and price of the product may be an insurmountable barrier for the young researcher. Secondly, company-developer never provides the source code, considering that this is not necessary for an ordinary user. Therefore, we can not modify the built-in algorithms to detect and eliminate heterosquadity.

Another drawback of the above program products is the outdated conceptual approaches to econometric methods, which are constantly being improved.

For example, the program products SPP and MICROSTAT calculate the coefficient of multiple correlations as the square root of the coefficient of determination. STATGRAPHICS calculates it as the square root of the adjusted coefficient of determination [13]. While in theory the coefficients of multiple correlations is estimated using elements of the correlation matrix [2].

Another important aspect that should be taken into account is the existence of different algorithms to identify heteroskedasticity and the specific problem of division by zero [14].

Ideal option would be to create your own software product that would take into account the research tasks.

However, to write such a program, the economist should be an expert in algorithmic programming. But this happens rarely.

In this article, we carry out a comparative analysis of the tests most often used to detect heteroskedasticity [1, 2, 14] and give their source code. The use of program code allows you to modify the program in accordance with the objectives of the study.

2 Analysis of literary data and the formulation of the problem

Before starting the construction of the regression model, it is necessary to verify whether the conditions of the Gauss-Markov theorem are fulfilled.

One of the main methods of preliminary research on heteroskedasticity is a visual analysis of the graph of residues. On these graphs, the scattering of points can vary depending on the value of the independent variables [14, 15].

To estimate heteroskedasticity, are used such quantitative tests [15 – 17] as the White test, the Goldfeld – Quandt test, the Breusch – Pagan test, the Park test, the Glazer test and also the Spearman test. Unlike other tests the Spearman rank correlation test is a nonparametric statistical test for the heteroskedasticity of random errors in the econometric model. The test algorithm can be studied in detail in [18, 19]. However, it is still not implemented in software products which are used to build multiple models [20 – 27].

In this paper we examined the software packages most commonly used in economic activity, which contain tests for heteroskedasticity [15, 28]. Indeed, these software products do not contain the Spearman rank correlation test.

The most widely used for evaluating heteroscedasticity is the Park test [20, 21]. However, the Park test contains the assumption that the change in the remnants of the model is described by a functional dependence of a certain type. It was noted in [24, 25] that this can lead to unreasonable conclusions. Therefore, the authors propose to consider the Park test together with other tests.

The software implementation of the Park test for multiple models also does not exist [28]. As far as we know software implementation of the Park test for multifactor models also does not exist.

Another test that the authors of the article implemented in the MATLAB environment is the Goldfeld – Quandt test. This test to check for heteroskedasticity of random errors is used when there is reason to believe that the standard deviation of errors is proportional to some variable.

The test statistics has a Fisher distribution [18, 27]. The Goldfeld – Quandt test can also be used if there is an assumption of intergroup heteroskedasticity, when the variance of errors takes, for example, only two possible values. In this case, for the application of the test, there is a need for its software implementation, since applied commercial software has not taken this possibility into account [25, 28].

In scientific articles ~~on~~ for the problem of detecting heteroskedasticity, the Breusch – Pagan test is often considered [10, 29]. We also carried out research this problem. But it oversteps this article.

Analysis of literature sources shows that all tests of heteroskedasticity detection are difficult for 'manual' application and require the development of special software. In turn, the software of econometric research does not contain built-in functions for heteroskedasticity testing with open source code.

That is way the authors of this article attempted to implement the above tests for heteroskedasticity in the construction of multifactor econometric models in the MATLAB software environment.

It should be noted that MATLAB does not contain ready-made software implementation to verify compliance of homoskedasticity. We chose it as a programming environment. For this purpose, other programming environments can also be used, for example, ~~a~~ the free software environment R.

The authors have chosen MATLAB by the following reasons. First, MATLAB is used as a high-level programming language for writing scripts (Spearman.m, Parks.m and Gold_Quan.m). Secondly, MATLAB includes built-in functions for constructing regression models (Econometric toolbox), which gave the authors relief from the duty of programming the standard functions of regression analysis. Thirdly, the authors worked with data structures based on matrices.

3 Aims and objectives of the study

The purpose of the article is to present functions to check for heteroskedasticity in multifactor regression models. The implementation is made in MATLAB.

To achieve this purpose, it is necessary to solve a number of problems. Namely:

- writing the program code in the MATLAB programming environment;
- planning and execution of computer calculations;
- completion of programs;
- analysis and interpretation of results;
- comparison with the results of calculations using software products of leading companies.

4 Practical implementation of the criteria for the detection of heteroskedasticity in econometric models in the MATLAB

4.1 Spearman's rank correlation test for multiple regression models

The use of the Spearman's test assumes that the variance of model errors will increase (or decrease) with increasing values of the independent variable.

This means that the absolute values of errors ε_i $(i=\overline{1,n})$ and the values x_{ij} of the independent variable x_j $(j=\overline{1,m})$ will correlate with each other.

To check whether heteroskedasticity is statistically significant the Spearman's test provides for the following stages:

1) Estimation of the parameters of the econometric model by the OLS:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_m x_{im}, \quad (5)$$

where $\hat{y}_i$ is the predicted response in accordance with the model when the independent variables are $(x_{i1};\ x_{i2};\ ...x_{im})$;

2) Calculate model errors as the difference between the empirical and the ratchet value of the dependent variable:

$$\varepsilon_i = y_i - \hat{y}_i \tag{6}$$

where y_i is the value of the dependent variable in the i^{th} experiment;

3) The pairs (x_{ij}, ε_i) are ranked in order of increasing values of the independent variable x_j;

4) The coefficient of rank correlation between ε_i and x_{ij} is calculated as

$$r_{x\varepsilon} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}, \tag{7}$$

where d_i is the difference between the two ranking;

5) The significance of $r_{x\varepsilon}$ is tested by using $t-$ statistic:

$$t = \frac{r_{x\varepsilon}\sqrt{n-2}}{\sqrt{1-r_{x\varepsilon}^2}} \tag{8}$$

6) In accordance with the predetermined confidence probability p (where $\alpha = 1 - p$) the tabulated value of $t_{cr.} = t_{0.5\alpha}(n-2)$ is found. Then the calculated value is compared with the critical one.

If the *t*-statistic value is greater than the critical value, we must say that heteroscedasticity is statistically significant. Here α is the significance level which is chosen to test the null hypothesis: $\rho_{x\varepsilon} = 0$. In the opposite case, the null hypothesis is non-contradictory.

As an example of the implementation of this test, we can suggest the following m-file named *Spearman*:

```
========================================
% initialization:
X1 = load('data1.scv');
X2 = load('data2.scv');
X3 = load('data3.scv');
X4 = load('data4.scv');
X5 = load('data5.scv');
Y = load('data.scv');
```

% Formation of the source data array:

```
X = [ones(n,1) X1' X2' X3' X4' X5'];
% Construction of a linear multifactor
% model by OLS - method:
[b,bint,r,rint,stats] = regress(Y,X,0.05);
y_p = b(1) + b(2).*X1 + b(3).*X2+
b(4).*X3+b(5).*X4+b(6).*X5
sprintf('Model:')
fprintf('y_p = %f + %f *X1+%f *X2+%f *X3+%f
*X4+%f *X5',b)
% Calculation of model remains:
e = Y - y_p';
% Preparing an array for further work:
X = [X1' X2' X3' X4' X5'];
[n,m] = size(X);% Determining the size of the
source data
%=======================================
%% Spearman rank correlation test
% Ranking of factors:
[Xs I] = sort(X)
Dx = zeros(n,m);
for j = 1:m
    for i = 1:n
        Dx(i,j) = i;
    end
end
TMP = zeros(n,m);
% Filling an array of factors with ranks
% taking into account their sequence numbers:
for j = 1:m
    for i = 1:n
        i1 = I(i,j);
        TMP(i1,j) = Dx(i,j);
    end
end
X = [X TMP]% Output array
% Ranking of remains:
[es I] = sort(e);
es = [es ones(size(e),1)];
e = [e ones(size(Y),1)];
sprintf(' critical values t:','\n')
t_r(:,j) = (r(:,j)*sqrt(n-1))/sqrt(1 -
r(:,j)^2);
end
t_r % output array t-Statistics by Spearman
% Comparative analysis and conclusions:
c = 0;
for i = 1:size(e)
    es(i,2) = i;
end
% Filling an array of remains with ranks
% taking into account their sequence numbers:
for i=1:size(e)
    e(I,2) = es(:,2);
end
e% an array of remains which contains ranks
r = zeros(1,m);
d = zeros(n,m);
% Calculating the difference of ranks
for j = 1:m
    for i = 1:n
        d(i,j) = TMP(i,j) - e(i,2);
    end
end
d % difference in rank
% The square of the difference of ranks:
for j = 1:m
    d(:,j) = d(:,j).^2;
end
d
Sd = zeros(1,m);
% The sum of the difference of ranks squares
% by the corresponding columns of ranks:
for j = 1:m
    Sd(:,j) = sum(d(:,j));
end
Sd % output array
% Calculating Spearman's Statistics:
for j = 1:m
    r(:,j) = 1 - (6*Sd(:,j))/(n*(n^2-1));
end
r % Output array
t_r = zeros(1,m);
%% Testing of the significance of the Spearman
coefficient:
t_t = tinv(0.975,n-2)% tabulated value t
for j = 1:m
  if abs(t_r(:,j)) < abs(t_t)
    sprintf(' Heteroskedasticity is absent ')
  else
    sprintf(' Heteroscedasticity is present ')
    c = c + 1;
  end
end
========================================
```

4.2 Park's test for multiple regression models

R. Park proposed a test to check for heteroskedasticity, which is based on some formal dependencies. Namely, it assumes that the heteroskedasticity may be proportional to some power of an independent variable x_j in the multiple models.

Since the variance of errors $\sigma_i^2 = \sigma^2(\varepsilon_i)$ is a function of the $i-$th value x_{ij} of the explanatory variable x_j, and for its description Park proposed the this dependence: $\sigma_i^2 = \sigma^2 x_{ij}^{\beta} \varepsilon^{v_i}$.

After computing its logarithms, we obtain the following relation: $\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln x_{ij} + v_i$. Since the variances σ_i^2 are usually unknown, they are replaced by their estimates ε_i^2.

The Park's test provides for such effectuation stages:

1) Estimation of the parameters of the econometric model by the OLS (Equation 5);

2) Calculation of the value $\ln \varepsilon_i^2 = \ln(y_i - \hat{y}_i)^2$ for each observation;

3) Building the regression model:

$$\ln \varepsilon_i^2 = \alpha + \beta \ln x_{ij} + \upsilon_i, \tag{9}$$

where $\alpha = \ln \sigma^2$. For the case of multiple regressions, this dependence is constructed for each explanatory variable;

4) Verification of statistical significance of the coefficient β on the basis of $t-$statistics:

$$t = \left| \beta / \sigma_{\beta} \right|. \tag{10}$$

5) In accordance with the predetermined confidence probability p (where $\alpha = 1 - p$) the tabulated value of $t_{cr.} = t_{\alpha}(n-m-1)$ is found. Then the calculated value is compared with the critical one.

If $t > t_{\alpha}(n-m-1$, then at the level of significance α the coefficient β is statistically significant and there is a link between $\ln \varepsilon_i^2$ and $\ln x_i$. It means that heteroskedasticity is present in statistical data.

The M-file named *Park's* which is implementation of the Park test has the form:

```
=======================================
% initialization:
X1 = load('data1.scv');
X2 = load('data2.scv');
X3 = load('data3.scv');
X4 = load('data4.scv');
X5 = load('data5.scv');
Y = load('data.scv');
% Formation of the source data array:
X = [ones(n,1) X1' X2' X3' X4' X5'];
[n, m] = size(X);
% ==========  Park Test Algorithm =======
%  1 stage of the Park test
% Construction of a linear multifactor
% model by OLS - method:
[b,bint,r,rint,stats] = regress(Y,X,0.05);
y_p   =   b(1)   +   b(2).*X1   +   b(3).*X2+
b(4).*X3+b(5).*X4+b(6).*X5
sprintf('Model:')
fprintf('y_p  =  %f  +  %f  *X1+%f  *X2+%f  *X3+%f
*X4+%f *X5')
%   2 stage of the Park test
ln_eps = log((Y' - y_p).^2)
%   3 stage of the Park test
for j=1:m
    for i = 1:n
        X(i,j) = log(X(i,j));
    end
end
%  4 stage of the Park test
for i = 2:m
 [bet, dev,stat] = glmfit(X(:,i),ln_eps);
    t_t = tinv(0.95, n-2);
    t_r = stat.t(2);
% Comparative analysis and conclusions:
        if abs(t_r) < abs(t_t)
            sprintf('  Heteroskedasticity  of  %i
factor is absent \n',i-1)
        else
            sprintf('  Heteroskedasticity  of  %i
factor is present\n',i-1)
        end
end
=======================================
```

The Park test's weakness is that it assumes the heteroskedasticity has a particular functional form.

4.3 Goldfeld – Quandt test for multiple regression models

When using the Goldfeld-Quandt test for heteroscedasticity, it is assumed that model errors σ_{ε} depend on one of the external variables x_j: $\sigma_{\varepsilon_i}^2 = \sigma^2 x_{ij}^2$

It is also assumed that errors ε_i are distributed according to the normal law, there is no autocorrelation.

The Goldfeld-Quandt test provides for such effectuation stages:

1) Estimation of the parameters of the econometric model by the OLS (Equation 5);

2) Ranking of all n observations in magnitude of the independent variable x_j;

3) Segregation this ordered sample into three approximately equal parts k, $n-2k$, k, respectively;

4) For each part of the sample that has a volume k, its regression equation is constructed and the sums of the squares of the deviations determine:

$$RSS_1 = \sum_{i=1}^{k} \varepsilon_i^2 \tag{11}$$

and

$$RSS_3 = \sum_{i=n-k+1}^{n} \varepsilon_i^2 . \tag{12}$$

Than empirical meaning of the $F-$statistic is calculated:

$$F = \frac{RSS_1/(k-m-1)}{RSS_3/(k-m-1)} \quad (13)$$

5) Evidence of heteroskedasticity is based on a comparison of the residual sum of squares (*RSS*) using the F – statistic. The calculated value is compared with the critical value $F_{cr.} = F_\alpha(k-m-1; k-m-1)$ in accordance with the predetermined confidence probability p (where $\alpha = 1-p$).

If $F < F_\alpha(k-m-1; k-m-1)$, this means that at the level of significance α the hypothesis that there is no heteroskedasticity does not have grounds to reject. In the opposite case, the hypothesis of the absence of heteroskedasticity is rejected.

For multiple regressions, we performed tests for all factors. The M-file named *Gold_Quan* which is the implementation of the Goldfeld – Quandt test has the form:

```
=======================================
% initialization:
X1 = load('data1.scv');
X2 = load('data2.scv');
X3 = load('data3.scv');
X4 = load('data4.scv');
X5 = load('data5.scv');
Y = load('data.scv');
% Formation of the source data array:
X = [ones(n,1) X1' X2' X3' X4' X5'];
[n, m] = size(X);
%=======================================
%% Goldfeld – Quandt test:
[Xsort Is] = sort(X);
for i=1:size(Y)
    Ysort(i,1) = Y(Is(i),1);
end
Dat = [Xsort Ysort];
c = fix(4*n/15);
k = fix((n - c)/2);
if floor(k) > 0.4
    k = k+1;
end
k
% Selective aggregate 1:
Dat1 = Dat(1:k,:);
[b1,dev1,stats1] = glmfit(Dat1(:,1),Dat1(:,2));
S1 = sum(stats1.resid.^2);
% Selective aggregate 2:
Dat2 = Dat(n-k+1:n,:);
[b2,dev2,stats2] = glmfit(Dat2(:,1),Dat2(:,2));
S2 = sum(stats2.resid.^2);
% Testing the hypothesis:
if S1 > S2
    Fp = S1/S2;
else
    Fp = S2/S1;
end
Ft = finv(0.95,k-m-1,k-m-1);
if Fp > Ft
     sprintf(Heteroscedasticity is present ')
else
    sprintf(Heteroscedasticity is absent ')
end
=======================================
```

A weakness of the Goldfeld – Quandt test is that the result is dependent on the criteria chosen for separating the sample measurements into their representative groups.

5 Results of numerical experiments

The problem of detecting heteroskedasticity in various multifactor econometric models was considered.

For carrying out numerical simulation experiments we used both the models of the Department of Higher Mathematics, Economic and Mathematical Methods of KhNEU [30 – 33], and econometric models which were published recently by leading journals [34 – 36].

To check for heteroscedasticity, we used real data. This is one of the advantages of this paper. However, it is possible to use the data obtained with the Monte Carlo simulation [6, 7, 37 – 39].

Numerical experiments were performed on the configuration AMD Athlon 64 3200+1.5Gb Ram, graphic accelerator – Nvidia GeForce GTX 560 2Gb with using technology NVIDIA CUDA 4.2.

Let's look at a concrete example of what happens to an eccentric model, if you do not take into account heteroskedasticity.

As a model problem, the linear regression model was calculated for the cost of electronic textbooks produced by the Department Higher Mathematics and Mathematical Methods in Economy. The initial data and designations used in the process of correlation-regression analysis are shown in Figure 1, where Y is the resulting factor Y (cost of the electronic textbook).

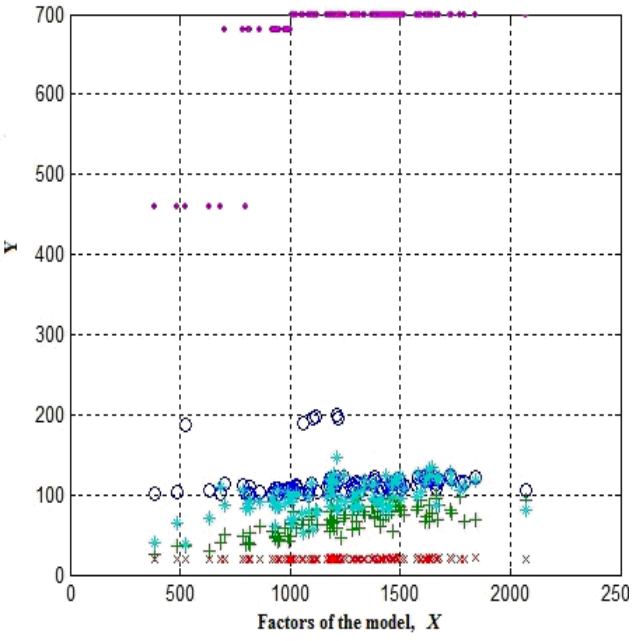

Figure 1: Initial data for model building

Figure 1 shows the dependence of the cost of the electronic textbook (Y) on such external factors:

○ - X1 (average cost of developers' wages);
+ - X2 (publication volume);
× - X3 (average CD recording price);

* - X4 (storage and distribution costs);
• - X5 (cost of the use of licensed software).

The regression model was constructed using the built-in function Matlab-regress (y, X, alpha) with the code:

```
=========================================
% The program for multiple regression model building, if heteroskedasticity is not taken into account :
[b,bint,r,rint,stats] = regress(Y,X,0.05);
y_p = b(1) + b(2).*X1 + b(3).*X2+ b(4).*X3+b(5).*X4+b(6).*X5;
sprintf(' Heteroskedasticity is not taken into account:')
fprintf('y_p = %f + %f *X1+%f *X2+%f *X3+%f *X4+%f *X5',b)
=========================================
```

The program for constructing multiple regressions, if you do not take into account heteroskedasticity, gives such a result:

$$\hat{y} = -1864.06 + 0.33 \cdot x_1 + 10.61 \cdot x_2 + \\ + 70.90 \cdot x_3 + 3.33 \cdot x_4 + 0.87 \cdot x_5. \quad (14)$$

The results of calculating the errors of the model represented by the Equation 10 are shown in Figure 2.

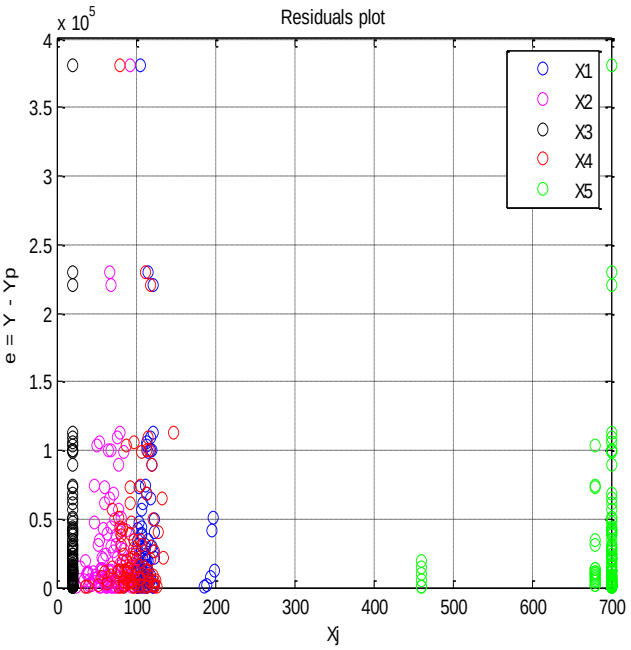

Figure 2: Graphic illustration of the remnants of the model

Analysis of the remnants of the model indicates that for this model the dispersion of remnants increases with an increasing of the value of external factors, that is, heteroskedasticity can not be ignored.

Using the program procedures developed by the authors to identify heteroskedasticity, the following results were obtained:

```
=========================================
   ans = Heteroskedasticity 1 is absent
   ans = Heteroskedasticity 2 is absent
   ans = Heteroskedasticity 3 is absent
   ans = Heteroskedasticity 4 is absent
    ans = Heteroskedasticity 5 is present
=========================================
```

The construction of the regression model, which takes into account the heteroskedasticity, was performed using the built-in function MATLAB: robustfit (X, y, wfun, tune,const).

It should be emphasized that the presence or absence of heteroskedasticity in the initial data is determined automatically by using the check box.

For this we used the code:

```
=========================================
%c is a parameter that takes the value 0 or 1
%(where 0 - Heteroscedasticity is absent, 1 -
% Heteroscedasticity is present),
%c depends on the result of the scripts' work
if c > 0
X = [X1' X2' X3' X4' X5'];
[b,stats3] = robustfit(X,Y,'fair',0.001,'on');
y_p = b(1) + b(2).*X1 + b(3).*X2+
b(4).*X3+b(5).*X4+b(6).*X5;
sprintf('Heteroskedasticity is taken into account:')
fprintf('y_p = %f + %f *X1+%f *X2+%f *X3+%f
*X4+%f *X5',b)
   end
=========================================
```

The program for multiple regression model building, if heteroskedasticity is taken into account yields this result:

$$\hat{y} = 27.85 + 0.94 \cdot x_1 + 10.33 \cdot x_2 - \\ - 29.16 \cdot x_3 + 4.18 \cdot x_4 + 0.80 \cdot x_5. \quad (15)$$

Thus, the above procedure allows eliminating heteroskedasticity. In this case, the resulting models will be able to adequately reflect the reality.

Table 1 shows the results of numerical experiments on testing of programs which are presented in this article on various multifactor models.

As can be seen from Table 1, software products developed by us using MATLAB can be proposed both for constructing multifactor econometric models, and for investigating the latter for the presence of heteroskedasticity.

In doing so, we used new numerical algorithms, developed on the basis of well-known tests of heteroskedasticity detection.

Open source code allows the researcher to use this software to solve their own problems.

Multiple Models	Theoretical results	The results of the work of the authors' programs		
		Spirmen.m	*Park.m*	*Goldfeld – Quandt.m*
Model [28]: Linear approximation	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Power approximation	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Hyperbolic approximation	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present
Model [30]	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present
Model [31]	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present
Model [33]	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Model [34]	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Model [35]	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is absent*	Heteroskedasticity is present

** The conclusion is not justified, since the test uses a monotonically increasing function*

Table 1: Results of testing programs on multiple models

6 Conclusion and future work

The article examined one of the key problems of regression analysis, which consists in verifying the fulfillment of the requirement of homoskedasticity of the remainders of the model. To this end we used various statistic tests.

Analysis of literature sources and our own studies confirm the complexity of using all existing tests for detecting heteroskedasticity in the 'manual account' mode. Therefore, we gave our own implementation in MATLAB for tests used for detecting heteroskedasticity.

This problem was successfully solved, as shown results of numerical experiments which are presented in the article. We represent all software products we have created with open source code, which enables each researcher to customize the program to their problems.

In conclusion, we want to note that the work presented in this article is an on going work having the final purpose to create a complete and effective software for detecting-heteroskedasticity in regression models.

Another further development consists in developing a complete econometric toolbox in MATLAB.

7 Reference

[1] Dougherty C (2016). Elements of econometrics http://www.londoninternational.ac.uk/sites/default/files/programme_resources/lse/lse_pdf/subject_guides/ec2020_ch1-4.pdf

[2] Dougherty C (2016). *Introduction to Econometrics* (5th edition) University Press: Oxford

[3] Hansen B (2018). Econometrics. University of Wisconsin http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf

[4] Brüggemann R, Jentsch C and Trenkler C (2016). Inference in VARs with conditional heteroskedasticity of unknown form *Journal of econometrics* 191 pp. 69-85. http://dx.doi.org/10.1016/j.jeconom.2015.10.004

[5] Cordeiro G (2008). Corrected maximum likelihood estimators in linear heteroskedastic regression models *Brazilian Review of Econometrics* 28 pp. 11–16.

[6] Hayakawa K and Pesaran H (2015). Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity *Journal of econometrics* 188 pp. 111-134. http://dx.doi.org/10.1016/j.jeconom.2015.03.042

[7] Chen S, Khan S and Tang X (2016). Informational content of special regressors in heteroskedastic binary response models *Journal of econometrics* 193 pp. 162-182. http://dx.doi.org/10.1016/j.jeconom.2015.12.018

[8] Kai B, Li R and Zou H (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models *The annals of statistics* 39 pp. 305-332

[9] Pelenis J (2014) Bayesian regression with heteroscedastic error density and parametric mean function *Journal of econometrics* 178 pp. 624-638. http://dx.doi.org/10.1016/j.jeconom.2013.10.006

[10] Norets A (2015). Bayesian regression with nonparametric heteroskedasticity *Journal of econometrics* 185 pp. 409-419. http://dx.doi.org/10.1016/j.jeconom.2014.12.006

[11] Wei C and Wan L (2015). Efficient estimation in heteroscedastic varying coefficient models *Econometrics* 3 pp. 1-7.

[12] Shen S, Cui J and Wang C (2014). Testing heteroscedasticity in nonparametric regression based on trend analysis *Journal of Applied Mathematics* 2014 pp. 1-5. http://dx.doi.org/10.1155/2014/435925

[13] Pen R (2015). Planirovaniye experimenta v Statgraphics Centurion. Mezdunarodnye zurnal eksperimentalnoy obrazovanija pp. 160–161 (in Russian)

[14] Malyarets L (2014). *Economico-mayematechni metody i modeli*. KhNEU im. S Kuznetz: Kharkiv (in Ukrainian)

[15] Williams R (2015). Heteroskedasticity University of Notre Dame https://www3.nd.edu/~rwilliam/stats2/l25.pdf
[16] Kolchinskaya E (2015). Vlijanie transportnoj infrastrucrury na promyshlennoe razvitie regionov Rossii. *Aktualnye problemy ekonomiki* 34 pp. 77-82 (in Russian)
[17] Radkovskaya E (2015). Matematicheskie metody v sovremennyh ekonomicheskih issledovaniyah. *Vestnik Yugorskogo gosudarstvennogo universiteta* 37 pp. 37-40 (in Russian)
[18] Gmurman V (2001). *Teorija verojatnostej i matematicheskaya statistika* (7th edition). Vyshaya shkola: Moscow (in Russian)
[19] Heteroskedasticity http://gauss.stat.su.se/gu/e/slides/Lectures%208-13/Heteroscadasticity.pdf
[20] Redace R (2017). Use the Park test to check for heteroskedasticity http://www.dummies.com/education/economics/econometrics/use-the-park-test-to-check-for-heteroskedasticity/
[21] Mazorchuk M (2014). Osobennosti vybora metodov izmereniya nadezhnosti pedagogicheskih tekstov. *Radioelectrohhi i komp'jutorni sistemy* pp. 131-137 (in Russian)
[22] Kim D, El-Tawil S and Naaman A (2007). Correlation between single fiber pullout and tensile response of FRC composites with high strength steel fibers. Fifth International RILEM Workshop on High Performance Fiber Reinforced Cement Composites (HPFRCC5). RILEM, ed H W Reinhardt and A E Naaman: Paris pp. 67-76.
[23] Baranov N and Sorokin L (2015). Komp'yuternye prikladnye programmy v formatirovanii stilja myshleniya budushchego spetsialista. *Mezhdunarodnyj nauchno-issledovatekskij zhurnal* 42 pp. 60-62 (in Russian)
[24] Kazanskaya A and Kompanietz V (2009) Opyt issledovanija metodov klasternogo analiza iz paketa Statistica 6. 0 na primere vuborki gorodov. *Izvestiya YuFU, Tehnicheskie nayki* pp. S103-110 (in Russian)
[25] Kosonogov V (2014). The psychometric properties of the Russian version of the empathy *Quotient Psychology in Russia* pp. 196-104
[26] Yüce M (2017). An Asymptoic test for the detection of heteroskedasticity http://eidergisi.istanbul.edu.tr/sayi8/iueis8m2.pdf
[27] Redace R (2017). Test for heteroskedasticity with the Goldfeld – Quandt test http://www.dummies.com/education/economics/econometrics/test-for-heteroskedasticity-with-the-goldfeld-quandt-test/
[28] Krasilnikov D (2011). Programmnoe obespechenie ekonometricheskogo issledovanija Econometric Software. *Vestnik Nizhegorodskogo universiteta im. N I Lobachevskogo* pp. 231-238 (in Russian)
[29] Halunga A, Orme C and Yamagata T (2017). A heteroskedasticity robust Breusch–Pagan test for contemporaneous correlation in dynamic panel data models *Journal of econometrics* 198 pp. 209-230. https://doi.org/10.1016/j.jeconom.2016.12.005
[30] Ponomarenko V, Malyarets L and Dorokhov A (2011). Obespechenie kontrolya logisticheskoyj deyatelnosti s minimizatsiey logisticheskih zatrat. *Izvestija IGEA* pp. 137-142 (in Russian)
[31] Malyarets L (2011). *Matematychni metody v suchasnyh tkonomichnih doslidzhennyah* KhNEU im. S Kuznetz: Kharkiv (in Ukrainian)
[32] Kovaleva E (2015). Regressionnaya model sebestoimosti elektronnyh multimedijnih izdaniy *Vestnik NTU KhPI. Mehaniko-tehnologichni sistemy i kompleksy* pp. 55-60 (in Russian)
[33] Malyarets L (2016). *Matematychni metody i modeli v upravlinni ekonomichnymy protsesamy* KhNEU im. S Kuznetz: Kharkiv (in Ukrainian)
[34] Degtyareva T, Buresh O and Chepasov V (2003). Statisticheskiy analiz transportnogo kompleksa regiona na osnove regressionnyh modelej. *Voprosy statistiki* pp. 65-67 (in Russian)
[35] Jacob J and Lamari M (2012). Factors influencing research productivity in higher education: an empirical investigation *Foresight* 6 pp. 40-50
[36] Krasnobokaya I (2011). Analiz formirovaniya sebestoimosti produktsii proizvodstvennogo predprijatiya s ispolzovaniem mnogofaktornyh ekonometricheskih modeley *Ekonomicheskiy analiz: teoriya i praktika* pp. 38-47 (in Russian)
[37] Chao J, Hausman J, Newey W, Swanson N and Woutersen T (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity *Journal of econometrics* 178 pp. 15-21. https://doi.org/10.1016/j.jeconom.2013.08.003
[38] Bekker P and Crudu F (2015). Jackknife instrumental variable estimation with heteroskedasticity *Journal of econometrics* 185 pp. 332-342. https://doi.org/10.1016/j.jeconom.2014.08.012
[39] Cavalierea G, Nielsen M and Taylor A (2015). Bootstrap score tests for fractional integration in heteroskedastic ARFIMA models, with an application to price dynamics in commodity spot and futures markets *Journal of econometrics* 187 pp. 557-579. https://doi.org/10.1016/j.jeconom.2015.02.039