Al- Azhar University – Gaza
Faculty of Economics and Administrative sciences
Department of Applied Statistics

A Study on the Violation of Homoskedasticity Assumption in Linear Regression Models

دراسة حول انتهاك فرضية ثبات التباين في نماذج الانحدار الخطي

Prepared by

Mohammed Marwan T Barbakh

Supervised by

Dr. Samir Khaled Safi

Associate Professor of Statistics

A thesis submitted in partial fulfillment of requirements for the degree of Master of Applied Statistics

June -2012

To my parents
To my grandmother
To my wife
To my sons, daughter
To my brothers and sisters

Acknowledgements

I would like to thank so much in the beginning Allah and then Al-Azhar University of Gaza for offering me to get the Master Degree, and my thank to my professors, my colleagues at the Department of Applied Statistics.

I am grateful to my supervisor Dr. Samir Safi for his guidance and efforts with me during this study. Also, I would like to thank Dr.Abdalla El-Habil and Dr. Hazem El-Sheikh Ahmed for their useful suggestions and discussions.

My sincere gratitude to my family, especially my parents for their love and support and to my wife. I am also thankful to all my friends for their kind advice, and encouragement. Finally, I pray to Allah to accept this work.

Abstract

The purpose of this study was Regression analysis is used in many areas such as economic, social, financial and others. Regression model describes the relationship between the dependent variable and one or more of the independent variables. This study aims to present one of the most important problem that affects the accuracy of standard error of the parameters estimates of the linear regression models , such problem is called non-constant variance (Heteroskedasticity).

In the classical linear regression model it is standard to assume that given any value of the explanatory variables the disturbances are uncorrelated with zero mean, and constant variance. According to the Gauss-Markov Theorem, the Ordinary Least Squares (OLS) estimator provides the Best Linear Unbiased Estimator (BLUE) of β_i.

On the other hand, if we use the OLS estimator when the assumption of constant variance is violated, then our inferences about the parameter estimate will be incorrect. So, the standard OLS variance estimator will be biased estimator, consequently, the usual inference procedures based on T and F tests are no longer appropriate. The most serious implications of Heteroskedasticity is not only the resulting inefficiency of OLS but the misleading inference when standard tests are used.

In this study, we introduce the nature of the problem of Heteroskedasticity, the consequences of the problem, we also introduce several methods for the problem diagnostics. In addition, we present the most common remedies for the problem. We try to find the best way to detect and find remedy for this problem through various statistical techniques.

For practical case, we used real data on household consumer in the Gaza Strip in 2011. The dependent variable is the monthly consumption of the household, and the two independent variables are monthly income and the borrow of the household.

The most important finding for this study can be split into two issues. The first: for the detection tests, we found that the Goldfeld – Quandt test is the most significant one because it's keep the conditions of OLS method, then Park test and the Glejser test for large sample sizes. Secondly, for the remedy methods, we found that the redefining variable method is the best one for our data.

الملخص

يستخدم تحليل الانحدار في العديد من المجالات الاقتصادية والاجتماعية والمالية وغيرها. نموذج الانحدار يصف العلاقة بين متغير تابع وأخر مستقل أو أكثر. هذه الدراسة تهدف إلى تناول أهم المشاكل التي تؤثر على دقة تقدير الخطأ المعياري لمعالم نموذج الانحدار الخطي وهي مشكلة عدم ثبات التباين.

عادة في نماذج الانحدار الكلاسيكية (التقليدية) نفترض أن الأخطاء العشوائية لأي قيمة من قيم المتغيرات المستقلة غير مرتبطة ، متوسطها صفراً ولها تباين ثابت.

بفرض أن فرضيات الانحدار العامة متحققة فإن طريقة المربعات الصغرى العادية تعطي تقديرات مثالية.

وطبقاً لنظرية جاوس–ماركوف فإن مقدرات هذه الطريقة تعطي أفضل مقدر خطي غير متحيز لمعاملات الانحدار.

في المقابل، إذا تم استخدام هذه الطريقة في حالة انتهاك فرضية ثبات التباين، فإن الاستدلال الإحصائي للمقدرات سيكون خاطئاً. وذلك لأن تباين طريقة المربعات الصغرى يكون متحيزاً، وبناء على ذلك فإن طرق الاستدلال الإحصائي المعتمدة على اختبارات T و F تكون غير مناسبة.

في هذه الدراسة، تم عرض طبيعة مشكلة عدم ثبات التباين، الآثار المترتبة عليها، وذلك من خلال عدة طرق لاكتشاف وعلاج هذه المشكلة وذلك باستخدام أساليب إحصائية مختلفة.

وقد تم استخدام بيانات مقطعية تتعلق بحجم الاستهلاك الشهري للأسر في قطاع غزة لسنة 2011 كمتغير تابع، وكلاً من الدخل الشهري والديون المتراكمة على الأسرة كمتغيرين مستقلين

من أهم النتائج التي تم التوصل إليها هو أن اختبار Goldfeld – Quandt يعتبر أفضل اختبار لاكتشاف مشكلة عدم ثبات التباين لأنه يحافظ على شروط طريقة المربعات الصغرى ثم كلاً من اختباري Park، Glejser في حالة العينات الكبيرة. أما بالنسبة لعلاج هذه المشكلة فقد توصلنا إلى أن طريقة إعادة تعريف المتغيرات هي الأفضل .

ABBREVIATIONS

AR	Autoregressive
ANOVA	Analysis of Variance
BLUE	Best Linear Unbiased Estimator
CI	Confidence Interval
CLT	Central Limit Theorem
DW	Durbin – Watson
GLM	Generalized Linear Model
GLS	Generalized Least Square
HC	Heteroskedasticity Corrected
HCSE	Heteroskedasticity Corrected Standard Error
MLR	Multiple Linear Regression
MULR	Multiple Linear Regression
SST	Total Sum of Squares
SSR	Regression Sum of Squares
SSE	Error Sum of Squares
SLR	Simple Linear Regression
RF	Redefining Variables
OLS	Ordinary Least Square
VIF	Variance Inflation Factor
WLS	Weight Least Square

Contents

Chapter 1

1.1 Introduction

Regression model is one of the most tools and methods in the process of statistical analysis. It is concerned with describing and evaluating the relationship between a variable called the dependent variable and one or more other known variables are called independent variables. The regression model has a good predictive ability by estimating the coefficient using the least squares method and investigating of the assumptions : linearity, constant variance (Homoskedasticity), normality, and the independence of the disturbances. When one of these assumptions is violated, the classical tests such as T and F are no longer appropriate. In this study we are going to study one of the assumptions in the regression model which is the constant variance (Homoskedasticity). and then going to discuss some of the most common and appropriate detection and remedies methods in order to diagnose and solve the problem of (Heteroskedasticity) to get a good model for prediction.

1.2 Detection Methods:

There are two methods that will be used for detecting (diagnosing) the problem.

1.2.1 Evidentiary Methods.(Formal Methods)

This method used several statistical tests:

- Levene's Test
- Park Test
- Glejser Test
- Goldfeld-Quandt Test
- Breusch-Pagan Test
- White's General Heteroskedasticity Test

1.2.2 Remedy Method

- Weighted Least Squares
- White's Heteroskedasticity-Consistent Variances and Standard Errors
- Redefining the Variable

1.3 Research problem:

The research problem is to detect and find a remedy the assumption of Heteroskedasticity in the regression models by using different statistical techniques.

1.4 Research Objective :

1-Building regression model and examining the classical assumptions of the regression model.

2- Indicate causes of Heteroskedasticity

3-Discuss the consequences of Heteroskedasticity

4-Introduce the methods of detecting of Heteroskedasticity

5-Indicate the methods of remedies (solutions) of Heteroskedasticity

1.5 Literature review

Andrew F. Hayes (2009), This Study focuses on investigating the Homoskedasticity as an important assumption in ordinary least squares (OLS) regression. Although the estimator of the regression parameters in OLS regression is unbiased when the homoskedasticity assumption is violated, the estimator of the covariance matrix of the parameter estimates can be biased and inconsistent under heteroskedasticity, which can produce significance tests and confidence intervals that can be liberal or conservative.

After a brief description of heteroskedasticity and its effects on inference in OLS regression, we discuss a family of heteroskedasticity-consistent standard error estimators for OLS regression and argue investigators should routinely use one of these estimators when conducting hypothesis tests using OLS.

Mario Francisco· Juan M. Vilar (2007), This Study focuses on two new tests for heteroskedasticity in nonparametric regression are presented and compared. The first of these tests consists in first estimating non parametrically the unknown conditional variance function and then using a classical least-squares test for a general linear model to test whether this function is a constant. The second test is based on using an overall distance between a nonparametric estimator of the conditional variance function and a parametric estimator of the variance of the model under the assumption of homoscedasticity.

A bootstrap algorithm is used to approximate the distribution of this test statistic. Extended versions of both procedures in two directions, first, in the context of dependent data, and second, in the case of testing if the variance function is a polynomial of a certain degree, are also described. A broad simulation study is carried out to illustrate the finite sample performance of both tests when the observations are independent and when they are dependent.

Xu Zheng (2009), This paper presents new nonparametric tests for heterokedasticity in nonlinear and nonparametric regression models. The tests have an asymptotic standard normal distribution under the null hypothesis of homoscedasticity and are robust against any form of heteroskedasticity. Amonte Carlo simulation with critical values obtained from the wild bootstrap procedure is provided to asses the finite sample performances of the tests. A real application of testing interest rate volatility functions illustrates the usefulness of the tests proposed.

Muhammad Aslam and Gulam Rasool Pasha (200), This Study focuses on the estimation of linear regression models in the presence of heteroskedasticity of unknown form, method of ordinary least squares does not provide the estimates with the smallest variances.

In this situation, adaptive estimators are used, namely, nonparametric kernel estimator and nearest neighbour regression estimator. But these estimators rely on substantially restrictive conditions. In order to have accurate inferences in the presence of heteroskedasticity of unknown form, it is a usual practice to use heteroskedasticity consistent covariance matrix (HCCME). Following the idea behind the construction of HCCME, Them formulate a new estimator. The Monte Carlo results show the encouraging performance of the proposed estimator in the

sense of efficiency while comparing it with the available adaptive estimators especially in small samples that makes it more attractive in practical situations.

Leonor Ayyangar (2007) This paper briefly describes the assumptions of the OLS regression model. SAS/STAT Version 9.1 procedures that can be employed to test these assumptions are described and illustrated by sample codes. The consequences of violating these assumptions are enumerated. Finally, solutions are recommended. PROC REG is a useful tool for detecting violations of the assumptions of an OLS regression model. It can output information into PROC PLOT to develop graphs that are useful in the detection of data disturbances. In our health care cost studies, we often see violations of one or more assumptions. As such, when a decision is made to use OLS regression models, we employ a combination of the solutions described above. Note that when several violations occur, the use of OLS regression models for health care cost data is difficult to justify. However, an OLS model is a good beginning model because it is easy to understand. Furthermore, with large samples, we can employ the Central Limit Theorem to justify our choice. It is important to know what the model's purpose is – is it descriptive or predictive? OLS regression models for CABG surgery data have performed as well as other models in identifying and examining the impact of factors associated with cost (Austin, et al). OLS models have the advantage of simplicity and clarity as well as being easy to code. These models, however, do not perform as well as other models in predicting cost for future patients. PROC ROBUSTREG, a new procedure in SAS was developed to handle data with outliers, and can be useful for OLS regression modeling of cost data.

James H. Stock and Mark W. Watson (2006), The conventional heteroskedasticity-robust (HR) variance matrix estimator for cross-sectional regression (with or without a degrees of freedom adjustment), applied to the fixed effects estimator for panel data with serially uncorrelated errors, is inconsistent if the number of time periods T is fixed (and greater than two) as the number of entities n increases. Them provide a bias-adjusted HR estimator that is $\sqrt{nT}$ -consistent under any sequences (n,T) in which n and/or T increase to ∞. This estimator can be extended to handle serial correlation of fixed order.

O. Baser (2007) Log models are widely used to deal with skewed outcome such as health expenditure. They improve the precision of the estimates and diminish the influence of outliers. Retransformation is generally required after estimation and the evidence of heteroskedasticity complicates the process. Smearing estimation suggested in the literature only works for homoskedastic errors or heteroskedastic errors due to categorical variables. Generalized linear models have been proposed as an alternative approach for log models when there exists unknown forms of heteroskedasticity. Recent literature shows that log models are superior to generalized linear models under certain conditions. Them present a method for applying transformation that accounts for any form of heteroskedasticity. Our proposed model assumes that errors achieve normality. Hetersokedasticity is modeled seperately. Simulation studies are conducted.We also used the Medstat MarketScan Database to estimate healthcare costs for asthma patients. Finally, a comparison of the method with smearing estimators and generalized linear model (GLM) estimators is established. Log-transformed health care costs of asthma patients were normal. There was an evidence of heteroskedasticity. The simulation study, heteroskedasiticity adjusted retransformed costs had the lowest mean squared error relative to estimators from smearing retransformation or generalized linear model. This study shows that if log-transformed costs are normally distributed, heteroskedasticity adjusted retransformation produces more efficient results.

Donald W. K. Andrews and Patrik Guggenberger (2011), This paper introduces a new confidence interval (CI) for the autoregressive parameter (AR) in an AR(1) model that allows for conditional heteroskedasticity of general form and AR parameters that are less than or equal to unity. The CI is a modification of Mikusheva's (2007a) modification of Stock's (1991) CI that employs the least squares estimator and a heteroskedasticity-robust variance estimator. The CI is shown to have correct asymptotic size and to be asymptotically similar (in a uniform sense). It does not require any tuning parameters. No existing procedures have these properties. Monte Carlo simulations show that the CI performs well in finite samples in terms of coverage probability and average length, for innovations with and without conditional heteroskedasticity.

Joris Pinkse (2006) This paper provides a nonparametric method of correcting for heteroskedasticity in linear regression models with independent and identically distributed (i.i.d.) observations. The new estimator requires an empiricist to select a small set (or index) of variables which are deemed to be the most important in explaining the presence of heteroskedasticity. The new estimator is the most efficient estimator in a wide class of estimators for which the heteroskedasticity correction can only depend on the variables chosen. The nonparametric correction uses $k-$nearest neighbor (KNN) estimation.

Hausman, Newey, Woutersen, Chao, and Swanson (2009) This paper gives a relatively simple, well behaved solution to the problem of many instruments in heteroskedastic data. Such settings are common in microeconometric applications where many instruments are used to improve efficiency and allowance for heteroskedasticity is generally important. The solution is a Fuller (1977) like estimator and standard errors that are robust to heteroskedasticity and many instruments. We show that the estimator has finite moments and high asymptotic efficiency in a range of cases. The standard errors are easy to compute, being likeWhite's (1982), with additional terms that account for many instruments. They are consistent under standard, many instrument, and many weak instrument asymptotics. Based on a series of Monte Carlo experiments, we find that the estimators perform as well as LIML or Fuller (1977) under homoskedasticity, and have much lower bias and dispersion under heteroskedasticity, in nearly all cases considered.

Andreea Halunga,Chris D. Orme andTakashi Yamagata (2011) This paper proposes a heteroskedasticity-robust Breusch-Pagan test of the null hypothesis of zero cross-section (or contemporaneous) correlation in linear panel data models. The procedure allows for either.xed, strictly exogenous and/or lagged de-pendent regressor variables, as well as quite general forms of both non-normality and heteroskedasticity in the error distribution. Whilst the asymptotic validity of the test procedure, under the null, is predicated on the number of time series observations, T, being large relative to the number of cross-section units, N, independence of the cross-sections is not assumed. Across a variety of experimental designs, a Monte Carlo study suggests that, in general (but not always), the predictions from

asymptotic the-ory provide a good guide to the finite sample behavior of the test. In particular, with skewed errors and/or when N/T is not small, discrepancies can occur. However, for all the experimental designs, any one of three asymptotically valid wild bootstrap approximations (that are considered in this paper) gives very close agreement between the nominal and empirical significance levels of the test. Moreover, in comparison with wild bootstrap.the original Breusch-Pagan test (Godfrey and Yamagata, 2011) the corresponding version of the heteroskedasticity-robust Breusch-Pagan test is more reliable. As an illustration, the proposed tests are applied to a dynamic growth model for a panel of 20 countries.

Timo Teräsvirta(2011) This paper contains a brief survey of nonlinear models of autoregressive conditional heteroskedasticity. The models in question are parametric nonlinear extensions of the original model by Engle (1982). After presenting the individual models, linearity testing and parameter estimation are discussed. Forecasting volatility with nonlinear models is considered. Finally, parametric nonlinear models based on multiplicative decomposition of the variance receive attention.

1.6 Summary:

In the present chapter, we introduced the problem of the violation of *homoskedasticity*, We mentioned the detection methods of this problem by using plots and tests, then the remedy methods to got rid-off this problem. Finally, we reviewed the basic literature related to this subject

Chapter 2
Regression Models

2.1 Introduction

Regression analysis is a statistical methodology that utilize the relation between two or more quantitative variable so that a response or outcome variable can be predicted from the other, or other. This methodology is widely used in business, the social and behavioral sciences, the biological sciences, and many other disciplines. In part I we take up regression analysis when a single predictor variable is used for predicting the response or outcome variable of interest. In this part II, we consider regression analysis when two or more variables are used for making predictions. In this chapter, we consider the basic ideas of regression analysis and discuss the estimation of the parameters of regression models containing a single predictor variable and two or more variables .

2.2 The Simple Regression Model

The simple regression model can be used to study the relationship between two variables. For reasons we will see, the simple regression model has limitations as a general tool for empirical analysis (Wooldridge, 2003).

2.2.1 Definition of The Simple Regression Model

Much of applied econometric analysis begins with the following premise: y and x are two variables, representating some population, and we are interested in "explaining y in terms of x," or in "studying how y varies with changes in x". We can resolve these ambiguities by writing down an equation relating y to x. A simple equation is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.1}$$

where ε is the error term (Wooldridge, 2003).

This implies:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned} \tag{2.2}$$

$$Y_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} X_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{bmatrix} \beta_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \tag{2.3}$$

Now we write (2.2) in matrix terms compactly as follows :

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times 2}{\boldsymbol{X}}\ \underset{2\times 1}{\boldsymbol{B}} + \underset{n\times 1}{\varepsilon} \tag{2.4}$$

Since :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \tag{2.5}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ . \\ . \\ . \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ . \\ . \\ . \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix} \tag{2.6}$$

Note that $X\beta$ is the vector of the expected values of the Y_i observations since $E(Y_i) = \beta_0 + \beta_1 X_1$; hence :

$$\underset{n\times 1}{E(Y)} = \underset{n\times 1}{X\beta} \tag{2.7}$$

Where $E(Y_i)$ is defined in

$$E(Y)_{n\times 1} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \cdot \\ \cdot \\ \cdot \\ E(Y_n) \end{bmatrix} \tag{2.8}$$

The column of 1s in the X matrix may be viewed as consisting of the constant $X_0 \equiv 1$ in the alternative regression model

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i \qquad \text{where } X_0 \equiv 1$$

Thus, the X matrix may be considered to contain a column vector consisting of ls and another column vector consisting of the predictor variable observations X_i (Kutnur et. al., 2004).Regression model assume that $E\{\varepsilon_i\} = 0$, $\sigma^2\{\varepsilon_i\} = \sigma^2$, and that the ε_i are independent normal random variables. The condition $E\{\varepsilon_i\} = 0$ in matrix term is:

$$E\{\varepsilon_i\}_{n\times 1} = 0_{n\times 1} \tag{2.9}$$

Since states:

$$E(\varepsilon)_{n\times 1} = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \cdot \\ \cdot \\ \cdot \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \tag{2.10}$$

The condition that the error terms have constant variance σ^2 and that covariances $\sigma\{\varepsilon_i, \varepsilon_j\}$ for $i \neq j$ are zero (since the ε_i are independent) is expressed in matrix terms through the variance- covariance matrix of the error terms:

$$\sigma^2 \underset{n\times n}{\{\varepsilon\}} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{2.11}$$

Since this a scalar matrix, we know form the earlier example that it can be expressed in the following simple fashion:

$$\underset{n\times n}{\sigma^2\{\varepsilon\}} = \underset{n\times n}{\sigma^2 I} \tag{2.12}$$

Thus, the normal error regression model in matrix term is :

$$Y = X\beta + \varepsilon \tag{2.13}$$

where ε is a vector of independent normal random variables with $E\{\varepsilon_i\} = 0$ and

$$\sigma^2\{\varepsilon\} = \sigma^2 I \tag{2.14}$$

2.2.2 Least Squares Estimation of Regression Parameters:

The normal equation in matrix term are :

$$\underset{2\times 2}{X'X}\ \underset{2\times 1}{b} = \underset{2\times 1}{X'Y} \tag{2.15}$$

Where b is the vector of the least squares regression coefficients :

$$\underset{2\times 1}{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \tag{2.16}$$

To see this recall that we obtained $X'X$ and $X'Y$ Equation thus state :

$$\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \tag{2.17}$$

or

$$\begin{bmatrix} nb_0 + b_1 \sum X_i \\ b_0 \sum X_i + b_1 \sum X_i^2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \tag{2.18}$$

These are precisely the normal equation(Kutner et. al., 2004).

2.2.3 Estimated Regression Coefficients:

To obtain the estimated regression coefficient from the normal equations by matrix methods, we premultiply both sides by the inverse of $X'X$ (we assume this exists)

$$(X'X)^{-1}X'Xb = (X'X)^{-1}X'Y \tag{2.19}$$

We then find, since $(X'X)^{-1}X'X = I$ and $Ib = b$:

$$\underset{2\times 1}{b} = \underset{2\times 2}{(X'X)^{-1}}\,\underset{2\times 1}{X'Y} \tag{2.20}$$

The estimators b_0 and b_1 in b are the same as those given earlier in (2.20).

2..2.4 Fitted Values and Residuals

Let the vector of the fitted values $\hat{Y}_i$ be denoted by $\hat{Y}$:

$$\underset{n\times 1}{\hat{Y}} = \begin{bmatrix} \hat{Y}_i \\ \hat{Y}_i \\ \vdots \\ \hat{Y}_i \end{bmatrix} \tag{2.21}$$

In matrix notation, we then have :

$$\underset{n\times 1}{\hat{Y}} = \underset{n\times 2}{X}\,\underset{2\times 1}{b} \tag{2.22}$$

Because:

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1X_1 \\ b_0 + b_2X_2 \\ \vdots \\ b_0 + b_nX_n \end{bmatrix} \tag{2.23}$$

- **Residuals:**

Let the vector of the residuals $\varepsilon_i = Y_i - \hat{Y}_i$ be denoted by ε:

$$\underset{n\times 1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.24}$$

In matrix notation, we then have:

$$\underset{n\times 1}{\varepsilon} = \underset{n\times 1}{Y} - \underset{n\times 1}{\hat{Y}} = \underset{n\times 1}{Y} - \underset{n\times 1}{Xb} \tag{2.25}$$

2.2.5 Analysis of Variance :

We begin with the total sum of squares SST. It will be convenient to use an algebraically equivalent expression :

$$SST = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \tag{2.26}$$

We know from (2.26) that:

$$Y'Y = \sum Y_i^2 \tag{2.27}$$

The subtraction term $\frac{(\sum Y_i)^2}{n}$ in matrix from use J, as follows:

$$J = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

$$\frac{(\sum Y_i)^2}{n} = \left(\frac{1}{n}\right) Y'JY \tag{2.28}$$

Hence if follows that :

$$SST = Y'Y - \left(\frac{1}{n}\right) Y'JY \tag{2.29}$$

$SSE = \sum \varepsilon_i^2 = \sum (Y_i - \hat{Y}_i)^2$ can be represented as follows:

$$SSE = \varepsilon'\varepsilon = (Y - Xb)'(Y - Xb) \tag{2.30}$$

Which can be shown to equal:

$$SSE = Y'Y - b'X'Y \tag{2.31}$$

Finally, it can be shown that:

$$SSR = b'X'Y - \left(\frac{1}{n}\right)Y'JY \tag{2.32}$$

2.2.6 Inferences in Regression Analysis

- **Regression Coefficients**

The variance –covariance matrix of b:

$$\underset{2\times 2}{\sigma^2\{b\}} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} \end{bmatrix} \tag{2.33}$$

That is:

$$\underset{2\times 2}{\sigma^2\{b\}} = \sigma(X'X)^{-1} \tag{2.34}$$

$$\underset{2\times 2}{\sigma^2\{b\}} = \begin{bmatrix} \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\Sigma(X_i - \bar{X})^2} & \frac{-\bar{X}^2\sigma^2}{\Sigma(X_i - \bar{X})^2} \\ \frac{-\bar{X}^2\sigma^2}{\Sigma(X_i - \bar{X})^2} & \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} \end{bmatrix} \tag{2.35}$$

When MSE is substituted for σ^2 in (2.35), we obtain the estimated variance-covariance matrix of b, denoted by $S^2\{b\}$:

$$\underset{2\times 2}{S^2\{b\}} = MSE(X'X)^{-1} = \begin{bmatrix} \frac{MSE}{n} + \frac{\bar{X}^2 MSE}{\Sigma(X_i - \bar{X})^2} & \frac{-\bar{X}^2 MSE}{\Sigma(X_i - \bar{X})^2} \\ \frac{-\bar{X}^2 MSE}{\Sigma(X_i - \bar{X})^2} & \frac{MSE}{\Sigma(X_i - \bar{X})^2} \end{bmatrix} \tag{2.36}$$

2.2.7 Prediction of New Observation

The estimated variance $S^2\{pred\}$:

$$S^2\{pred\} = MSE(1 + X_h'(X'X)^{-1}X_h) \tag{2.37}$$

2.3 General Linear Regression Model in matrix Terms:

To express general linear regression model :

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots\cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \tag{2.38}$$

In matrix terms, we need to define the following matrices :

$$\underset{n\times 1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad , \quad \underset{n\times p}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

In matrix terms, the general linear regression model is :

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{n\times p}{\beta} + \underset{n\times 1}{\varepsilon} \tag{2.39}$$

Where

Y : is a vector of responses

β : is a vector of parameters

X : is a vector of constants

ε : is a vector of independent normal random variables with expectation

$E\{\varepsilon\} = 0$ and variance-covariance matrix :

$$\underset{n\times n}{\sigma^2\{\varepsilon\}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I \tag{2.40}$$

Consequently, the random vector Y has expectation:

$$\underset{n\times n}{E\{Y\}} = X\beta \tag{2.41}$$

And the variance-covariance matrix of Y is the same as that of ε :

$$\underset{n\times n}{\sigma^2\{\varepsilon\}} = X\beta \tag{2.42}$$

TABLE 2.1 ANOVA Table for General Linear Regression Model

Source Variance	*SS*	*df*	*MS*
Regression	$SSR = b'X'Y - \left(\frac{1}{n}\right)Y'JY$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = Y'Y - b'X'Y$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SST = Y'Y - \left(\frac{1}{n}\right)Y'JY$	$n-1$	

To test there is a regression relation between the response variable Y and the set of X variables $X_1 \cdots\cdots X_{p-1}$, i.e., to choose between the alternative :

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
H_a : *not all* $\beta_k (k = 1, \ldots, p-1)$ *equal zero*

We use the test statistic:

$$F^* = \frac{MSR}{MSE}$$

The decision rule to control the type I error at α is:

If $F^* \le F(1-\alpha\ ;\ p-1\ , n-p)$, conclude H_0

If $F^* > F(1-\alpha\ ;\ p-1\ , n-p)$, conclude H_a

2.3.1 Coefficient of Multiple Determination

The coefficient of multiple determination, denoted by R^2, is defined as follows :

$$R = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \tag{2.43}$$

$$0 \leq R^2 \leq 1$$

Where R^2 assumes the value 0 when all $b_k = 0 (k = 1, \cdots, p-1)$, and the value 1 when all Y observations fall directly on the fitted regression surface, i.e., when $Y = \hat{Y}_i$ for all i. Adding more X variables to the regression model can only increase R^2 and never reduce it, because SSE can never become larger with more X variables and SST is always the same for a given set of responses. Since R^2 usually came be made larger by including a larger number of predictor variables. it is sometimes suggested that a modified measure be used that adjusts for the number of X variables in model. The adjusted coefficient of multiple determination. denoted by R_a^2 adjusts R^2 by dividing each sum of square by its associated degrees of freedom :

$$R_a^2 = \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-p}{n-1}\right)\frac{SSE}{SST} \tag{2.44}$$

This adjusted coefficient of multiple determination may actually become smaller when another X variable is introduced in to the model. because any decease in SSE may be more than off set by the loss of a degree of freedom in the denominator $n - p$.

2.4 Assumptions of the Model

The assumptions of MLR relevant to estimation via OLS mirror those for SLR with perhaps a couple of exceptions(DeMaris & Alfred,2004):

- The relationship between y and the xs is linear in the parameters; that is,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \qquad for\ i = 1,2,....., n \tag{2.45}$$

- The observations are sampled independently.

- Y is approximately interval level, or binary (although the ideal procedures when Y is binary are probit or logistic regression). The X's are approximately interval-level, or dummy variables. Dummy variables are binary-coded X's that are used to represent qualitative predictors or predictors that are to be treated as qualitative.

- The $x's$ are fixed over repeated sampling. As in the case of SLR, this requirement can be waived if we are willing to make our results conditional on the observed sample values of the X's.

- $E(\varepsilon_i) = 0$ at each covariate pattern. This is the orthogonality condition.

- $V(\varepsilon_i) = \sigma^2$ at each covariate pattern.

- $Cov(\varepsilon_{i,}\varepsilon_j) = 0 \quad for\ i \neq j,$ or the errors are uncorrelated with each other.

Again, this is equivalent to assumption 2 if the data are cross-sectional.

- The errors are normally distributed.

- None of the X_k is a perfect linear combination, or weighted sum, of the other X's in the model. That is, if we regress each X_k on all of the other X's in the model, no such MULR would produce an R^2 of 1.0. Should we find an R^2 of 1.0 for the regression of one or more X's on the remaining X's, we say that there is an exact collinearity among the X's; that is, at least one of the X's is completely determined by the others.

Under certain assumptions, the minimum SSE criterion has the characteristics of unbiasedness, consistency, and efficiency these assumptions and their consequences follow:

1. If the noise term for each observation, ε, is drawn from a distribution that has a mean of zero, then the sum of squared errors criterion generates estimates that are unbiased and consistent. That is, we can imagine that for each observation in the sample, nature draws a noise term from a different probability distribution. As long as each of these distributions has a mean of zero (even if the distributions are not the same), the minimum SSE criterion is unbiased and consistent. This assumption is logically sufficient to ensure that one other condition holds namely, that each of the

explanatory variables in the model is uncorrelated with the expected value of the noise term.

2. If the distributions from which the noise terms are drawn for each observation have the same variance, and the noise terms are statistically independent of each other (so that if there is a positive noise term for one observation, for example, there is no reason to expect a positive or negative noise term for any other observation), then the sum of squared errors criterion

gives us the best or most efficient estimates available from any linear estimator (defined as an estimator that computes the parameter estimates as a linear function of the noise term, which the SSE criterion does). If assumptions (2) are violated, the SSE criterion remains unbiased and consistent but it is possible to reduce the variance of the estimator by taking account of what we know about the noise term. For example, if we know that the variance of the distribution from which the noise term is drawn is bigger for certain observations, then the size of the noise term for those observations is likely to be larger. And, because the noise is larger, we will want to give those observations less weight in our analysis. The statistical procedure for dealing with this sort of problem is termed "generalized least squares,".

2.5 Summary:

In this chapter, we discussed the simple and multiple linear regression with their formula equations, estimation method using OLS method, fitted values equations and the definitions of residuals using specific equations, ANOVA regression, In addition we introduced the assumptions of the linear regression model.

Chapter 3

Heteroskedasticity

3.1 Introduction

Heteroskedasticity is the violation of classical assumption, which states that the observations of the error term are drawn from a distribution that has a constant variance. The assumption of constant for different observations of the error term (Homoskedasticity) is not always realistic. In general, heteroskedasticity is more likely to take place in cross- sectional models than in time series models. The focus on cross- sectional models is not to say that heteroskedasticity in time series models is impossible, though (McCulloch, 1985).

Heteroskedasticity, like serial correlation, can be divided into pure and impure version.

Pure heteroskedasticity is caused by the error term of the correctly specified equation; impure heteroskedasticity is caused by a specification error such as an omitted variable.

Definitions 3.1 Pure Heteroskedasticity:

Pure heteroskedasticity refers to heteroskedasticity that is a function of the error term of a correctly specified regression equation. Such pure heteroskedasticity occurs when the variance of the error term is violated in correctly specified equation, assumption of homoskedasticity assumes that:

$$VAR(\varepsilon_i) = \sigma^2 \qquad \textbf{for } i = 1,2,.....,n \tag{3.1}$$

If this assumption is met, all the observations of the error term can be though of as being drawn from the same distribution. A distribution with a mean of zero and a variance of σ^2. This σ^2 does not change for different observations of error term ; this property is called homoskedasticity.

With heteroskedasticity, this error term variance is not constant ; instead, the variance of the distribution of the error term depends on exactly which observation is being discussed:

$$VAR(\varepsilon_i) = \sigma^2 \qquad \textbf{for } i = 1,2,.....,n \tag{3.2}$$

In this model of heteroskedasticity, the variance of the error term is related to an exogenous variable Z_i for a typical regression equation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \quad (3.3)$$

The variance of the otherwise classical error term ε might be equal to :

$$VAR(\varepsilon_i) = \sigma_i^2 Z_i^2 \quad (3.4)$$

where Z may or may not be one of the $x's$ in the equation.

The variable Z is called a proportionality factor because the variance of the error term change a proportionality to the square of Z_i. The higher the value Z_i, the higher the variance of the distribution of the i^{th} observation of the error term. There would be n different distributions one for each observation, form which the observation of the error term could be drawn depending on the number of different values that Z takes.

Definitions 3.2 Impure Heteroskedasticity:

Heteroskedasticity that is caused by an error in specification, such as an omitted variable, is referred to impure heteroskedasticity. An omitted variable can cause a heteroskedastic error term because the portion of the omitted effect not represented by one of the included explanatory variable must be absorbed by the error term. If this effect has a heteroskedastic component, the error term of the miss specified equation might be heteroskedastic even if the error term of the true equation is not. This distinction is important because with impure heteroskedasticity the correct remedy is to find the omitted variable and include it in the regression. It's therefore important to be sure that your specification is correct before trying to detect or remedy pure heteroskedasticity.

3.2 The consequences of Heteroskedasticity.

If the error term of an equation is known to be heteroskedastic, there are three major consequences:

1. Pure heteroskedasticity does not cause bias in the coefficient estimates. Even if the error term of an equation is known to be purely heteroskedastic, that heteroskedasticity will not cause bias in the OLS estimates of the coefficient. This is true because with pure heteroskedasticity, none of the independent variable is

correlated with the error term. As a result, we can say pure heteroskedasticity still has property that:

$$E(\hat{\beta}) = \beta \qquad \text{for all } \beta \text{ s}$$

Lack of bias does not guarantee " accurate " coefficient estimates, especially since heteroskedasticity increases the variance of the estimates, but the distribution of the estimates is still centered around the true β. Equation with impure heteroskedasticity caused by an omitted variable, of course, will have possible specification bias.

2. Heteroskedasticity increases the variances of the $\hat{\beta}$ distribution.. If the error term of an equation is heteroskedastic with respect to a proportionality factor Z :

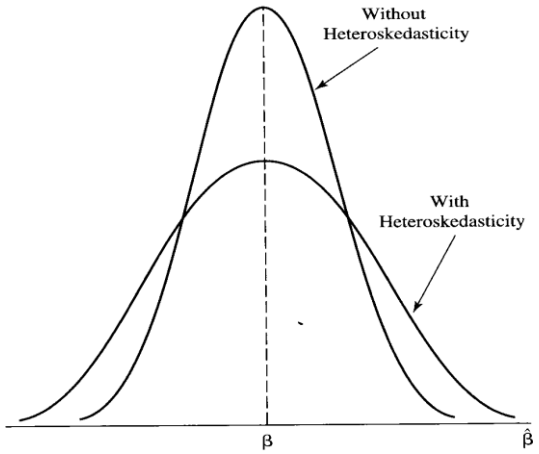
$$VAR(\varepsilon_i) = \sigma_i^2 Z_i^2 \tag{3.5}$$

. the minimum variance portion of the Gauss- Markov Theorem cannot be proven because there are other linear unbiased estimators that have smaller variances. This is because the heteroskedastic error term causes the dependent variable to fluctuate in a way that the OLS estimation procedure attributes to the independent variable. Thus, OLS is more likely to misestimate the true β in the face of heteroskedasticity. On balance, the $\hat{\beta}s$ are still unbiased because overestimates are just as likely as underestimates ; however, these error increase the variance of the distribution of the estimates, increasing the amount that any given estimate is likely to differ from the β (see figure 3.1)

Figure 3.1

Distribution of β

Without Heteroskedasticity and With Heteroskedasticity



3. Heteroskedasticity causes OLS to tend to underestimate the variance and standard error of the coefficients. As result, neither the T statistic nor the F statistic can be relied on in the face of uncorrected heteroskedasticity. In practice, OLS usually ends up with higher T scores than would be obtained if the error term were homoscedastic, some times leading researchers to reject null hypotheses that shouldn't be rejected. OLS estimator is still unbiased in the face of heteroskedasticity. The heteroskedasticity has caused the $\hat{\beta}s$ to be farther from the true value, however and so the variance of the distribution of the $\hat{\beta}s$ has increased.

3.3 Testing For Heteroskedasticity

Heteroskedasticity poses potentially severe problems for inferences based on least squares. It is useful to be able to test for homoskedasticity and if necessary, modify our estimation procedures accordingly.(Ohtani & Toyoda 1980) Several types of tests have been suggested. Most of the tests for heteroskedasticity are based on the following strategy. OLS is a consistent estimator of β even in the presence of heteroskedasticity.

Therefore, tests designed to detect heteroskedasticity will, in most cases, be applied to the ordinary least squares residuals.

3.3.1 White's General Test

To formulate most of the available tests, it is necessary to specify, at least in rough terms,
the nature of the heteroskedasticity. It would be desirable to be able to test a general hypothesis of the form

$$\begin{aligned} &H_0 : \sigma_i^2 = \sigma^2 \qquad \textit{for all } i\ , \\ &H_1 : \textit{Not } H_0 \end{aligned} \tag{3.6}$$

In view of our earlier findings on the difficulty of estimation in a model with n unknown

parameters, this is rather ambitious. Nonetheless, such a test has been devised by White 1980. The correct covariance matrix for the least squares estimator is

$$Var[b|X] = \sigma^2 [X'X]^{-1} [X'\Omega X][X'X]^{-1} \tag{3.7}$$

which, as we have seen, can be estimated using.

$$Var[b] = \frac{1}{n}\left[\frac{X'X}{n}\right]^{-1}\left[\frac{1}{n}\sum_i^n e_i^2 x_i x_i'\right]\left[\frac{X'X}{n}\right]^{-1} \tag{3.8}$$

The conventional estimator is $Var = s^2[X'X]^{-1}$. If there is no heteroskedasticity, then V will give a consistent estimator of $Var[b|X]$, whereas if there is, then it will not. White has devised a statistical test based on this observation. A simple operational version of his test is carried out by obtaining nR^2 in the regression of e_i^2 on a constant and all unique variables contained in X and all the squares and cross products of the variables in X. The statistic is asymptotically distributed as chi-squared with $P-1$ degrees of freedom, where P is the number of regressors in the equation, including the constant (Greene 1993).

The White test is extremely general. To carry it out, we need not make any specific assumptions about the nature of the heteroskedasticity. Although this characteristic is a virtue, it is, at the same time, a potentially serious shortcoming. The test may reveal heteroskedasticity, but it may instead simply identify some other specification error (such as the omission of X^2 from a simple regression). Except in the context of a specific problem, little can be said about the power of White's test;

it may be very low against some alternatives. In addition, unlike some of the other tests we shall discuss, the White test is nonconstructive. If we reject the null hypothesis, then the result of the test gives no indication of what to do next.

3.3.2 The Park Test

How do we test for pure heteroskedasticity of the form that we assumed in the previous section That form ' as we outlined in the previous section 'is:

$$VAR(\varepsilon_i) = \sigma^2 Z_i^2 \tag{3.9}$$

Where :

ε : the error term of the equation being estimated

σ^2 : the variance of the homoskedastic error term

Z : The proportionality factor

The Park test is a formal procedure that attempts to test residuals for this heteroskedasticity in a manner similar to the way that the Durbin _Watson d statistic tests residuals for serial correlation. The Park test has three basic steps. First, the regression equation is estimated by OLS and the residuals are calculated. Second, the log of the squared residuals is used as the dependent variable of an equation whose sole explanatory variable is the log of the proportionality factor Z. Finally, the results of the second regression are tested to see if there is any evidence of heteroskedasticity. If there is reason to suspect heteroskedasticity, it's appropriate to run a Park test. Since the Park test is not run automatically by computer regression package, you should known how to run the test by yourself by the following:

1. Obtain the residuals of the estimated regression equation. The first step is to estimate the equation with OLS and then find the residuals from their estimation:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \tag{3.10}$$

These residuals, which are printed out by most computer regression package, are the same ones used to calculate the Durbin –Watson d statistic to test for serial correlation.

2. Use these residuals to form the dependent variable in a second regression. In particular, the Park test suggests that you run the following double – log regression :

$$In(e_i^2) = \alpha_0 + \alpha_1 InZ_i + \varepsilon_i \qquad (3.11)$$

Where:

e_i : the residual from the *ith* observation from Equation

Z_i : your best choice as the possible proportionality factor (Z)

ε_i : a classical (homoskedastic) error term

3. Test the significance of the coefficient of Z in Equation (3.8) with a t-test. The last step is to use the t-statistic to test the significance of InZ in explaining $In(e^2)$ in equation (3.8) If the coefficient of Z is significantly different from zero, this is evidence of heteroskedastic patterns in the residuals with respect to Z ; otherwise, heteroskedasticity related to this particular Z is not supported by the evidence in these residuals. However, it's impossible to prove that a particular equation's error term are homoskedastic. The Park test is not always easy to use. Its major problem is the identification of the proportionality factor Z. Although Z is often an explanatory variable in the original regression equation, there is no guarantee of that. A particular Z should be chosen for your Park test only after investigating the type of potential heteroskedasticity in your equation. A good Z is a variable that seems likely to vary with the variance of the error term.

3.3.3 The Goldfeld – Quandt Test

By narrowing our focus somewhat, we can obtain a more powerful test. Two tests that are relatively general are the Goldfeld–Quandt (1965) test and the Breusch–Pagan (1979) Lagrange multiplier test. For the Goldfeld–Quandt test, we assume that the observations can be divided into two groups in such a way that under the hypothesis of homoskedasticity, the disturbance variances would be the same in the two groups, whereas under the alternative, the disturbance variances would differ systematically. The most favorable case for this would be the group wise heteroskedastic model of a model such as $\sigma_i^2 = \sigma^2 X_i^2$ for some variable X. By ranking the observations based on this x, we can separate the observations into

those with high and low variances. The test is applied by dividing the sample into two groups with n_1 and n_2 observations. To obtain statistically independent variance estimators, the regression is then estimated separately with the two sets of observations. The test statistic is

$$F[n_1 - k, n_2 - k] = \frac{e_1' e_1 / (n_1 - k)}{e_2' e_2 / (n_2 - k)} \tag{3.12}$$

where we assume that the disturbance variance is larger in the first sample. (If not, then reverse the subscripts.) Under the null hypothesis of homoscedasticity, this statistic has an F distribution with $n_1 - k$ and $n_2 - k$ degrees of freedom. The sample value can be referred to the standard F table to carry out the test, with a large value leading to rejection of the null hypothesis. To increase the power of the test, Goldfeld and Quandt suggest that a number of observations in the middle of the sample be omitted. The more observations that are dropped, however, the smaller the degrees of freedom for estimation in each group will be, which will tend to diminish the power of the test. As a consequence, the choice of how many central observations to drop is largely subjective. Evidence by (Harvey & Phillips,1974) suggests that no more than a third of the observations should be dropped. If the disturbances are normally distributed, then the Goldfeld–Quandt statistic is exactly distributed as F under the null hypothesis and the nominal size of the test is correct. If not, then the F distribution is only approximate and some alternative method with known large-sample properties, such as White's test, might be preferable.

3.3.4 The Breusch –Pagan/godfrey LM

The Goldfeld–Quandt test has been found to be reasonably powerful when we are able to identify correctly the variable to use in the sample separation. This requirement does limit its generality, however. For example, several of the models we will consider allow he disturbance variance to vary with a set of regressors. Breusch and Pagan have devised a Lagrange multiplier test of the hypothesis that $\sigma_i^2 = \sigma^2 f(\beta_0 + \beta' z_i)$,where z_i is a vector of independent variables. The model is homoskedastic if $\alpha = 0$. The test can be carried out with a simple regression: $LM = \frac{1}{2}$ explained sum of squares in the regression of $e_i^2 / (e'e/n)$ on z_i For

computational purposes, let Z be the $n \times P$ matrix of observations on $(1, z_i)$, and let g be the vector of observations of $g_i = e_i^2 / (e'e/n)^{-1}$. Then

$$LM = \frac{1}{2}[g'Z(Z'Z)^{-1}Z'g] \tag{3.13}$$

Under the null hypothesis of homoskedasticity, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in z_i. This test can be applied to a variety of models, (Harvey, 1976). It has been argued that the Breusch–Pagan Lagrange multiplier test is sensitive to the assumption of normality. (Koenker, 1981) and (Koenker & Bassett, 1982) suggest that the computation of LM be based on a more robust estimator of the variance of ε_i^2

$$V = \frac{1}{2}\sum_{i=1}^{n}\left[e_i^2 - \frac{e'e}{n}\right]^2 \tag{3.14}$$

The variance of ε_i^2 is not necessarily equal to $2\sigma^4$ if ε_i is not normally distributed. Let u equal $(e_1^2, e_2^2, ..., e_n^2)$ and i be an $n \times 1$ column of 1s. Then $\bar{u} = \frac{e'e}{n}$. With this change, the computation becomes

$$LM = \left[\frac{1}{V}\right](u - \bar{u}i)'Z(Z'Z)^{-1}Z'(u - \bar{u}i) \tag{3.15}$$

Under normality, this modified statistic will have the same asymptotic distribution as the Breusch–Pagan statistic, but absent normality, there is some evidence that it provides a more powerful test. (Waldman, 1983) has shown that if the variables in z_i are the same as those used for the White test described earlier, then the two tests are algebraically same.

3.3.5 Glejser Test

In least-squares analysis, the most widely used tests for the presence of heteroskedasticity examine whether the squared residuals are correlated with some other variables. The latter, in particular, focused their attention on the test proposed by Glejser (1969), which is designed to test whether the regression residuals in absolute value are correlated with some other variables. It has been an issue in the literature whether the Glejser test is valid when the error density is not symmetric by Talwar (1983).

Consider a linear model:

$$Y_i = X_i\beta + \varepsilon_i \qquad i = 1, 2....., n \tag{3.16}$$

where x_i is the $1\times k$ vector of the explanatory variables, and β is the $k\times 1$ parameter vector of interest $\{(y_i, x_i)\}: i = 1,2....., n\}$ is an independently and identically distributed random sequence. the upper case letters indicate the data matrix which stacks the observations for $i = 1,2....., n$. A set of basic assumptions necessary for least-squares analysis is

Assumption 1. $E(x_i', \varepsilon_i) = 0$.

Assumption 2. $E(x_i', x_i)$ is positive definite.

Assumption 3. (y_i, x_i). has finite fourth moment.

These assumptions are central to the least-squares analysis, ensuring the consistency and the asymptotic normality of the OLS estimator of β obtained as

$$\hat{\beta}_i = (X'X)^{-1} X'Y. \tag{3.17}$$

Heteroskedasticity is present if

$$E(x_i' x_i \varepsilon_{i)}^2) \neq \sigma^2 E(x_i' x_i), \tag{3.18}$$

where $\sigma^2 = E(\varepsilon_i^2)$.We are interested in testing the null hypothesis (Judge et al. 1985)

$$H_0 : E\left[x_i' x_i (\varepsilon_i^2 - \sigma^2)\right] = 0 \tag{3.19}$$

Therefore, the problem is whether the squared-error sequence $\left\{\varepsilon_i^2\right\}$ is correlated. Glejser(1969) proposed a test designed to see whether $|e_i|$ is correlated with z_i as a test of heteroskedasticity. The statistic is obtained as in the regression ;

$$|\hat{e}_i| = \beta_0 + Z_i X + \varepsilon \tag{3.20}$$

3.3.6 Leven's Test

The modified Levene test can be used to evaluate the validity of the assumption of constant variance. It has been shown to be reliable even when the residuals do not follow a normal distribution.

The test is constructed by grouping the residuals according to the values of X. The number of groups is arbitrary, but usually, two groups are used. In this case, the absolute residuals of observations with low values of X are compared against those with high values of X. If the variability is constant, the variability in these two groups of residuals should be equal. The test is computed using the formula

$$L = \frac{\bar{d}_1 - \bar{d}_2}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.21}$$

Where

$$S_p = \sqrt{\frac{\sum_{j=1}^{n1} (d_{j1} - \bar{d}_1)^2 + \sum_{j=1}^{n2} (d_{j2} - \bar{d}_2)^2}{n_1 + n_2 - 2}} \tag{3.22}$$

$$d_{j1} = \left| e_{j1} - \tilde{e}_1 \right|$$
$$d_{j2} = \left| e_{j2} - \tilde{e}_2 \right|$$

and $\tilde{e}_1$ is the median of the group of residuals for low values of X and $\tilde{e}_2$ is the median of the group of residuals for high values of X. The test statistic L is approximately distributed as a t statistic with $N-2$ degrees of freedom (Jerry L. Hintze &Kaysville, Utah 2007)

3.4 Remedies for Heteroskedasticity

The first thing to do if the White's General Test and Park test, Goldfeld – Quandt Test, Breusch –Pagan/godfrey LM, Glejser Test and Leven's Test indicates the possibility of heteroskedasticity is to examine the equation carefully for specification errors. Although you should never include an explanatory variable simply because a test indicates the possibility of heteroskedasticity, you ought to rigorously think through the specification of the equation. If this rethinking allows you to discover a variable that should have been in the regression from the beginning, then that variable should be added to the equation.

However, if there are no obvious specification errors, the heteroskedasticity is probably pure in nature, and one of the remedies described in this section should be considered:

1. Weighted Least Squares (WLS)
2. Heteroskedasticity – Corrected Standard Errors
3. Redefining the Variable

3.4.1 Weighted Least Squares (WLS)

Take an equation with pure heteroskedasticity caused by proportionality factor Z :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{3.22}$$

Where the variance of the error term, instead of being constant, is :

$$VAR(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2 \tag{3.23}$$

Where σ^2 is the constant variance of a classical (Homoskedastic) error term ε_i and Z_i is the proportionality factor. Given that pure heteroskedasticity exists, then equation can be shown to be equal to :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i \varepsilon_i \tag{3.24}$$

The error in equation $Z_i \varepsilon_i$, is heteroskedastic because $\sigma^2 Z_i^2$, its variance, is not constant.

The easiest method is to divide the entire equation through by the proportionality factor Z_i, resulting in an error term ε_i, that has a constant variance σ^2. The new equation satisfies the Classical Assumption, and a regression run on this new equation would no longer be expected to have heteroskedasticity. This general remedy to heteroskedasticity is called WLS, which is actually a version of GLS. WLS involves dividing equation through by whatever will make the error term homoskedastic and then rerunning the regression on the transformed variables (A.H.Studenmund,2010).

Given the commonly assumed form of heteroskedasticity in equation this means that the technique consists of three steps:

1. Dividing equation through the proportionality factor Z obtaining :

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + \varepsilon_i \tag{3.25}$$

The error term of equation is now ε_i, which is homoskedasticity

2. Recalculate the data for the variables to conform to equation

3. Estimate equation with OLS

This third step in WLS, the estimation of the transformed equation, is fairly tricky, because the exact details of how to complete this regression depend on whether the proportionality factor Z is also an explanatory variable in equation. If Z is not an explanatory variable in equation ,then the regression to be run in step 3 might seem to be :

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + \varepsilon_i \tag{3.26}$$

Note, however, that this equation has no constant term. Most OLS computer Packages can run such a regression only if the equation is forced through the origin by specifically suppressing the intercept with an instruction to the computer.

However, the omission of the constant term forces the constant effect of omitted variables, nonlinearities, and measurement error in to the other coefficient estimates. To avoid having these constant term elements forced in to the slope coefficient estimates, one alternative approach to equation is to add a constant term before the transformed equation is estimated. Consequently, when Z is not identical to one of the Xs in the original equation, then we suggest that the following specification be run as step 3 in Weighted Least Squares :

$$\frac{Y_i}{Z_i} = \alpha_0 + \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + \varepsilon_i \tag{3.27}$$

If Z is an explanatory variable in equation, then no constant term need be added because one already exists. Look again at equation. If $Z = X_1$ (or, similarly, if $Z = X_2$), then one of the slope coefficients becomes the constant term in the transformed equation because $\frac{X_1}{Z} = 1$:

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \beta_1 + \frac{\beta_2 X_{2i}}{Z_i} + \varepsilon_i \tag{3.28}$$

If this form of Weighted Least Squares is used, however, coefficient obtained from an estimation of equation must be interpreted very carefully. Notice that β_1 is now the intercept term of equation even though it is a slope coefficient in equation and that β_0 is a slope in equation, even though it is the intercept in equation. As a result,

a researcher interested in an estimate of the coefficient of X_1 in equation would have to examine the intercept of equation, and are searcher interested in an estimate of the intercept term of equation would have to examine the coefficient of $\frac{1}{Z_i}$ in equation. The computer will print out $\hat{\beta}_0$ as a " slope coefficient" and $\hat{\beta}_1$ as a " constant term " when in reality they are estimates of the opposite coefficients in the original equation.

There are three other major problems with using Weighted Least Squares:

1.The job of identifying the proportionality factor Z is, as has been pointed out, quite difficult.

2.The functional form that relates the Z factor to the variance of the error term of the original equation may not be our assumed squared function of equation. When some other functional relationship is involved, a different transformation is required. For more on these advanced transformations

3. Some times Weighted Least Squares is applied to an equation with impure heteroskedsticity. In such cases, it can be shown that the WLS estimates reduce some what the bias from an omitted variable, but the estimates are inferior to those obtained from the correctly specified equation.

3.4.2 Heteroskedasticity Corrected Standard Error

The most popular remedy for heteroskedasticity is heteroskedasticity corrected standard error, which take a completely different approach to the problem. It focuses on improving the estimation of the SE($\hat{\beta}$)s without changing the estimates of the slope coefficient. The logic behind this approach is powerful. Since heteroskedasticity causes problems with the SE($\hat{\beta}$)s but not with the $\hat{\beta}$s, it makes sense to improve the estimation of the SE($\hat{\beta}$)s in a way that doesn't alter the estimates of the slope coefficient. This differs from our other two remedies because both WLS and reformulating the equation affect the $\hat{\beta}$s as well as the SE($\hat{\beta}$)s.

Thus, heteroskedasticity correct (HC) standard error are SE($\hat{\beta}$)s that have been calculated specifically to avoid the consequences of heteroskedasticity. For a linear regression model with one independent variable in which both X and Y are

measured as deviations from the mean, the HC estimator of the variance of the estimated slope coefficient is:

$$\frac{\sum_{i=1}^{n} X_i^2 \varepsilon_i^2}{(\sum_{i=1}^{n} X_i^2)^2} \tag{3.29}$$

Where ε_i is the OLS residual for observation n

While they are biased, are generally more accurate than uncorrected standard error for large samples in the face of heteroskedasticity. As a result, the HCSE can be use in t-test and other hypothesis tests in most samples without the error of inference potentially caused by heteroskedasticity. Typically, the HCSE are greater than the OLS $\text{SE}(\hat{\beta})$s, thus producing lowest t-scores and decreasing the probability that a given estimated coefficient will be significantly different from zero (Halbert White 1980).

There are a few problems with using heteroskedasticity correct standard error. First, as mentioned, the technique works best in large samples. Second, details of the calculation of the HCSE are beyond the scope of this text and imply a model that is substantially more general than the basic theoretical construct, $VAR(\varepsilon_i) = \sigma^2 Z_i^2$.

3.4.3 Redefining the Variables

Another approach to an equation of heteroskedasticity is to go back to the basic underlying theory of the equation and redefining the variables in a way that avoids heteroskedasticity. A redefinition of the variables often is useful in allowing the estimated equation to focus more on the behavioral aspect of the relationships. Such is a rethinking is a difficult and discouraging process because it appears to dismiss all the work already done. However, once the theoretical work has been reviewed, the alternative approaches that are discovered are often exciting in that they offer possible ways to avoid problems that had previously seemed insurmountable.

In some cases, the only redefinition that's needed to rid an equation of heteroskedasticity is to switch from a linear functional form to a double-log functional form. The double-log form has inherently less variation than the linear form, so its less likely to encounter heteroskedasticity. In addition, there are many research topics for which the double-log is just as theoretically logical as the linear form. This is especially true if the linear form was chosen by default, as is often the

case. In other situation, it might be necessary to completely rethink the research project in terms of underlying theory.

3.5 Summary

In the present chapter, we discussed the tests that detect the heteroskedasticity as mentioned as follows: Levene's Test, Park Test, Glejser Test, Goldfield- Quandt Test, Cook Weisburg (or Breusch-Pagan-Godfrey) Test, and White's General Heteroskedasticity Test. Then we discussed the remedies of the Heteroskedasticity by as follows: Weighted Least Squares, Heteroskedasticity Corrected Standard Errors and Redefining the Variables.

Chapter4

Case study

4.1 Description Data :

The Data represents 50 observations for households consumption in 2011, in Gaza city. The dependent variable is the households consumption (Consumption) and the independent variables are households consumption obtaining or receiving a loan (Borrow) and the annual income of the households consumption (Income). Resources are a questioner to find out family consumption in Gaza Strip.

The previous studies showed that there exist significant positive relationship between households consumption with income and borrow.

We want to regress the households consumption on borrow and income, using the multiple linear regression model :

$$E(CONSUMPTION) = \beta_0 + \beta_1 BORROW + \beta_2 INCOME$$

4.2 Data Analysis :

4.2.1 Descriptive Data :

Table 4.1 shows the households consumption ranges between 1,300 NIS through 10,150 NIS with mean equal 3,628.40 NIS and standard deviation equal 1,705.42 and the borrow ranges between 120 NIS through 1,700 NIS with mean equal 233.40 NIS and standard deviation equal 338.06 and the income ranges between 850 NIS through 10,500 NIS with mean equal 4,083.00 NIS and standard deviation equal 2,144.89 NIS

Table (4.1) : Descriptive Statistics for the variables

Variable	Mean	Std.Dev	Min	Max
Consumption	3,628.40	1,705.42	1,300	10,150
Borrow	233.40	338.06	120	1,700
Income	4,083.00	2,144.89	850	10,500

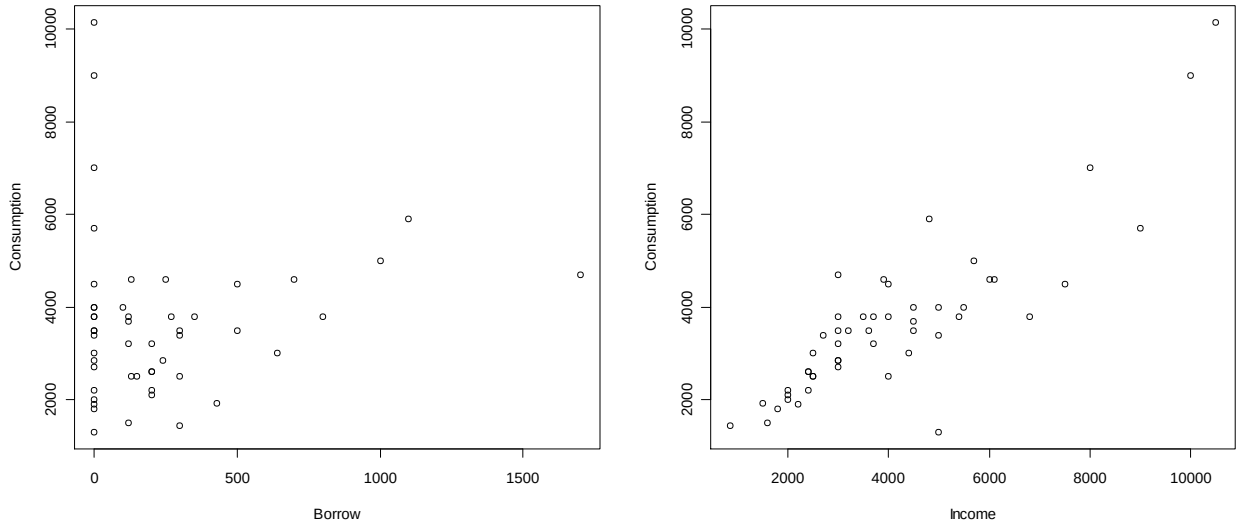
4.2.2 Scatter plots for the variables :

E-views software program is used to analyze the data, we perform statistical analysis for all independent variables and dependent variable.

Figure 4.1 represents the scatter plots for households consumption versus borrow and income respectively. The plots show there exists positive strong relationship between households consumption and borrow and income.

Figure (4.1)

scatter plot



4.2.3 Correlation coefficients

Table 4.2 shows the correlation matrix for households consumption, borrow and income. The correlation coefficient between households consumption and borrow equals 0.611 with p-value < 0.0001. Also The correlation coefficient between households consumption and income equals 0.611 with p-value < 0.0001.

We conclude there is a significant positive relationship between households consumption with each borrow and income

Table 4.2 : Correlation Matrix

family consumption	Borrow	Income
Pearson Correlation	0.611**	0.855**
Prob.	<0.0001	<0.0001

**Correlation is significant at the 0.01 level (2-tailed).

4.2.3 linear regression model result

Table 4.3 shows the linear regression results. Adjusted R-square equals 0.88 which means that 88% of variation in households consumption is explained by borrow and income and the remaining percentage 12% due to other factors, that are not included in the model, Since F=107.19 and p-value < 0.0001. Then it is significant relationship between households consumption with each borrow and income variable because p-value < 0.0001 so, the regression equation is:

$$\hat{C}onsumption = 0.778 + 0.175 * Borrow + 0.656 * Income. \quad (4.1)$$

Table 4.3 : Regression equation result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.777885	0.186771	4.164906	0.0003
BORROW	0.174900	0.029797	5.869732	0.0000
INCOME	0.656468	0.049986	13.13313	0.0000

R-squared	0.888140	Mean dependent var	3.504819
Adjusted R-squared	0.879854	S.D. dependent var	0.150939
S.E. of regression	0.052318	Akaike info criterion	-2.968294
Sum squared resid	0.073905	Schwarz criterion	-2.828174
Log likelihood	47.52441	F-statistic	107.1868
Durbin-Watson stat	1.883234	Prob(F-statistic)	0.000000

4.3 Assumptions Validation

In this section we examine the regression equation assumptions. The assumptions are normality of households consumption, heteroskedasticity, autocorrelation and multicollinearity.

4.3.1 Normality Assumption

One of the assumptions of linear regression analysis is that the residuals are normally distributed.

Figure 4.2 and 4.3 and table 4.4 show the diagnostic method for the normality of the residuals. All the plots (Histogram, Q-Q plots) suggest that the residuals are normally distributed. The tests of normality " Shapiro test" is not significant(p-value > 0.05).

Figure 4.2

histogram

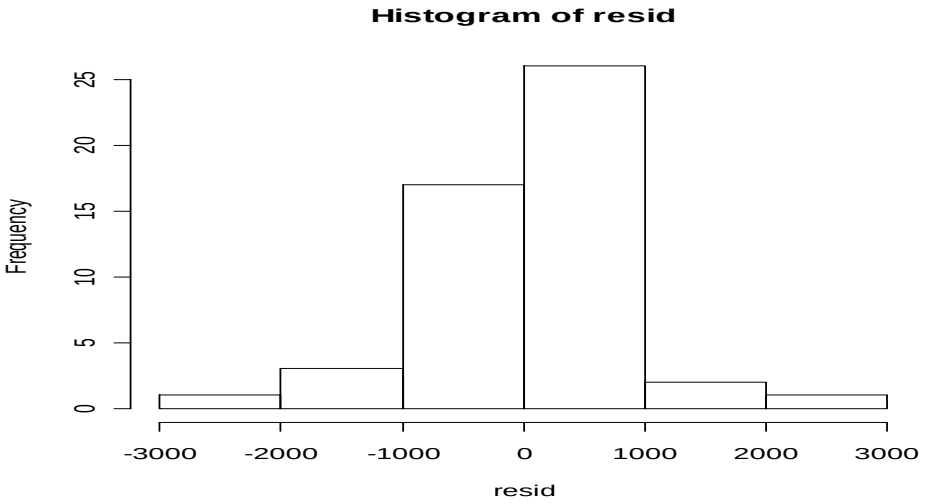


Figure 4.3
Normal Q-Q plot of Regression Standardized Residual

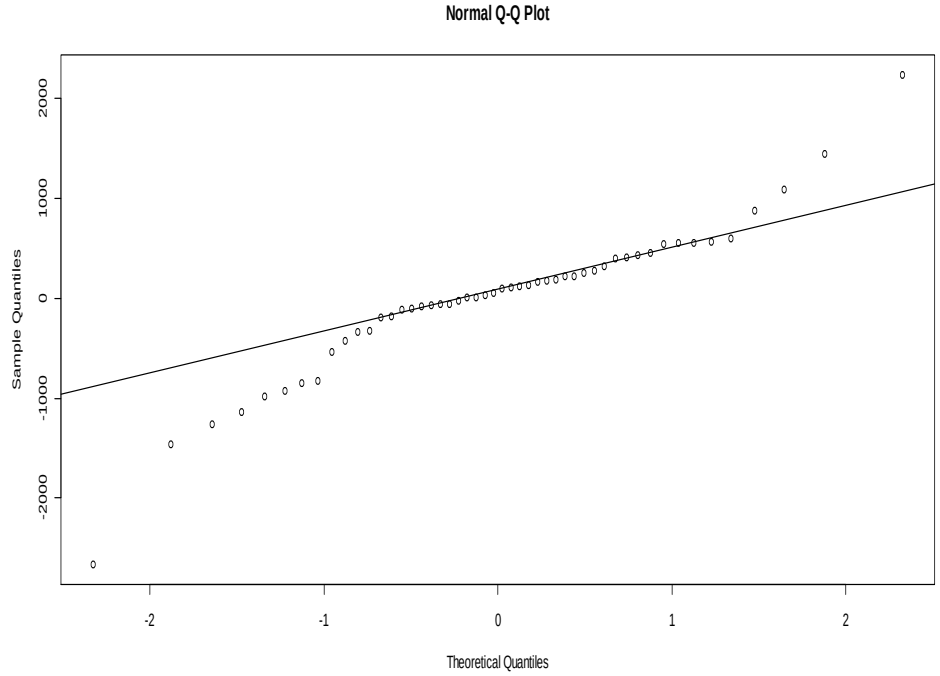


Table 4.4 : Normality Test Result

	Kolmogorov -Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.161	50	.224	.920	50	.122

a. Lilliefors Significance Correction

4.3.2 Independence of disturbances variables Assumption:

When there is a perfect linear relationship among the predictors, the estimates for a regression model cannot be uniquely computed. The term collinearity implies that two variables are near perfect linear combinations of one another. When more than two variables are involved it is often called multicollinearity, although the two terms are often used interchangeably.

The primary concern is that as the degree of multicollinearity increases, the regression model estimates of the coefficients become unstable and the standard errors for the coefficients can get wildly inflated

Table 4.5 shows the "tolerance" and *variance inflation factor* (VIF) values for each predictor as a check for multicollinearity. The "tolerance" is an indication of the percent of variance in the predictor that cannot be accounted for by the other predictors, hence very small values indicate that a predictor is redundant, and values that are less than 5 may merit further investigation. The VIF, is (1 / tolerance) and as a rule of thumb, a variable whose VIF values is greater than 5 may merit further investigation. The "tolerance" and "VIF" values are all quite acceptable because all VIFs are less than 5.

Table4.5 :VIF Results

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Tolerance	VIF
C	0.777885	0.186771	4.164906	0.0003		
BORROW	0.174900	0.029797	5.869732	0.0000	0.969	1.031
INCOME	0.656468	0.049986	13.13313	0.0000	0.969	1.031

4.3.3 Independence of Residuals Assumption:

Another assumption of OLS regression is the independence of the residuals can be broken, when data are collected on the same variables over time. This is known as autocorrelation, The Durbin-Watson (DW) statistic to test for correlated residuals.

The DW statistic has a range from 0 to 4 with a midpoint of 2. Table 4.6 shows that the observed value of DW=1.88 which is less than 2 this result is not surprising since the data is not truly time-series.

Table 4.6 : DW Results

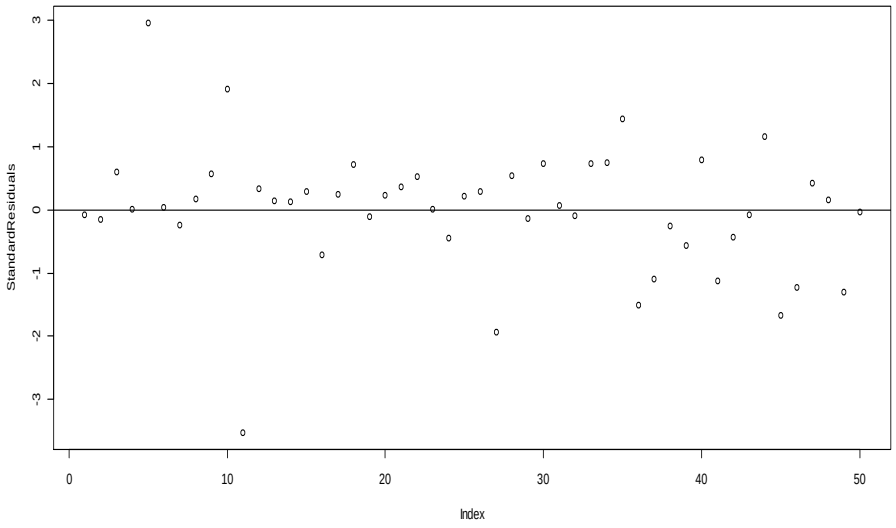
R-squared	0.888140	Mean dependent var	3.504819
Adjusted R-squared	0.879854	S.D. dependent var	0.150939
S.E. of regression	0.052318	Akaike info criterion	-2.968294
Sum squared resid	0.073905	Schwarz criterion	-2.828174
Log likelihood	47.52441	F-statistic	107.1868
Durbin-Watson stat	1.883234	Prob(F-statistic)	0.000000

4.3.4 Homoskedasticity Assumption

Another assumption of OLS regression is that the variance of the residuals is homogeneous across levels of the predicted values, also known as homoscedasticity. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant then the residual variance is said to be "heteroskedastic." Below we illustrate graphical methods for detecting heteroskedasticity. A commonly used graphical method is to use the residuals the value versus fitted (predicted) values.

Figures 4.4 shows the residuals of regression equation not scattering randomly around about zero with same pattern so, that means the residuals are not constant variance (heteroskedasticity), which means the assumption of constant variance is not violated. We perform many tests to check this problem of non-constant variance (heteroskedasticity).

Figure (4.4)
Scatter plot of the residuals



4.4 Detection Tests

In this section, use some tests to detect the Heteroskedasticity problem .

4.4.1 White test

Table 4.7 shows the probability printed to the right of the value in the E-views output for White's heteroskedasticity test equal 0.005234 represents the probability that you would be incorrect if you rejected the null hypothesis of no heteroskedasticity. The F-statistic is an omitted variable test for the joint significance of all cross products, excluding the constant. It is displayed above White's test statistic for comparison purposes

Table 4.7 : White test result

F-statistic	4.389704	Probability	0.002501
Obs*R-squared	16.64065	Probability	0.005234

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	64033.14	881343.6	0.072654	0.9424
BORROW	-716.0073	2215.297	-0.323211	0.7481
BORROW^2	0.467155	0.884630	0.528079	0.6001
BORROW*INCOME	-0.054202	0.422408	-0.128317	0.8985
INCOME	11.65159	347.5294	0.033527	0.9734
INCOME^2	0.027489	0.029677	0.926264	0.3594

4.4.2 Park test

Table 4.8 shows process of the equation between log(residuals^2) and log(income), so we have the model is significance because p-value is less than 0.0001, which means the constant variance is violated

Table 4.8 : Park test result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-18802.77	2340.602	-8.033305	<0.0001
BORROW	1.078768	0.429244	2.513185	0.0155
LOG(INCOME)	2708.546	283.5789	9.551295	<0.0001

R-squared	0.664595	Mean dependent var	3628.400
Adjusted R-squared	0.650322	S.D. dependent var	1705.416
S.E. of regression	1008.473	Akaike info criterion	16.72839
Sum squared resid	47799846	Schwarz criterion	16.84311
Log likelihood	-415.2097	F-statistic	46.56448
Durbin-Watson stat	0.894659	Prob(F-statistic)	<0.000001

4.4.3 The Goldfeld – Quandt test

We split data into three groups and remove the observation lies in the mid. (second group) and building the regression equation from the first group from 1 to 17

Table 4.9 : Goldfeld – Quandt test

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	214.0011	292.9767	0.730437	0.4772
BORROW	0.869758	0.365079	2.382383	0.0319
INCOME	0.889126	0.120383	7.385828	0.0000
R-squared	0.796156	Mean dependent var		2307.647
Adjusted R-squared	0.767036	S.D. dependent var		529.2038
S.E. of regression	255.4273	Akaike info criterion		14.08254
Sum squared resid	913403.7	Schwarz criterion		14.22958
Log likelihood	-116.7016	F-statistic		27.34006
Durbin-Watson stat	2.163179	Prob(F-statistic)		0.000015

Building the regression equation from the third group from 34 to 50

Table 4.10 : Goldfeld – Quandt test result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2192.701	1068.546	-2.052041	0.0594
BORROW	2.100109	0.846420	2.481167	0.0264
INCOME	1.050856	0.150227	6.995108	<0.0001
R-squared	0.777559	Mean dependent var		4967.647
Adjusted R-squared	0.745781	S.D. dependent var		2124.187
S.E. of regression	1071.017	Akaike info criterion		16.94939
Sum squared resid	16059094	Schwarz criterion		17.09643
Log likelihood	-141.0698	F-statistic		24.46896
Durbin-Watson stat	1.807393	Prob(F-statistic)		0.000027

Table 4.9 and 4.10 show the regression of the two models after arrangement the interest variable (income) that causes the non- constant variance and remove (n/3) from the middle data in variable income, after that we get SSE for each model which are the SSE(1) equal 16059094 and SEE(2) equal 913403.7 , we get $F_{stat} = \frac{SSE_2}{SSE_1} = \frac{16059094}{913403.7} = 17.58$ comparing with the critical value of

$F_{(0.05,15,15)} = 2.009$ So, F(stat) is greater than F(Critical) which means there is a violation in the constant variance.

4.4.4 Bresuch pagan test

Table 4.11 shows the results of Bresuch pagan test that $Q_B = \frac{175.3368}{2} = 87.6684$ comparing with the critical value of chi-square test with one degree of freedom equal 3.841459, so there is a violation in a constant variance

Table 4.11 : Bresuch pagan test result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.189974	0.657782	-1.809071	0.0768
BORROW	-0.000623	0.000829	-0.751930	0.4558
INCOME	0.000572	0.000131	4.378007	0.0001
R-squared	0.314270	Mean dependent var		1.000000
Adjusted R-squared	0.285090	S.D. dependent var		2.284347
S.E. of regression	1.931469	Akaike info criterion		4.212563
Sum squared resid	175.3368	Schwarz criterion		4.327284
Log likelihood	-102.3141	F-statistic		10.77006
Durbin-Watson stat	2.334615	Prob(F-statistic)		0.000141

4.4.5 Glejser test

Table 4.12 shows the result of Glejser test that we process the equation between (abs(residuals)) and log(Income), so we have the model is significant because p-value is less than 0.0001, which means the constant variance is violated

Table 4.12 : Glejser test result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.555191	1.307545	-5.013359	<0.0001
BORROW	-0.000232	0.000240	-0.966964	0.3385
LOG(INCOME)	0.886786	0.158417	5.597782	<0.0001
R-squared	0.420151	Mean dependent var		0.652291
Adjusted R-squared	0.395476	S.D. dependent var		0.724581
S.E. of regression	0.563370	Akaike info criterion		1.748363
Sum squared resid	14.91711	Schwarz criterion		1.863084
Log likelihood	-40.70907	F-statistic		17.02777
Durbin-Watson stat	2.304559	Prob(F-statistic)		0.000003

4.4.6 Levene's tes

Table 4.13 shows the result of levene's test that depends on the difference between the independent variable and the residual such that splitting the independent variable in two groups that the first is less than the mean of its variable and the second is greater its mean, so we have p-value is less than 0.0001, that means the constant variance is violated.

Table 4.13 : Leven's test result

Levene's Test for Equality of Variances		
Residual	F	Sig.
Equal variances assumed	15.127	<0.0001
Equal variances not assumed		

4.5 Remedies Methods

In this section , we introduce the remedies methods to solve the problem of heteroskedasticity

4.5.1 Weighted Least Square method

Table 4.14 shows the result of WLS method to correct the violation of Heteroskedasticity by dividing the proportionality factor that is income of each component in the regression model. Such that, p-value is less 0.0001 of each variable which means the variables is significant.

Table 4.14 : WLS result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	518.0095	126.4378	4.096951	0.0002
BORROW	1.290066	0.205381	6.281330	<0.0001
INCOME	0.691116	0.041757	16.55082	<0.0001
Weighted Statistics				
R-squared	0.670189	Mean dependent var		3015.748
Adjusted R-squared	0.656154	S.D. dependent var		788.2378
S.E. of regression	462.2097	Akaike info criterion		15.16804
Sum squared resid	10040978	Schwarz criterion		15.28276
Log likelihood	-376.2010	F-statistic		154.0471
Durbin-Watson stat	2.207132	Prob(F-statistic)		<0.000001

We use White test to diagnose the problem of heteroskedasticity. Table 4.15 shows p-value equals 0.666675 which is greater of than 0.05. So, the WLS method is successfully solved the violation in constant variance in the original data.

Table 4.15 : White test result

White Heteroskedasticity Test:

F-statistic	0.604989	Probability	0.696378
Obs*R-squared	3.216320	Probability	0.666675

4.5.2 Heteroskedasticity-Consistent Standard Errors & Covariance

Table 4.16 shows HCSE result by correcting the standard error in the regression model. So, the

p-value in each variable is less than 0.05 that means the variables is significant

Table 4.16 : HCSE result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	374.7572	262.7500	1.426288	0.1604
BORROW	1.383342	0.331117	4.177807	<0.0001
INCOME	0.717798	0.052189	13.75392	<0.0001
R-squared	0.803692	Mean dependent var		3628.400
Adjusted R-squared	0.795338	S.D. dependent var		1705.416
S.E. of regression	771.5223	Akaike info criterion		16.19273
Sum squared resid	27976596	Schwarz criterion		16.30745
Log likelihood	-401.8183	F-statistic		96.20975
Durbin-Watson stat	2.344052	Prob(F-statistic)		<0.000001

Table 4.17 shows White test result using White Heteroskedasticity-Consistent Standard Errors & Covariance by correcting the standard error in the regression model, which p-value equals 0.005234 which is less than 0.05 so, the problem still be violated and the HCSE is not efficient to remedy the violation**.**

Table 4.17 : White test result

White Heteroskedasticity Test:

F-statistic	4.389704	Probability	0.002501
Obs*R-squared	16.64065	Probability	0.005234

4.5.3 Redefining The Variables

Table 4.18 shows RF result by redefining the variables Income and Borrow which make the violation of homoskedasticity such that we transform income and borrow to log(equation). So, after transforming this variable we have the best transforming for removing the heteroskedasticity in original data. So, the p-value in each variable is less than 0.0001 that means the variables are significant.

Table 4.18 : RF result

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.323029	0.064586	5.001500	<0.0001
LOG(BORROW)	0.122039	0.021870	5.580099	<0.0001
LOG(INCOME)	0.656657	0.049618	13.23416	<0.0001
R-squared	0.886389	Mean dependent var		1.253226
Adjusted R-squared	0.877974	S.D. dependent var		0.043650
S.E. of regression	0.015248	Akaike info criterion		-5.434084
Sum squared resid	0.006278	Schwarz criterion		-5.293964
Log likelihood	84.51126	F-statistic		105.3268
Durbin-Watson stat	1.833085	Prob(F-statistic)		<0.000001

.

We use White test to diagnose the problem of heteroskedasticity Table 4.19 shows p-value equals 0.312428 which is greater of than 0.05. So, the RF method is successfully solved the violation in constant variance in the original data. So, the best transforming formula of the original data is given by the equation of log of each two side.

Final formula of a regression equation:

$$\log(\hat{C}onsumption) = 0.323029 + 0.122039\log(Borrow) + 0.656627\log(Income) \quad (4.5.1)$$

Table 4.19: White test result

F-statistic	1.184215	Probability	0.346105
Obs*R-squared	5.936692	Probability	0.312428

By comparing WLS and RF variables methods, we choose the RF approach technique because in addition of solving the problem of heteroskedasticity, also it has best R^2 ,Which equals to 87.9%.

4.6 Check The Assumptions of The Final Model in (4.5.1)

4.6.1 Normality Assumption:

Figure 4.5 represents the histogram of the residuals. It shows the residuals has randomly pattern (no pattern) over the residuals values. So, there is no violation on redefining variables

Figure 4.5

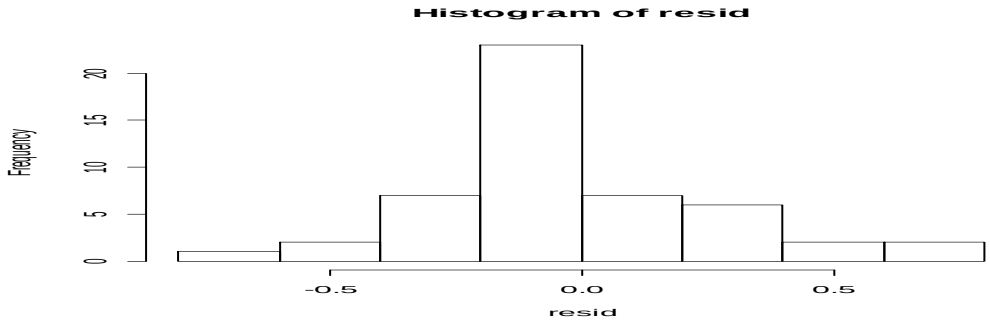


Table 4.20 Shows the normality test of the residuals is achieved because p-value is greater than 0.05 So, the residuals is normally distributed.

Table 4.20 : Normality test

Tests of Normality

	Kolmogorov -Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.169	30	.228	.946	30	.129

a. Lilliefors Significance Correction

4.6.2 Independence of Explanatory Variables Assumption:

Table 4.21 Shows that multicollinearity does not appear because VIF < 5

Table 4.21 : VIF Result

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Tolerance	VIF
C	0.323029	0.064586	5.001500	<0.0001		
LOG(BORROW)	0.122039	0.021870	5.580099	<0.0001	0.988	1.002
LOG(INCOME)	0.656657	0.049618	13.23416	<0.0001	0.988	1.002

4.6.3 Independence of disturbances Assumption:

The DW statistic has a range from 0 to 4 with a midpoint of 2. Table (4.22) shows that the observed value of DW=1.83 which is less than 2 this result is not surprising since the data is not truly time-series

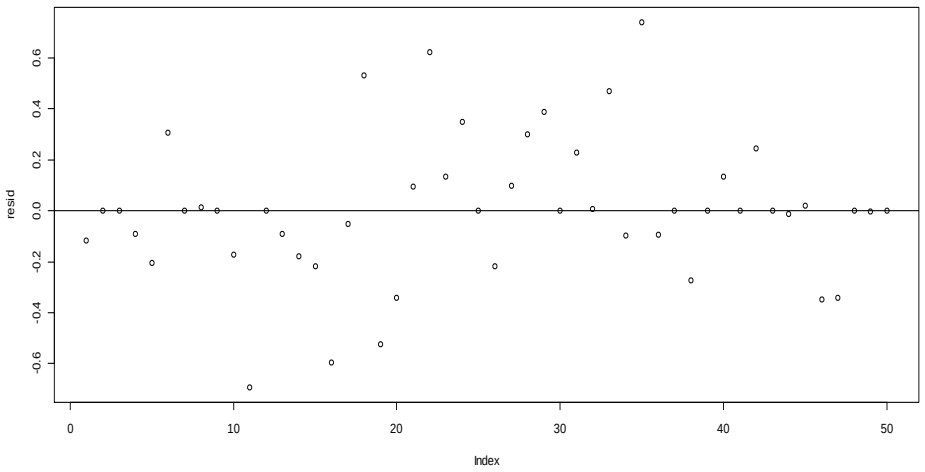
Table 4.22 : Durbin – Watson Result

R-squared	0.886389	Mean dependent var	1.253226
Adjusted R-squared	0.877974	S.D. dependent var	0.043650
S.E. of regression	0.015248	Akaike info criterion	-5.434084
Sum squared resid	0.006278	Schwarz criterion	-5.293964
Log likelihood	84.51126	F-statistic	105.3268
Durbin-Watson stat	**1.833085**	Prob(F-statistic)	<0.000001

4.6.4 Homoskedicity Assumption:

Figures 4.6 shows the residuals of regression equation scattering randomly around about zero with same pattern so, that means the residuals are constant variance (homoskedasticity), which means the assumption of constant variance is not violated.

Figure 4.6



4.6.5 Final regression model is :

$$\text{family consumption} = 0.323029 + 0.122039 * \log(\text{Borrow}) + 0.656657 * \log(\text{Income})$$

4.7 Summary:

We used different methods of detection the heteroskedasticity assumption and we found the best method of detect the heteroskedasticity assumption are Golfeld - Quandt test and Park test in this data. Also, we used different methods to remedy the volation of constant variance which the best method of remedy is redefining the variables.

Chapter 5

5.1 Conclusion

The researcher came to know the theoretical foundations for the regression model plan and realized the main problem that it proposes in which it shows his predictable ability.

The research problem is to see the problem constant variance and ways of finding and means of tacking them remedy methods .

This is through many different procedures, so the data regression model was done. The researcher detection that the model suffers from this problem and that he followed remedy methods and the results are the following:

- Goldfeld – Quandt test is considered the best way at the detection tests that problem and this is due to Goldfeld – Quandt test which keep to OLS conditions .
- The White test does not make any assumptions about the particular form of heteroskedasticity, and so it is quite general in application.
- After that park test was used because it depends on remedy methods for the model two sides which shows the relation between the residual and the problem independent variable
- The White test does not make any assumptions about the particular form of heteroskedasticity, and so it is quite general in application
- However, if you use White's standard errors, eradicating the heteroskedasticity is less important.
- The remedy methods, we found the redefining variable method is the best one for our data
- The result conditions were confirmed according to OLS showing that the chosen model given the best predictable one for our data

5.2 Recommendation

For further researchers we recommend to do the following :

- Testing the Heteroskedasticity In Nonlinear Regression models
- Extend Linear Regression using Local Polynomial Regression
- Testing Heteroskedasticity in cross- section data and time series data
- Testing Autocorrelation in cross- section data and time series data
- Testing Generalized Autoregressive conditional heteroskedasticity in the time series.

REFERENCE

1- A.H. Studenmud(2010), Using Econometrics : Apractical Guide, 6th Edition. Addison Wesley Longman.

2- Alfred DeMaris (2004), Regression With Social Data : Modeling Continuous and Limited Response Variables John Wiley & Sons, Mathematics - 534 pages

3- Breusch, T. S., and A. R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," Econometrica 50, 987–1007.

4- Goldfeld, Stephen M.; Quandt, R. E. (1965), "Some Tests for Homoscedasticity". Journal of the American Statistical Association.

5- Graybill, F. A. (2002), Matrices with Applications in statistics .2nd ed.Belmont, calif :Wadsworth .

6- Greene, W.H. (1993), Econometric Analysis, Second Edition, New York: Macmillan Publishing Company.

7- Harvey, A. (1990), The Econometric Analysis of Economic Time Series. 2d ed. Cambridge.

8- Harvey, A. (1979), "Estimating Regression Models with Multiplicative Heteroskedasticity." Econometrica, 44, pp. 461–465.

9- Harvey, A., and G. Phillips.(1974), "A Comparisonof the Power of Some Tests for Heteroskedasticity in the General Linear" University of California, Berkeley, Department of Economics

10- H. White. (1980), "A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity." Econometrica, 48, 817-838.

11- Judge, G., W. Griffiths, C. Hill, and T. Lee. (1985) The Theory and Practice of Econometrics. New York: JohnWiley and Sons

12- Koenker, R., and G. Bassett.(1982), "Robust Tests for Heteroskedasticity Based on Regression Quantiles." Econometrica, 50,pp43-61

13- Koenker,R. (1981), "ANote on Studentizing aTest for Heteroskedasticity." Journal of Econometrics, 17, pp. 107–112

14- J. Huston McCulloch,(1985), "Miscellanea On Heteroskedasticity", Econometrica, Vol. 53.

15- Jerry l. Hintze and Kaysville, Utah (2007), " Gene Expression Statistical System" GESS- USA.

16- MacKinnon, J.G. and H. White. (1985), ‘Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties’. Journal of Econometrics, 29, 53-57

17- M. Kuter, C.Nachtsheim and J.Neter (2004), " Applied Linear Regression Models " 4th

18- Ohtani, Kazuhiro & Toyoda, Toshihisa, (1980), "Estimation of regression coefficients after a preliminary test for homoscedasticity," Journal of Econometrics, Elsevier, vol. 12(2), pages 151-159, February

S.M. Goldfeld and R.E. Quandt, (1965), "Some Tests for Homoscedasticity," Journal of the American Statistical Society, Vol.60.

19- T.S. Breusch and A.R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," Econometrica, Vol. 47.

20- Talwar ,(1983), "detecting a shift in location " . Journal of Econometrics23,53-367

21- Waldman,D.(1983), “ANote on the Algebraic Equivalence of White’s Test and a Variant of the Godfrey/Breusch-Pagan Test for White, H., ed. “Non-Nested Models.” Journal of Econometrics, 21, 1, pp. 1–160

22- Woolridge, J.(2003), " Introductory Econometrics: A Modern Approach". New York: Southwestern Publishers.

Research Reference

1- Andrew F. Hayes (2009), " Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software "

2- Mario Francisco · Juan M. Vilar (2007), " Two tests for heteroscedasticity in nonparametric regression

3- Xu Zheng (2009), " Testing Heteroscedasticity in Nonlinear and Nonparametric Regression"

4- Muhammad Aslam1 and Gulam Rasool Pasha2 (2000), "Adaptive Estimation of Heteroscedastic Linear Regression Models Using Heteroscedasticity Consistent Covariance Matrix

5- Leonor Ayyangar, (2007), “ Skewness, Multicollinearity, Heteroskedasticity - You Name It, Cost Data Have It! Solutions to Violations of Assumptions of Ordinary Least Squares Regression Models”

6- James H. Stock and Mark W. Watson,(2006), “Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression”

7- O. Baser,(2007), “Modeling Transformed Health Care Cost with Unknown Heteroskedasticity”

8- Donald W. K. Andrews and Patrik Guggenberger,(2011), ”ConditionalHeteroskedasticity-Robust Confidence Interval for the Autoregressive Parameter”

9- Joris Pinkse,(2006) “Heteroskedasticity Correction and Dimension Reduction”

10- Hausman, Newey, Woutersen, Chao, and Swanson,(2009), “Instrumental Variable Estimation with Heteroskedasticity and Many Instruments

11- Andreea Halunga,Chris D. Orme andTakashi Yamagata, (2011) ”A Heteroskedasticit Robust Breusch-Pagan Test for Contemporaneous Correlation in Dynamic Panel Data Models

12- Timo Teräsvirta,(2011) "Nonlinear models for autoregressive conditional heteroskedasticity"