YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

THE VIOLATION FOR ASSUMPTIONS OF MULTIPLE REGRESSION MODEL

BY

MAY THU

M.Econ (Statistics)

Roll No.4

NOVEMBER, 2019

YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

THE VIOLATION FOR ASSUMPTIONS OF MULTIPLE REGRESSION MODEL

Thesis submitted as a partial fulfillment towards

the Degree of Master of Economics

BY

MAY THU

M.Econ (Statistics)

Roll No.4

NOVEMBER, 2019

YANGON UNIVERSITY OF ECONOMICS
DEPARTMENT OF STATISTICS

THE VIOLATION FOR ASSUMPTIONS OF MULTIPLE REGRESSION MODEL

This thesis is submitted to board of Examination as partial fulfillment of the requirement for degree of M.Econ (Statistics).

Approved by the Board of Examiners

Supervised by:

Daw Saw Nan Swe
Lecturer
Department of Statistics
Yangon University of Economic

Submitted by:

Ma May Thu
Roll No.4
M.Econ (Statistics)

NOVEMBER, 2019

YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

This is certify that the thesis entitled "**THE VIOLATION FOR ASSUMPTIONS OF MULTIPLE REGRESSION MODEL"** submitted as a partial fulfillment towards the requirements of Master of Economics (Statistics) has been accepted by the Board of Examiners.

BOARD OF EXAMINERS

(Chairman)

Professor Dr. Tin Win

Rector

Yangon University of Economics

Professor Dr. Ni Lar Myint Htoo

Pro-Rector

Yangon University of Economics

(Chief Examiner)

Prof. Dr. Maw Maw Khin

Professor and Head

Department of Statistics

Yangon University of Economics

(External Examiner)

Dr. Swe Swe Zin

Associate Professor

Department of Shipping Management

Myanmar Maritime University

(Examiner)

Daw Khin Nu Win

Associate Professor

Department of Statistics

Yangon University of Economics

(Examiner)

Daw Aye Aye Maw

Associate Professor

Department of Statistics

Yangon University of Economics

NOVEMBER, 2019

ABSTRACT

The study intends to apply some of the most common and appropriate detections and remedies methods to meet the assumptions of a multiple linear regression model. When the assumptions are violated, then the inferences about the parameter estimate will be incorrect. The secondary data for maize (1998-2018), wheat (1998-2018), rice (1966-2018) and sesame (1989-2018) of Myanmar. Maize data for linearity assumption is used to detect and remedy. Wheat data for normality assumption is used to apply in the detection and remedial ways. Rice data for homoscedasticity assumption is used and sesame data for micronumerosity assumption, multicollinearity assumption, and the nature of independent variables assumption, autocorrelation assumption are used to diagnosis and resolving ways.

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest gratitude to Professor Dr. Tin Win, Rector of Yangon University of Economics, for allowing me to develop this thesis.

I am also thanks to Dr. Ni Lar Myint Htoo, Pro-Rector of Yangon University of Economics, for supporting to carry out this thesis.

I am greatly indebted to Professor Dr. Maw Maw Khin, Head of the Department of Statistics, Yangon University of Economics, for her permission, valuable suggestions and recommendations to prepare this thesis study.

I would like to express my indebtedness to Professor Dr. Mya Thandar, Department of Statistics, Yangon University of Economics, for her valuable suggestions and recommendations to improve my thesis.

I would like to express my indebtedness to Professor Dr. Ching Do Nem, Department of Statistics, Yangon University of Economics, for her valuable suggestions and recommendations to improve my thesis.

Special thanks go to my external examiner, Dr. Swe Swe Zin, Associated Professor, Department of Shipping Management, Myanmar Maritime University, for her valuable guidance and helpful advice.

Special thanks go to my co-supervisor, Dr. Aye Thida, Associated Professor, Department of Statistics, Yangon University of Economics, for her valuable guidance, helpful advice and supervision.

Special thanks go to my supervisor, Daw Saw Nan Swe, Lecturer, Department of Statistics, Yangon University of Economics, for her valuable guidance, helpful advice and supervision.

Finally, I would like to special thank my parents for supporting and encouraging me to study in the cause of Master of Economics (Statistics) without worries.

CONTENTS

Pages

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BLUE	=	Best Linear Unbiased Estimator
OLS	=	Ordinary Least Square
MR	=	Multiple Regression
TOL	=	Tolerance
VIF	=	Variance Inflation Factor
PROD	=	Production
WR	=	Weir
SN	=	Sown
QUALI	=	Quality Seeds
HAR	=	Harvested
IRRI	=	Irrigation
PESTI	=	Pesticides
SUBSOWN	=	Substitution Sown
YIE	=	Yield
LA	=	Loan

CHAPTER I

INTRODUCTION

1.1 Rationale of the Study

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. Although the regression problem may be solved by the number of techniques, the most-used method is Ordinary Least Squares (OLS). The regression model has a good predictive ability by estimating the coefficient using OLS method. According to the Gauss-Markov Theorem, the OLS estimator provides the Best Linear Unbiased Estimator (BLUE).

All parametric tests in a statistical analysis assume some certain assumptions about the data. The classical multiple regression model has several assumptions. When one of these assumptions is violated, the classical tests such as t and F are no longer appropriate. A violation of any of these assumptions changes the conclusion of the research and interpretation of the results. Therefore all research must check and adhere to these assumptions for accurate interpretation and model integrity. The classical multiple regression model is based on several simplifying assumptions, which are as follows.

The first one of the assumption of linear regression analysis is the relationship between dependent variable and independent variables to be linear in the parameters. If linearity is violated, the results of the regression analysis will under- or over-estimate the true relationship and increase the risk of Type I and Type II errors (Osborne & Waters, 2002). Violation of this assumption threatens the meaning of the parameters estimated in the analysis (Keith, 2006).

In the second assumption, the values of the independent variables, the X's are fixed, or X values are independent of the error term, that is, the covariance between the disturbance terms and each independent variable must be zero. Violation this assumption biases the coefficient estimate. When an independent variable correlates with the error term, OLS incorrectly attributes some of the variance that the error term actually explains to the independent variable instead.

The third assumption is that for given independent variables (X's), the mean value of the disturbance terms must be zero. If this assumption is not fulfilled, the original intercept cannot be estimated.

The forth assumption is that any disturbances have the same variance. That is, there is homoscedasticity or no heteroscedasticity. If the error term of an equation is known to be heteroscedastic, there are three major consequences: Even if the error term of an equation is known to be purely heteroskedastic, that heteroscedasticity will not cause bias in the OLS estimates of the coefficient, heteroscedasticity increases the variances of the $\hat{\beta}$ coefficient and heteroskedasticity causes OLS to tend to underestimate the variance and standard error of the coefficients.

The fifth assumption is that any two errors are independent of each other. This assumption can be broken when data are collected on the same variables over time. This is known as autocorrelation. If the disturbances are auto-correlated, the regression coefficients remain unbiased and consistent, but are not efficient and regression model is less reliable.

The sixth assumption is that the number of observations n must be greater than the number of parameters to be estimated. Although the best linear unbiased estimator (BLUE), the OLS estimators have large variances and covariances, making precise estimation difficult. The t-ratio of one or more coefficients tends to be statistically insignificant. Although the t ratio of one or more coefficients is statistically insignificant, R^2, the overall measure of goodness of fit can be very high. The OLS estimators and their standard errors can be sensitive to small changes in the data.

In the seventh assumption, the independent variables X values in a given sample must not all be the same, that is, the variables must vary. If all the X values are identical, then $X_i = \bar{X}$ and the denominator of that equation will be zero, making it impossible to estimate β_2 and β_1. The variance of independent variable must be a positive number. If there is very little variation in the independent variables, much of the variation in dependent variable can be explain. Furthermore, there can be no outliers in X values. When there is outliers in the values of X variables, this observation influence its own prediction and its means and inflate the standard error with which it is being standardized.

The eighth assumption is that any independent variables are not correlated with each other. Multicollinearity occurs when the independent variables are too

highly correlated with each other. Multicollinearity is a phenomenon in with one independent variable in a multiple regression model can be linearly predicted from the other. In this situation, the estimates of linear regression coefficient may change in response to small changes in the model or the data. Multicollinearlity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set.

The ninth assumption is that the model is correctly specified, so there is no specification bias. The usual confidence interval and hypothesis testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters. The estimated parameters in overfitting model will be generally inefficient.

The final assumption is that the stochastic disturbance terms is normally distributed. When scores on variables are skewed, correlations with other measures will be attenuated, and when the range of scores in the sample is restricted relative to the population correlations with scores on other variables will be attenuated (Hoyt et al., 2006). Non-normally distributed variables can distort relationships and significance tests (Osborne & Waters, 2002). Outliers can influence both Type I and Type II errors and the overall accuracy of results (Osborne & Waters, 2002).

In this thesis, the above assumptions in the classical regression model are studied and some of the most common and appropriate detections and remedies methods are discussed in order to diagnose and solve these problems.

1.2 Objectives of the Study

The objectives of the study are

(i) To develop the classical multiple linear regression model.

(ii) To check the violation of the classical multiple linear regression model.

(iii) To modify the model by using the requiring remedies.

1.3 Method of study

In this study, the descriptive methods and multiple regression analysis are used to detect whether the assumptions are violated and to resolve. Linearity is examined through scatter plots, residual plot and remedy way is the transforming variables. Normality is checked with box plot and Kolmogorov-Smirnov test and transforming variable is used to remedy the normality. In detection of heteroscedasticity, White's

general heteroscedasticity test is used and the remedy methods of heteroscedasticity are redefining the variables. Micronumersity problem are remedied by adding the observations. Examination of tolerance (TOL) and Variance Influence Factor (VIF) are used in the detection of multicollinearity and then the variables assume that had the multicollinearity is removed. The outliers of the X variables are exposed with the box-plots and the influence observations are observed with Cook's distance. The Durbin-Watson test is used to detect the autocorrelation between the disturbances and remedy way is generalized difference equation method.

1.4 Scope and Limitations of the Study

In the scope of this study, yearly secondary data of sesame production and rice production are used. The study periods of maize production and wheat production are from 1998 to 2018 in Myanmar, rice production is from 1966 to 2018 in Myanmar and sesame production is from 1989 to 2018 in Myanmar. The main sources of these data are various statistical yearbooks published from Central Statistical Organization (CSO) and Agricultural Statistics.

1.5 Organization of the Study

This study consists of five chapters. Chapter I is introductory chapter including of rationale, objectives, method, scope and limitation and organization of the study. Chapter II is literature review and theoretical background is discussed in Chapter III. Chapter IV presents detection and remedy of the assumptions of multiple linear regression model. Finally, conclusion is presented in chapter V.

CHAPTER-II

LITERATURE REVIEW

The study by Andrew F. Hayes (2009) focuses on investigating the Homoscedasticity as an important assumption in OLS regression. Although the estimator of the regression parameters in OLS regression is unbiased when the homoscedasticity assumption is violated, the estimator of the covariance matrix of the parameter estimates can be biased and inconsistent under heteroscedasticity, which can produce significance tests and confidence intervals that can be liberal or conservative. After a brief description of heteroscedasticity and its effects on inference in OLS regression, this study discuss a family of heteroscedasticity-consistent standard error estimators for OLS regression.

Mario Francisco Juan M. Vilar (2007), the study focuses on two new tests for heteroscedasticity 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 constant. The second test is based on using an overall distance between a nonparametric estimators 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 content 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 heteroscedasticity 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 heteroscedasticity. Amonte Carlo simulation with critical values obtained from the wild bootstrap procedure is provided to assess 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 (2000), this study focuses on the estimation of linear regression models in the presence of heteroscedasticity of unknown form, method of OLS does not provide the estimates with the smallest variances. In this situation, adaptive estimators are used, namely, nonparametric kernel estimator and nearest neighbor regression estimator. But these estimators rely on substantially restrictive conditions. In order to have accurate inferences in the presence of heteroscedasticity of unknown form, it is a usual practice to use heteroscedasticity consistent covariance matrix (HCCME). Following the idea behind the construction of HCCME, they 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.

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 heteroscedasticity complicates the process. Smearing estimation suggested in the literature only works for homoscedastic errors or heteroscedastic errors due to categorical variables. Generalized linear models have been proposed as an alternative approach for log models when there exists unknown forms of heteroscedasticity. Recent literature shows that log models are superior to generalized linear models under certain conditions. They present a method for applying transformation that accounts for any form of heteroscedasticity. The proposed model assumes that errors achieve normality. Heteroscedasticity is modeled separately. Simulation studies are conducted. The Medstat Market Scan Database are used to estimate healthcare costs for asthma patients. Finally, a comparison of the method with smearing estimators is estimated. Log-transformed health care costs of asthma patients were normal. There was an evidence of heteroscedasticity. The simulation study, heteroscedasticity adjusted retransformed costs had the lowest mean squared error relative to estimators from smearing transformation or generalized linear model. This study shows that if log-transformed costs are normally distributed, heteroscedasticity 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 heteroscedasticity of general form and AR

parameters that are less than or equal to unity. The CI is a modification of Mikusheva's (2007) modification of Stock's (1991) CI that employs the least squares estimator and a heteroscedasticity-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 heteroscedasticity.

Joris Pinkse (2006), this paper provides a nonparametric method of correcting for heteroscedasticity 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 present of heteroscedasticity. The new estimator is the most efficient estimator in a wide class of estimators for which the heteroscedasticity 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 heteroscedastic data. Such settings are common in microeconometric applications where many instruments are used to improve efficiency and allowance for heteroscedasticity is generally important. The solution is a Fuller (1977) like estimator and standard errors that are robust to heteroscedasticity and many instruments. The estimator has finite moments and high asymptotic efficiency in a range of cases. The standard errors are easy to compute, being like White's (1982), with additional terms that account for many instruments. They are consistent under standard, many instrument, and many weak instrument asymptotic. Based on a series of Monte Carlo experiments, this study find that the estimators perform as well as LIML or Fuller (1977) under homoscedasticity, and have much lower bias and dispersion under heteroscedasticity, in nearly all cases considered.

Andreea Halunga, Chris D. Orme and Takashi Yamagata (2011) this paper proposes a heteroscedasticity-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 strictly exogenous and/ or lagged dependent regressor variables, as well as quite general forms of both non-normality and heteroscedasticity

in the error distribution. While the asymptotic validity of the test procedure, under the null, is predicated on the number of time series observation, T, being large relative to the number of cross-section units, N, independence of 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 theory 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 consider 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 Breush-Pagan test (Godfrey and Yamagata, 2011) the corresponding version of the heteroscedasticity-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 Terasvirta (2011), this paper contains a brief survey of nonlinear models of autoregressive conditional heteroscedasticity. 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 based on multiplicative decomposition of the variance receive attention.

P. Marshall, T. Szikszai, V. LeMay and A. Kozak (1995), this paper contains testing the distributional assumptions of least squares linear regression. The error terms in least squares linear regression are assumed to be normally distributed with equal variance (homoscedastic), and independent of one another. If any of these distributional assumptions are violated, several of the desirable properties of a least squares fit may not hold. A variety of statistical tests of the assumptions is available. This paper are recommended for reasons of ease of use and discriminating power: the K^2 test for testing for non-normality, either the Durbin-Watson test or the Q-test for testing autocorrelation and either Szroeter's or White's test for testing for heteroscedasticity. Violating any of the distributional assumptions of least squares linear regression can impact on the properties of the regression equation, most notably the efficiency and unbiasedness of the estimates of variance. A variety of tests exist that allow these assumptions to be tested. The order in which the tests are applied is important since violation of one of the assumptions may invalidate the results of subsequent tests.

Germa Coenders and Marc Saez (2000), this paper review some classic collinearity, heteroscedasticity and outlier diagnostics in multiple regression models. Some major problems are described in the Breusch-Pagan test, the condition number and the critical values for the studentized deleted residual and cook's distance. Alternatives are suggested which consist of computing the conditional number of the correlation matrix instead of the rescaled moment matrix, using the NR^2 statistic for the Breusch Pagan test, setting global-risk- based critical values for the studentized deleted residual, and drawing graphical displays for Cook's distance. Very large differences between the original and alternative diagnostics emerge both on simulated data and on real data from a work absenteeism study. In this paper, major weaknesses of some commonly used collinearity heteroscedasticity and outlier diagnostics. The classic diagnostics have been compared to the alternatives on an empirical data set. The differences were large enough to lead to completely different conclusions depending on which diagnostics were employed. Further robustness problems of these diagnostics are not solved by the suggested alternatives. Critical values for studentized deleted residuals are very sensitive to the normality assumption. Heteroscedasticity tests are very sensitive to the presence of outliers because they involve squaring the residuals, which makes outliers to have a more serious effect in the auxiliary than in the main regression.

Deanna Schreiber-Gregory (2018), this paper contains violation recognition and control of logistic and linear regression assumptions. Regression analyses are one of the first steps in any analytic plan, regardless of plan complexity. Therefore, it is worth acknowledging that the choice and implementation of the wrong type of regression model, or the violation of its assumptions can have detrimental effects to the results and future directions of any analysis. Considering this, it is important to understand the assumptions of these models and be aware of the processes that can be utilized to test whether these assumptions are being violated. Some logistic regression assumptions that will reviewed include: dependent variable structure, observation independence, absence of multicollinearity, linearity of independent variables and log odds, and large sample size. For linear regression, the assumptions that will be reviewed include: linearity, multivariate normality, absence of multicollinearity and autocorrelation, homoscedasticity and measurement level. In order to ensure that the model is appropriately interpreted, it is important to make sure that all assumptions have been tested and any violations have been corrected.

Oyeyemi, G. M., Bolakale, A., Folorunsho, A.I. and Garba, M.K., this paper looks at the problem of micronumerosity in classical linear regression (CLR) models in other to prescribe appropriate remedy to the problem if encountered at any CLR problem. This study is aimed at determining an optimum sample size such that when the number of observations of variables in CLR is greater than then micronumerosity is not a problem and to suggest means of correcting micronumerosity in CLR. The optimum sample size for a given number of independent variables and level of correlation between the dependent and independent variables was determined. If there is presence of Micronumerosity in a data set, then additional data should be obtained. If it is not possible to increase the sample size, then the best method of remedying micronumerosity is to use factor analysis regression or principal component regression.

CHAPTER III

THEORETICAL BACKGROUND

Multiple regression examines the relationship between a single outcome measure and several predictor or independent variables (Jaccard et al., 2006). Statistical tests rely upon certain assumptions about the variables used in an analysis (Osborne & Waters, 2002). The correct use of the multiple regression model requires that several critical assumptions be satisfied in order to apply the model and establish validity (Poole & O'Farrell, 1971). Inferences and generalizations about the theory are only valid if the assumptions in an analysis have been tested and fulfilled.

Multiple regression is widely used to estimate the size and significance of the effects of a number of independent variables on a dependent variable (Neale, Eaves, Kendler, Heath, & Kessler, 1994). Before a complete regression analysis can be performed, the assumptions concerning the original data must be made (Sevier, 1957). Ignoring the regression assumptions contribute to wrong validity estimates (Antonakis, & Deitz, 2011). When the assumptions are not met, the results may result in Type I or Type II errors, or over- or under-estimation of significance of effect size (Osborne & Waters, 2002).

Meaningful data analysis relies on the researcher's understanding and testing of the assumptions and the consequences of violations. The extant research suggests that few articles are reporting having tested the assumptions of the statistical tests they rely on for drawing their conclusions (Antonakis & Dietz, 2011; Osborne & Waters, 2002; Poole & O'Farrell, 1971). The validation and reliability of theory and future research relies on diligence in meeting assumptions of MR. This section specifically define each assumption, review consequences of assumption failure, and address how to test for each assumption, and the interpretation of results.

3.1 Linearity

This assumption is the most important, as it directly relates to the bias of the results of the whole analysis (Keith, 2006). Linearity defines the dependent variable as a linear function of the predictor (independent) variables (Darlington, 1968). Multiple regression can accurately estimate the relationship between dependent and independent variables when the relationship is linear in nature (Osborne & Waters, 2002).

Linearity can be interpreted in two different ways. The first interpretation is linearity in the variables. It is that the conditional expectation of Y is a linear function of X_i such as $(Y \mid X_i) = \beta_1 + \beta_2 X_i$. Geometrically, the regression curve in this case is a straight line. In this interpretation, a regression function such as;

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$$

is not a linear function because the variable X appears with a power or index of 2 (Gujarati).

The second interpretation of linearity is that the conditional expectation of Y, E (Y|X) is a linear function of the parameters, the β's; it may or may not be linear in the variable X. In this interpretation $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is a linear (in the parameter) regression model. The model $E(Y \mid X_i) = \beta_1 + \beta_2^2 X_i$ which is nonlinear in the parameter β_2 . This model is a nonlinear (in the parameter) regression model (Gujarati).

Linear in the parameters as well as the variables is a linear regression model and so is a model that is linear in the parameter but nonlinear in the variables. If a model is nonlinear in the parameters it is a nonlinear (in the parameter) regression model whether the variables of such a model are linear or not. For some models look nonlinear in the parameters but are inherently or intrinsically linear because with suitable transformation they can be made linear in the parameter regression models. But if such models cannot be linearized in the parameters, they are called intrinsically nonlinear regression model (Gujarati).

3.1.1 Consequences of Violation of Linearity

If linearity is violated, all the estimates of the regression including regression coefficients, standard errors, and tests of statistical significance may be biased (Keith, 2006). If the relationship between the dependent and independent variables is not linear, the results of the regression analysis will under- or over- estimate the true relationship and increase the risk of Type I and Type II errors (Osborne & Waters, 2002). When bias occurs it is likely that it does not reproduce the true population values (Keith, 2006). Violation of this assumption threatens the meaning of the parameters estimated in the analysis (Keith, 2006).

3.1.2 Detection of Linearity

Examination of the residual plots also indicate linear vs. curvilinear relationships (Keith, 2006; Osborne & Waters, 2002). Residual plots showing the standardized residuals vs. the predicted values are very useful in detecting violation in Linearity (Stevens, 2009). The residuals magnify the departures from linearity (Keith, 2006). If there is no departure from linearity, a random scatter about the horizontal line would be seen. Any systematic pattern or clustering of the residuals suggests violation (Stevens, 2009). If each parameter test is significant and R square is high, linearity assumption is satisfied. Another way for detection of linearity is scatter plot that must be a straight line. If not, linearity assumption is violated. In this paper, scatter plots and residual plots are used to diagnosis the linearity assumption.

3.1.3 Remedy of Violation of linearity

The relationship between the dependent and independent variables is linear. However, this is not always the case. The data without the linearity suggest a nonlinear (or curvilinear) as follows:

$$\hat{Y} = b_0 b_1 X \tag{3.1}$$

where b_0 and b_1 are constants.

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \tag{3.2}$$

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{3.3}$$

Without the linearity, the data must be transformed for one or both variables in order to display it as a linear model. There are many transformation ways such as double log transformation, semi log transformation, squared root transformation) to satisfy the linearity assumption. A common method of transformation is logarithmic transformation makes.

A linear relationship assumes that for every one-unit change on X, Y changes by a constant amount. A curvilinear model assumes that Y changes by a different amount each time. The rules of logarithms allow to express formula:

$$Ln\,(\hat{Y}) = lnb_0 + b_1 ln(X) \tag{3.4}$$

where,

$$b_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} \quad (3.5)$$

$$\ln b_0 = \frac{\sum \ln(Y)}{n} - b_1 \frac{\sum \ln(X)}{n} \quad (3.6)$$

3.2 Fixed X Values or X Values Independent of the Error Term

One of the assumptions was that the explanatory variables or regressors were either fixed or nonstochastic or if stochastic, they were independent of the error term. The X variable (s) can be assumed nonstochastic for the following reasons:

1. First, this is done initially to simplify the analysis and to introduce the reader to the complexities of regression analysis gradually.
2. Second, in experimental situations it may not be unrealistic to assume that the X values are fixed.
3. Third, even if the X variables are stochastic, the statistical results of linear regression based on the case of fixed regressors are valid when the X's are random, provided that some conditions are met. One condition is that regressor X and error u_i are independent.

Violation of this assumption can occur because there is simultaneity between the independent and dependent variables, omitted variable bias or measurement error in the dependent variables.

3.2.1 Consequences of Violation of Fixed X Values

Violation this assumption biases the coefficient estimates. To understand why this bias occurs, keep in mind that the error term always explains some of the variability in the dependent variable. However, when an independent variable correlates with the error term, OLS incorrectly attributes some of the variance that the error term actually explains to the independent variable instead.

3.2.2 Detection of Fixed X Values

This assumption is assumed that the X variables and the error term are independent. Therefore, the covariance of X_i and u_i must be zero. If $cov(X_i, u_i) \neq 0$, this assumption is violated.

3.2.3 Remedy of Violation of Fixed X Values

If an independent variable is correlated with the error term, the independent variable can be used to predict the error term which violates the notation that the error term represents unpredictable random error. The remedy ways of fixed X values or X values independent of the error terms are the omission of a variable and the imposition of any correct restriction.

3.3 Zero Mean Value of Disturbances u_i

This assumption states that the mean value of u_i conditional upon the given X_i is zero. This assumption should not be difficult to comprehend. The factors not explicitly included in the model and subsumed in u_i, do not systematically affect the mean value of Y. The positive u_i values cancel out the negative u_i values so that their average or mean effect on Y is zero. The assumption E ($u_i|X_i$) =0 implies that $E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$. Therefore, the two assumption are equivalent.

If the conditional mean of one random variable given another random variable is zero, the covariance between the two variables is zero and hence the two variables are uncorrelated. Therefore, this assumption implies that X_i and u_i are uncorrelated.

3.3.1 Consequences of Violation of Zero Mean Value of Disturbances u_i

If this assumption is not fulfilled or violated, the original intercept cannot be estimated because the original intercept added the error terms can be available.

3.3.2 Detection of Violation of Zero Mean Value of Disturbances u_i

Given the value of Xi, the mean or expected value of random disturbance term ui is zero. Symbolically, $E(u_i \mid X_i) = 0$ or if X is nonstochastic, $E(u_i) = 0$. If $E(u_i \mid X_i) \neq 0$ or $E(u_i) \neq 0$, this assumption is violated.

3.3.3 Remedy of Violation of Zero Mean Value of Disturbances u_i

When the population regression function (PRF) is expressed as

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (3.7)$$

X and u which represent the influence of all omitted variables have separate influences on Y. Thus, if X and u are correlated, it is not possible to assess their individual effect on Y. It is quite possible that the error term actually includes some

variables that should have been included as additional regressors in the model. This assumption is another way of stating that there is no specification error in the chosen regression model.

3.4 Homoscedasticity or Constant Variance of Disturbances u_i

One of the important assumptions of the classical linear regression model is that the variance of each disturbance term u_i conditional on the chosen value of the explanatory variables is some constant number equal to σ^2. This is the assumption of homoscedasticity or equal (homo) spread (scedsticity) that is equal variance.

Symbolically,

$$E(u_i^2) = \sigma^2 \qquad , i=1, 2, 3\ldots n \tag{3.8}$$

The conditional variance of Y_i increase as X increases. Here, the variances of Y_i are not the same. Hence, there is heteroscedasticity. Symbolically,

$$E(u_i^2) = \sigma_i^2 \qquad , i=1, 2, 3\ldots n \tag{3.9}$$

Heteroscedasticity can be divided into pure and impure version. Pure heteroscedasticity is caused by the error term of the correctly specified equation; impure heteroscedasticity is caused by a specification error such as an omitted variable.

3.4.1 Consequences of Heteroscedasticity

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

1. Pure heteroscedasticity does not cause bias in the coefficient estimates. Even if the error term of an equation is known to be purely heteroskedastic, that heteroscedasticity will not cause bias in the OLS estimates of the coefficient. This is true because with pure heteroscedasticity, none of the independent variable is correlated with the error term. As a result, pure heteroscedasticity still has property that:

$$E(\hat{\beta}) = \beta \text{ for all } \beta s \tag{3.10}$$

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

$$VAR\ (\varepsilon_i) = \epsilon_i^2 Z_i^2 \tag{3.11}$$

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 β'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.

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 uncorrelated heteroskedasticity. In practice, OLS usually ends up with higher t scores than would be obtained if the error term were homoscedastic, sometimes 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.

3.4.2 Detection of Heteroscedasticity

Heteroscedasticity can be detected by using the informal method such as graphical method and formal method such as Park test, Glejser test, Spearman's Rank Correlation test and Goldfeld-Quandt test, etc. In this study, White's General Heteroscedasticity test are used. Unlike the Goldfeld-Quandt test, which requires reordering the observations with respect to the X variable that supposedly caused heteroscedasticity, or the BPG test, which is sensitive to the normality assumption, the general test of heteroscedastcity proposed by White does not rely on the normality assumption and is easy to implement. Consider the regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \qquad (3.12)$$

The White test proceed as follows:

Step 1. Given the data, estimate Eq (3.12**)** and obtain the residuals,$\hat{u}_i$

Step 2. Run the following (auxiliary) regression:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \cdots + \alpha_k X_{ki} + \alpha_{k+1} X_{(k+1)i}^2 + \cdots + \alpha_{2k} X_{2i} X_{3i} + \cdots + v_i \quad (3.13)$$

Obtain the R^2 this (auxiliary) regression.

Step 3. Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R^2 obtained from the (auxiliary) regression asymptotically follows the chi-square distribution with degree of freedom equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is,

$$n.R^2 \sim x_{df}^2 \quad (3.14)$$

Step 4. If the chi-square value obtained in Eq (3.14) exceeds the critical chi-square value at the chosen level of significance, the conclusion that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression (3.13),

$\alpha_2 = \alpha_3 = \ldots = 0$.

3.4.3 Remedy of Heteroscedasticity

There are many remedial methods to remove the heteroscedasticity. There are weighted least squares method, White's heteroscedasticity consistent variances and standard errors and redefining method. A redefinition of the variables often is useful in allowing the estimated equation to focus more on the behavioral aspect of the relations. 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 often possible ways to avoid problem that had previously seemed insurmountable.

In some cases, the only redefinition that's needed to rid an equation of heteroscedasticity 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 it's less likely to encounter heteroscedasticity. In addition, there are many research topics for which the double-log is just as theatrically logical as the linear form. This is especially true if the linear form was chosen by default, as in often the case. In other situation, it might to necessary to completely rethink the research project in terms of underlying theory.

3.5 No Autocorrelation between the Disturbances or Independence of Errors

The term autocorrelation may be defined as correlation between members of series of the observations ordered in time or space. In the regression context, the classical linear regression model assumes that such autocorrelation does not exist in the disturbances u_i. Symbolically,

$$cov(u_i, u_j \mid x_i, x_j) = E(u_i u_j) = 0 \qquad i \neq j \qquad (3.15)$$

The classical model assumes that the disturbance term relating to any observation is not influenced by the disturbance term relating to any other observation. However, if there is such a dependence such autocorrelation does not exist in the disturbances u_i. Symbolically,

$$E(u_i u_j) \neq 0 \qquad i \neq j \qquad (3.16)$$

3.5.1 Consequences of Autocorrelation between the Disturbances

When independence of errors is violated standard scores and significance tests will not be accurate and there is increased risk of Type I error (Keith, 2006; Stevens, 2009). When data are not drawn independently from the population, the result is a risk of violating the assumption that errors are independent (Keith, 2002). This means that violations of this assumption can underestimate standard errors, and label variables as statistically significant when they are not (Keith, 2006). In the case of MR, effect sizes of other variables can be over-estimated if the covariate is not reliably measured (Osborne & Waters, 2002). Essentially what occurs is that the full effect of the covariate is not removed (Osborne & Waters, 2002). Violation of this assumption therefore threatens the interpretations of the analysis (Keith, 2006).

3.5.2 Detection of Autocorrelation between the Disturbances

The autocorrelation can be detected with graphical method, the Runs test, Durbin-Watson d test and the Breusch-Godfrey (BG) test. The most celebrated test for detecting autocorrelation is Durbin-Watson d statistic which is defined as

$$d = \frac{\sum_{t=2}^{t=n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n}\hat{u}_t{}^2} \qquad (3.17)$$

A great advantage of the d statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. It is important to note the assumptions underlying the d statistic.

1. The regression model includes the intercept term. If it is not present, as in the case of the regression through the origin, it is essential to rerun the regression including the intercept term to obtain the RSS.
2. The explanatory variables, the X's are nonstochastic or fixed in repeated sampling.
3. The disturbance u_t is generated by the first-order autoregressive scheme: $u_t = \rho u_{t-1} + \varepsilon_t$. Therefore, it cannot be used to detect higher order autoregressive schemes.
4. The error term u_t is assumed to be normally distributed.
5. The regression model does not include the lagged values of the dependent variable as one of the explanatory variables. Thus, the test is inapplicable in models of the following type:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \gamma Y_{t-1} + u_t \qquad (3.18)$$

   where, Y_{t-1} is the one period lagged value of Y.
6. There are no missing observations in the data.

The mechanics of the Durbin-Watson test are as follows, assuming that the assumptions underlying the test fulfilled:

1. Run the OLS regression and obtain the residuals.
2. Compute d from Eq (1)
3. For the given sample size and given number of explanatory variables, find out the critical d_L and d_U values.
4. The decision rules are also followed in the following table:

Durbin-Watson d test: Decision Rules

Null Hypothesis	Decision	If
No positive autocorrelation	Reject	$0 < d < d_L$
No positive autocorrelation	No decision	$d_L < d < d_U$
No negative autocorrelation	Reject	$4 - d_L < d < 4$
No negative autocorrelation	No decision	$4 - d_U < d < 4 - d_L$
No autocorrelation, positive or negative	Do not reject	$d_U < d < 4 - d_L$

3.5.3 Remedy of Autocorrelation between the Disturbances

In remedy of autocorrelation between the disturbances, there are two case: (1) ρ is known and ρ is not known. When ρ is known, generalized, or qusi, difference equation are used and when ρ is not known, first difference equation are used. Since the ρ can be gotten based on Durbin-Watson d statistic, the generalized difference equation are used in this study. The generalized difference equation are used to remedy of autocorrelation. Consider the k-variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + \cdots + \beta_k X_{t-k} + u_t \quad (3.19)$$

and assume that the error term follows the AR (1) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \quad (3.20)$$

If Eq (1) holds true at time t, it also holds true at time (t-1). Hence,

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + \cdots + \beta_k X_{t-1-k} + u_{t-1} \quad (3.21)$$

Multiplying eq (3.21) by ρ on both sides,

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \cdots + \rho\beta_k X_{t-1-k} + \rho u_{t-1} \quad (3.22)$$

Subtracting Eq (3.22) from Eq (3.19) gives

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \cdots + \beta_k(X_{t-k} - \rho X_{t-1-k}) + \varepsilon_t \quad (3.23)$$

Where, $\varepsilon_t = (u_t - \rho u_{t-1})$

Eq (3.23) can be expressed as

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t \quad (3.24)$$

where,

$\beta_1^* = \beta_1(1-\rho)$,$Y_t^* = (Y_t - \rho Y_{t-1})$, $X_t^* = (X_t - \rho X_{t-1})$ and $\beta_2^* = \beta_2$

Since the error term in Eq (3.24) satisfies the usual OLS assumptions, OLS can be applied to the transformed variables Y^* and X^* and estimators can be obtained with all the optimum properties, namely, BLUE. In this differencing procedure, one observation is lost because the first observation has no antecedent. To avoid this loss of one observation, the first observation on Y and X is transformed as follows: $Y_1\sqrt{1-\rho^2}$ and $X_1\sqrt{1-\rho^2}$.

If the first-difference transformation cannot be used because ρ is not sufficiently close to unity, an easy method of estimating ρ from the relationship between d and ρ has.

$$\rho \approx 1 - \frac{d}{2} \tag{3.25}$$

Thus, in reasonably large samples one can obtain ρ from Eq (3.25) and use it to transform the data as shown in generalized difference equation (3.23). The relationship between ρ and d given in Eq (3.25) may not hold true in small samples.

3.6 No Micronumersity

The problem of micronumerosity simply means small sample size. Micronumerosity is a situation whereby the sample size is not sufficient to obtain a precise (unbiased) estimate with relatively least standard errors. A regression model with Ordinary Least Squares (OLS) method cannot be estimated in a case of exact micronumerosity or having fewer observations than parameters to be estimated. Also, large standard errors have with near micronumerosity which means the number of observations barely exceeds the number of parameters to be estimated. Exact micronumerosity arises when n, the sample size, is zero, in which case any kind of estimation is impossible. Near micronumerosity arises when the number of observations barely exceeds the number of parameters to be estimated.

The number of observations n must be greater than the number of parameters to be estimated. If the sample size is less or equal to the number of predictors in a classical linear regression equations, it is impossible to estimate the regression parameters or fit an appropriate model to the data. If the sample size barely exceeds the number of predictors, there is lack of fit in the regression equation even if all other basic assumptions of classical linear regression hold.

3.6.1 Consequences of Micronumersity

Although the best linear unbiased estimator (BLUE), the OLS estimators have large variances and covariances, making precise estimation difficult. Thus, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” more readily. The t-ratio of one or more coefficients tends to be statistically insignificant. Although the t ratio of one or more coefficients is statistically insignificant, R^2, the overall measure of goodness of fit can be very high.

The OLS estimators and their standard errors can be sensitive to small changes in the data.

3.6.2 Detection of Micronumersity

The dependent variable is regressed on the independent variable (s) and the model diagnosis using Analysis of Variance (ANOVA) is also regressed. If the F-statistic computed is significant at 0.01 level of significance, then the sample size n used is accepted as the minimum sample size required to avert micronumerosity, otherwise, the sample size is rejected and another sample is taken by increasing the sample size until a significant model is obtained.

At the end of varying the sample size, the correlation between the dependent and independent variable(s) was also varied to see the effect of correlation on the sample size required.

3.6.3 Remedy of Micronumersity

If there is presence of micronumersity in a data set, then additional data should be obtained. If it is not possible to increase the sample size, then the best method of remedying micronumersity is to use factor analysis regression or principal component regression.

3.7 The Nature of X Variables

The X value in a given sample must not all be the same, that is, the variables must vary. Technically, var (X) must be a positive number. The variation in both X and Y is essential to use regression analysis as a research tool. Furthermore, there can be no outliers in the values of X variable. From a practical perspective, two types of outliers are problematic in regression analysis. On the one hand, some observations may fail to be predicted by the model with a reasonable degree of accuracy. This type of outliers may reveal the fact that several populations are mixed in the data set or that some relevant variables have been omitted. On the other hand, some observations may be influential in the sense that their presence in the data set substantially modifies the estimates. This type of outliers weakens the conclusions which may be drawn from the model.

3.7.1 Consequences of the Nature of X Variables

If all the X values are identical, then $X_i = \bar{X}$ and the denominator of that equation will be zero, making it impossible to estimate β_2 and β_1. If there is very little variation in X, much of the variation in Y can be explain. When there is outliers in the values of X variables, this observation influence its own prediction and its means and inflate the standard error with which it is being standardized.

3.7.2 Detection of the Nature of X Variables

Box plot is the best way in detecting the outliers. If the dot has outside the box plot, this plot is assumed that have the outliers. Cook's distance (Cook, 1977) is the usual statistic which is employed to detect influential observations. A particular multivariate distribution model should be assumed for X, which would often be unreasonable. Cook (1977) and Weisberg (1980) suggest using the 50^{th} percentile of the F distribution with k and N-k-1 degree of freedom.

3.7.3 Remedy of the Nature of X Variables

When the data has the outliers or influential observations, this observation is dropped. For each parameter's confidence interval, one could report the lowest value for the lower limits and the highest value for the upper limits found when dropping different observations (Leamer, 1979). The Cook's distance does not use any limits at all but drawing a scree plot of Cook's distances ordered from highest to lowest. This plot will be useful to separate the few most influential observations from the many least influential ones. A sensitivity analysis of the estimates should then be carried out by hand by sequentially dropping the identified observations and qualitatively evaluating the extent to which the conclusions to be drawn from the model change. This is what ultimately counts when evaluating influence of the observations and is far more useful than blindly using a fixed critical value. The removal of the observations with the highest Cook's distances did lead to some substantial changes in the model.

3.8 No Multicollinearity between the X variables

Multicollinearity (also called collinearity) refers to the assumption that the independent variables are uncorrelated (Darlington, 1968; Keith, 2006). The researcher is able to interpret regression coefficients as the effects of the independent variables on the dependent variables when collinearity is low (Keith, 2006; Poole &

O'Farrell, 1971). This means that inferences about the causes and effects of variables can be made reliably. Multicollinearity occurs when several independent variables correlate at high levels with one another, or when one independent variable is a near linear combination of other independent variables (Keith, 2006). The more variables overlap (correlate) the less able researchers can separate the effects of variables. In MR the independent variables are allowed to be correlated to some degree (Cohen, 1968; Darlington, 1968; Hoyt et al., 2006; Neale et al., 1994). The regression is designed to allow for this, and provides the proportions of the overlapping variance (Cohen, 2968). Multicollinearlity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set.

Many sources of multicollinearity are

1. The data collection method employed.
2. Constraints on the model or in the population being sampled.
3. Model Specification.
4. An overdetermined model.

3.8.1 Consequences of Multicollinearity

Multicollinearity can result in misleading and unusual results, inflated standard errors, reduced power of the regression coefficients that create a need for larger sample sizes (Jaccard et al., 2006; Keith, 2006). Interpretations and conclusions based on the size of the regression coefficients, their standard errors, or associated t-tests may be misleading because of the confounding effects of collinearity (Mason & Perreault Jr., 1991). The result is that the researcher can underestimate the relevance of a predictor, the hypothesis testing of interaction effects is hampered, and the power for detecting the moderation relationship is reduced because of the intercorrelation of the predictor variables (Jaccard et al., 2006; Shieh, 2010).

3.8.2 Detection of Multicollinearity

High R^2 but few significant t ratios, high pair- wise correlations among regressors, examination of partial correlations, auxiliary regression, eigenvalues and condition index, tolerance and variance inflation factor are tools to detect the multicollinearity. The effect of a given level of collinearity can be evaluated in conjunction with the other factors of sample size, R2, and magnitude of the

coefficients (Mason & Perreault Jr., 1991). Tolerance measures the influence of one independent variable on all other independent variables. Tolerance levels for correlations range from zero (no independence) to one (completely independent) (Keith, 2006). The VIF is an index of the amount that the variance of each regression coefficient is increased over that with uncorrelated independent variables (Keith, 2006). When a predictor variable has a strong linear association with other predictor variables, the associated VIF is large and is evidence of multicollinearity (Shieh, 2010). Small values for tolerance and large VIF values show the presence of multicollinearity (Keith, 2006).

3.8.3 Remedy of Multicollinearity

The remedial measures are a priori information, dropping a variable (s) and specification bias, transformation of variables and additional or new data, etc. When faced with several multicollinearity, one of the simplest things to do is to drop one of the collinear variables. When one of the collinear variables are dropped, the regression which shows that whereas in the original model the variable was statistically insignificant, it become highly significant is obtained.

But in dropping a variable from the model, a specification bias or specification error may be committed. Specification bias arises from incorrect specification of the model used in the analysis. Dropping a variable from the model to alleviate the problem of multicollinearity may lead to the specification bias. Hence, the remedy may be worse than the disease in some situations because, whereas multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may seriously mislead us as to the true values of the parameters.

3.9 No Specification Errors or No Specification Bias

One of the assumptions of the classical linear regression model (CLRM) is that the regression model used in the analysis is correctly specified: If the model is not correctly specified, the problem of model specification error or model specification bias. When one of the assumptions is violated, specification error is caused. A model chosen for empirical analysis should satisfy the following criteria:

1. Be data admissible.
2. Be consistent with theory.
3. Have weakly exogenous regressors.

4. Exhibit parameter constancy.
5. Exhibit data coherency.
6. Be encompassing.

In practice one is likely to commit various model specification errors. The type of specification errors are

1. Omission of a relevant variable(s).
2. Inclusion of an unnecessary variable(s).
3. Adoption of the wrong functional form.
4. Errors of measurement.
5. Incorrect specification of the stochastic error term.
6. Assumption that the error term is normally distributed.

3.9.1 Consequences of Specification Errors

The consequences of some type of specification errors are described as follow.

(i) The consequences of omitting variable are:

1. If the left-out, or omitted, variable is corrected with the included variable, the correlation coefficient between the two variables is nonzero and the coefficients in the underfitting model are biased as well as inconsistent. The bias does not disappear as the sample size gets larger.
2. Even if the independent variables are not correlated each other, the slope coefficient in the underfitting model is biased, although the intercept coefficient in this model is unbiased.
3. The disturbance variance is incorrectly estimated.
4. The conventionally measured variance of the intercept coefficient in the underfitting model is a biased estimator of the variance of the intercept coefficient in the true model.
5. The usual confidence interval and hypothesis testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.
6. The forecasts based on the incorrect model and the forecast confidence intervals will be unreliable.

(ii) The consequences of inclusion of an irrelevant variable are:

1. The OLS estimators of the parameters of the incorrect model are all unbiased and consistent.
2. The error variance is correctly estimated.
3. The usual confidence interval and hypothesis testing procedures remain valid.
4. However, the estimated parameters in overfitting model will be generally inefficient.

(iii) The consequences of errors of measurement are:

1. Although the errors of measurement in the dependent variable still give unbiased estimates of the parameters and their variances, the estimated are now larger than in the case where there are no such errors of measurement.
2. The errors of measurement in the explanatory variable can be shown that the OLS estimators are not only biased but also inconsistent, they remain biased even if the sample size increases indefinitely.

(iv) The consequences of other specification error are that the parameter is biased estimator, inconsistent, large standard error.

3.9.2 Detection of Specification Errors

(i) Detecting the Presence of Unnecessary Variables

Suppose a k-vaiables model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad (3.26)$$

In given model, whether one or more regressors are really relevant by the usual t and F tests. The t and F tests should not be used to build a model iteratively, that is, initially Y is not related to X_2 only because $\hat{\beta}_2$ is statistically significant and then expand the model to include X_3 and decide to keep that variable in the model if $\hat{\beta}_3$ turns out to be statistically significant and so on. This strategy of building a model is called data mining by the somewhat pejorative term.

(ii) Detecting the Omitted Variables and Incorrect Functional Form

Examination of residuals is a good visual diagnostic to detect autocorrelation or heteroscedasticit. But these residuals can also be examined especially in cross-sectional data for the model speciation errors such as omission of an important

variable or incorrect functional form. If in fact there are such errors, a plot of the residuals will exhibit distinct patterns.

By the Durbin-Watson d statistic Once Again, from Durbin-Watson tables if the estimated d value is significant, then one can accept the hypothesis of model mis-specification. The other test such as Ramsey's RESET test and Lagrange Multiplier (LM) test for adding variables are also used to detect omitted variable and incorrect functional form.

Most specification errors such as normality of the disturbances, error of measurement are also the same with some assumptions. The violation of the assumptions causes specification errors. Hence, detecting ways of the assumptions can also be used.

3.9.3 Remedy of Specification Errors

The way of the remedy for the presence of unnecessary variables is the dropping unnecessary variables and the way of the remedy for omitted variables and incorrect functional form are the adding the important variables and the using changed the function form. One way of the remedy for error of measurement is the use of instrumental or proxy variables that although highly correlated with the original X variables, are correlated with the equation and measurement of error terms. If such proxy variables can be found, then one can obtain a consistent estimate of β. The other specification errors type remedy as the ways of some assumptions that are same.

3.10 Normality of the Disturbance Terms

Multiple regression assumes that variables have normal distributions (Darlington, 1968; Osborne & Waters, 2002). This means that errors are normally distributed, and that a plot of the values of the residuals will approximate a normal curve (Keith, 2006). The assumption is based on the shape of normal distribution and gives the researcher knowledge about what values to expect (Keith, 2006). Once the sampling distribution of the mean is known, it is possible to make predictions for a new sample (Keith, 2006).

The several reasons that employ the normality assumption:

1. The u_i represent the combined influence of a large number of independent variables that are not explicitly introduced in the regression model.

2. A variant of central limit theorem (CLT) states that, even if the number of variables is not very large or if these variables are not strictly independent, their sum still be normally distributed.
3. With the normality assumption, the probability distributions of OLS estimators can be easily derived.
4. The normal distribution is a comparatively simple distribution involving only two parameters (mean and variance).
5. If sample size is small or finite, the normality assumption assumes a critical roles. It not only helps to derive the exact probability distributions of OLS estimators but also enables to use the t, F and χ^2 statistical tests for regression models.
6. In large samples, t and F statistics have approximately the t and F probability distributions so that the t and F tests that are based on the assumption that the error term is normally distributed can still be applied validly.

3.10.1 Consequences of Violation of Normality of the Disturbance Terms

When scores on variables are skewed, correlations with other measures will be attenuated, and when the range of scores in the sample is restricted relative to the population correlations with scores on other variables will be attenuated (Hoyt et al., 2006). Non-normally distributed variables can distort relationships and significance tests (Osborne & Waters, 2002). Outliers can influence both Type I and Type II errors and the overall accuracy of results (Osborne & Waters, 2002).

3.10.2 Detection of Normality of the Disturbance Terms

The researcher can test this assumption through several pieces of information: visual inspection of data plots, skew, kurtosis, and P-Plots (Osborne & Waters, 2002). Data cleaning can also be important in checking this assumption through the identification of outliers. Statistical software has tools designed for testing this assumption. Skewness and kurtosis can be checked in the statistic tables, and values that are close to zero indicate normal distribution. Normality can further be checked through histograms of the standardized residuals (Stevens, 2009).

3.10.3 Remedy of Violation of Normality of the Disturbance Terms

If normality assumption cannot be maintain on the basis of various normality tests, various transformation ways such as double log transformation, semi log transformation, square root transformation and reflect and square root transformation are used or bootstrapping is used. The basic idea underlying bootstrapping is to churn a given sample over and over again and then obtain the sampling distributions of the parameters of interest.

CHAPTER IV

DETECTING AND REMEDY OF THE ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION MODEL

In this chapter, descriptive analysis for production, pesticides, weir and quality seeds of Maize, for production, pesticides and substitution sown of Wheat, for production, sown and yield of Rice and for production, sown, harvested, irrigation, loan, quality seeds and pesticides of Sesame are described. Moreover, detecting and remedy of the assumptions of multiple linear regression model for each of crops (Maize, Wheat, Rice and Sesame) are also presented.

4.1 Descriptive Data Analysis

In this section, mean value, standard deviation (Std.Dev), minimum value (Min) and maximum value (Max) of each crop in Myanmar are expressed. The descriptive Statistics for production (PROD), pesticides (PESTI), weir (WR) and quality seeds (QUALI) of Maize in Myanmar is shown in following Table (4.1).

Table (4.1)

Descriptive Statistics for Maize in Myanmar

Variable	Mean	Std.Dev	Min	Max
PROD	1068.71	542.17	297.9	1909
PESTI	78570.67	210229.42	111	910543
WR	25.67	8.47	7	35
QUALI	8894.05	13461.00	1192	60641

Source: Appendix (B11)

In Table (4.1), maize production ranges between 297.9 ton and 1909 ton with mean 1068.71 ton and standard deviation 542.17 ton. Weir ranges between 7 thousand acres and 35 thousand acres with mean 25.67 thousand acres and with standard deviation 8.47 thousand acres. Pesticides ranges between 111 gallon and 910543 gallons with mean 78570.67 gallons and with standard deviation 210229.42 gallons and quality seeds ranges between 1192 baskets and 60641 baskets with mean 8894.05 baskets and with standard deviation 13461 baskets.

The descriptive Statistics for production (PROD), pesticides (PESTI) and substitution sown (SUBSOWN) of Wheat in Myanmar is presented in Table (4.2).

Table (4.2)

Descriptive Statistics for Wheat in Myanmar

Variable	Mean	Std.Dev	Min	Max
PROD	144.08	34.66	90.7	182.9
PESTI	290.00	230.98	22	894
SUBSOWN	7105.90	3227.94	665	13814

Source: Appendix (B12)

In Table (4.2), wheat production ranges between 90.7 ton and182.9 ton with mean 144.08 ton and standard deviation 34.66 ton. Pesticides ranges between 22 gallons and 894 gallons with mean 290 gallons and with standard deviation 230.98 gallons and substitution sown ranges between 665 thousand acres and 13814 thousand acres with mean 7105.9 thousand acres and with standard deviation 3227.94 thousand acres.

The descriptive Statistics for production (PROD), pesticides (PESTI) and yield (YIE) for Rice in Myanmar is presented in Table (4.3).

Table (4.3)

Descriptive Statistics for Rice in Myanmar

Variable	Mean	Std.Dev	Min	Max
PROD	17311.47	7899.68	6636.76	32164
SOWN	14567.92	2780.29	11530	20076
YIE	56.88	15.78	28.5	78.91

Source: Appendix (B13)

In Table (4.3), the rice production ranges between 6636.76 ton through 32164 ton with mean equal 17311.47 ton and standard deviation 7899.68 ton and sown ranges between 11530 thousand acres through 20076 thousand acres with mean 14567.92 thousand acres and with standard deviation equal 2780.29 thousand acres and yield ranges between 28.5 lb through 78.91 lb with mean 56.88 lb and with standard deviation 15.78 lb.

The descriptive Statistics for production (PROD), sown (SN), harvested (HAR), irrigation (IRRI), loan (LA), quality seeds (QUALI) and pesticides (PESTI) for Sesame in Myanmar is represented as shown in Table (4.4).

Table (4.4)

Descriptive Statistics for Sesame in Myanmar

Variable	Mean	Std. Dev	Min	Max
PROD	497.97	262.05	142.7	840
SN	3499.77	396.58	2557	4052
HAR	2976.20	723.63	1521	3863
IRRI	210.76	49.71	147.76	335.49
LA	7034.38	12645.55	59.72	58018.6
QUALI	1283.77	1216.96	68	4464
PESTI	64238.77	187153.36	111	928447

Source: Appendix (B14)

In Table (4.4), the sesame production ranges between 840 ton and 142.7 ton with mean 497.97 ton and standard deviation 262.05 ton. Sown ranges between 2557 acres and 4052 acres with mean 3499.77 acres and with standard deviation 396.58 acres and harvested ranges between 1521 acres and 3863 acres with mean 2976.2 acres and with standard deviation 723.63 acres. Irrigation ranges between 147.76 acres and 335.49acres with mean 210.76 acres and with standard deviation 49.71 acres and loan ranges between 59.72 kyats in thousand and 58018.6 kyats in thousand with mean 7034.38 kyats in thousand and with standard deviation 7034.38 kyats in thousand. Quality seeds range between 68 baskets and 4464 baskets with mean 1283.77 baskets and with standard deviation 1216.96 baskets and pesticides ranges between 111 gallons and 928447 gallons with mean 64238.77 gallons and with standard deviation 187153.36 gallons.

4.2 Detecting and Remedy of Linearity Assumption

Based on the time series data of maize production, pesticides and quality seeds during the period of 1998 to 2018, the proposed multiple regression model of maize

production in Myanmar on pesticides, weir and quality seeds are fitted. The dependent variable (Y_i) is production (PROD) and the independent variables (X_{ij}) are pesticides (PESTI), weir (WR) and quality seeds (QUALI).The regression equation results are shown in the following Table (4.5).

Table (4.5)

Results for Original Data of Maize Production

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF	Cov(X_{ij},u_i)
Constant	110.466	263.236	0.420	0.680			
PESTI	0.001	0.000	2.849	0.011**	0.914	1.094	0.07
WR	37.746	8.937	4.223	0.001***	0.626	1.597	0.00
QUALI	-0.009	0.005	-1.582	0.132	0.657	1.523	-0.01
R-squared			0.793	E (u_i)			0.00
Adjusted R-squared			0.756	F-statistic			21.646
Std. Error of the Estimate			267.857	Prob (F-statistic)			0.00***
Durbin-Watson			0.499				

*** denotes significant at 1% level and ** denotes significant at 5% level

Source: Appendix (A1)

The estimated regression equation for maize production is

$$PROD = 110.466 + 0.001PESTI + 37.746WR - 0.009QUALI \quad (4.1)$$

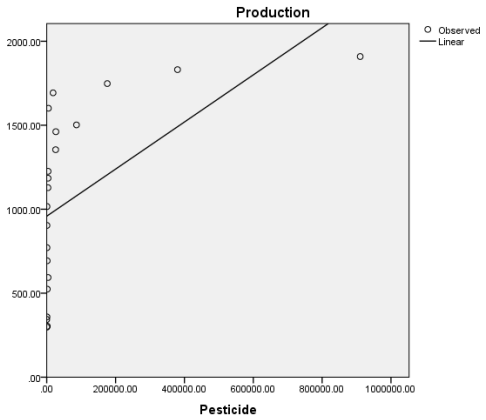
From the estimated regression equation (4.1) for maize production, it is found that maize production is positively related to pesticides and weir, and negatively related to quality seeds. It can be that holding the quality seeds and weir of maize is held constant; a 1 gallon increase in pesticides led on the average to about 0.001 ton increase in production. Similarly, holding the quality seeds and pesticides is held constant, a 1 basket increase in weir led on the average to about 37.746 thousand acreages increase in production and holding the pesticides and weir is held constant, a 1 basket increase in quality seeds led on the average to about 0.009 thousand acreages decrease in production.

The pesticides (2.849) is statistically significant at 5% level and weir (4.223) is also statistically significant at 1% level but quality seeds is not significant. The multiple linear regression's F-test (21.646) is highly significant at 1% level. Adjusted R-square is 0.756 and the R squared is 0.793 which means that 79.3% of variation in maize production is explained by pesticides, weir and quality seeds and the remaining percentage 20.7% due to other factors that are not included in the model. The standard error of estimate (267.857) is very large.

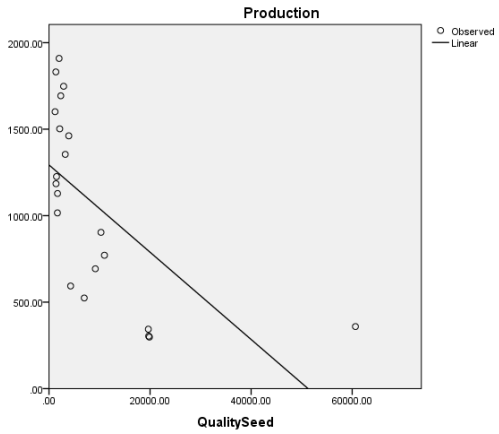
The value of Durbin-Watson statistic DW = 0.499 is less than the table value of the lower limit for Durbin-Watson statistic d_L = 1.026 at 95% confidence limit with the number of parameter 3. Therefore, there is autocorrelation. The mean of error is zero and the covariance values of error term and independent variables are 0.07, 0 and -0.01 which can be assumed to zero. Moreover, since VIF values of pesticides, weir and quality seeds are 1.094, 1.597 and 1.523 which is not greater than 5 and total VIF is not greater than 10 and tolerance values are 0.914, 0.626 and 0.657 which are closely to 1. Hence, there is no multicollinearity. In this section, since quality seeds is not significant and standard error is very large, assume that linearity assumption between production and pesticides, weir and quality seeds is violated. Then, detect the violation of the linearity assumption.

4.2.1 Detection of Linearity Assumption

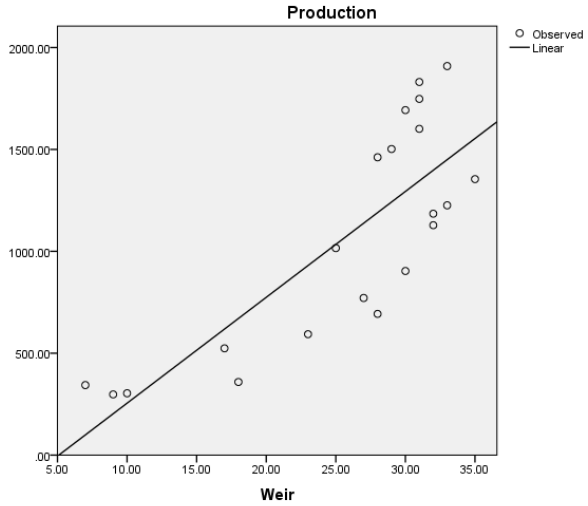
One of the assumptions of classical linear regression analysis is that the regression model is linear in the parameters. The relationship between the dependent and each independent variable need to be linear. The scatter plots between the dependent and each independent variable and residual plot showing the standardized residuals versus the predicted values are used to diagnosis linearity assumption in the following Figures (4.1) and (4.2).



(a)



(b)



(c)

Figure (4.1) Scatter Plots for Original Data

Source: Appendix (A1)

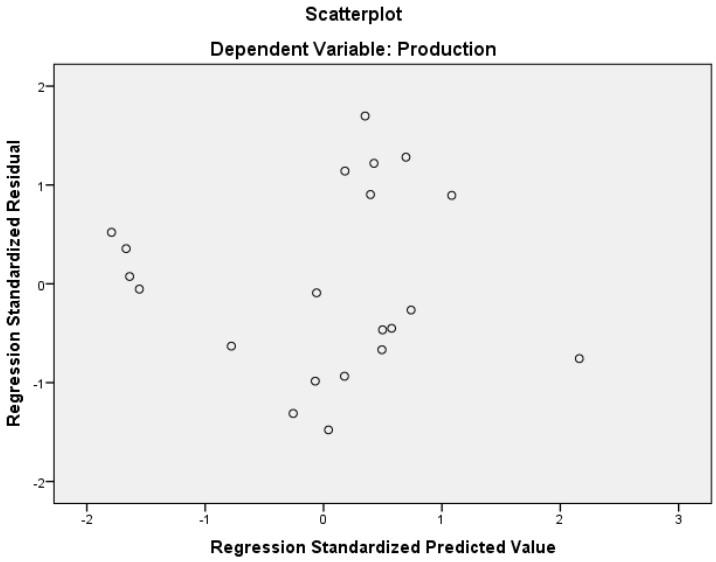


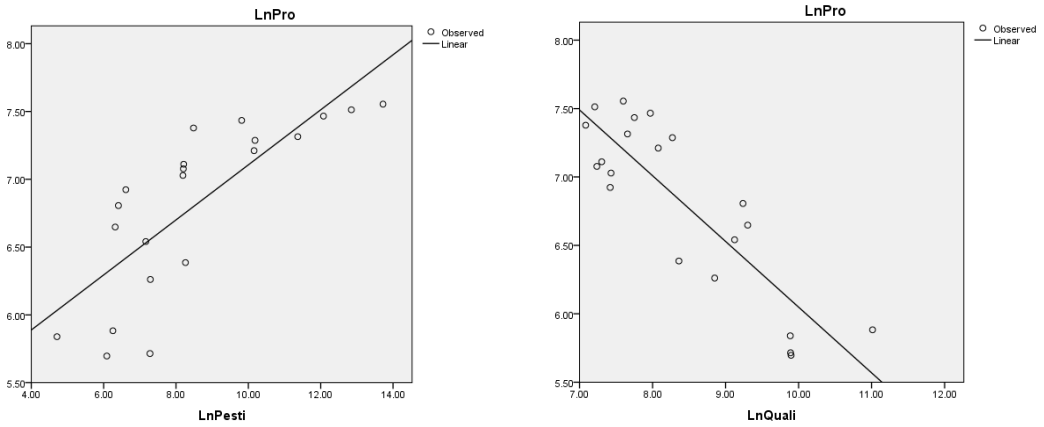
Figure (4.2) Residual Plot of Original Data

Source: Appendix (A1)

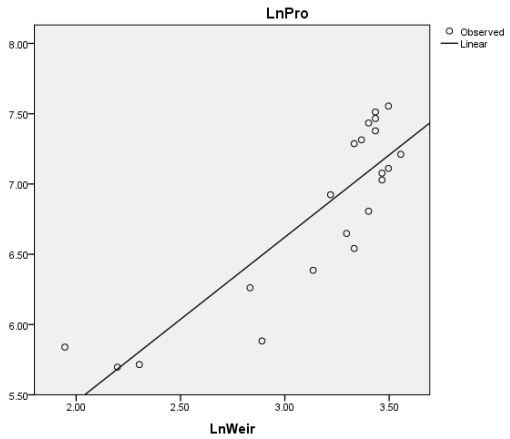
Linear relationship between production and pesticides, between production and weir between production and quality seeds are indistinct in Figures (4.1) (a), (b) and (c) and Figure (4.2) can be seen as the curve pattern. Therefore, these data can be concluded that violated the linearity assumption.

4.2.2 Remedy of Linearity Assumption

The variables are needed to transform to remedy the violation of linearity assumption. So, the variables are transformed by taking the double-log (log - log). The scatter plots and residual plot for transformation data are described the Figures (4.3) and (4.4).



(a) (b)



(c)

Figure (4.3) Scatter Plots for Transformation Data

Source: Appendix (A1)

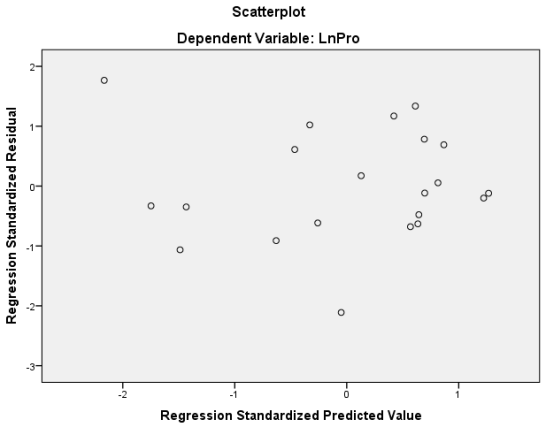


Figure (4.4) Residual Plot of Transformation Data

Source: Appendix (A1)

Linear relationship between production and pesticides, between production and weir, between production and quality seeds are obvious in Figures (4.3) (a), (b) and (c). In Figure (4.4), residual plot showing the standardized residuals versus the predicted values has no pattern. The double log model satisfies the linearity assumption between dependent and independent variables and the regression model is linear in the parameters.

4.2.3 Fitted Regression Model for Maize Production

Table (4.6) shows the results of transformation model by taking the double log variables of Maize production as follows.

Table (4.6)
Results for Transforming Data of Maize Production

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	6.000	0.843	7.120	0.000***		
Ln(PESTI)	0.079	0.023	3.340	0.003***	0.570	1.755
Ln(WR)	-0.201	0.058	-3.494	0.003***	0.422	2.369
Ln(QUALI)	0.579	0.134	4.304	0.000***	0.450	2.220
R-squared			0.923	F-statistic		68.378
Adjusted R-squared			0.910	Prob (F-statistic)		0.000***
Std. Error of the Estimate			0.188			
Durbin-Watson			1.272			

*** denotes significant at 1% level

Source: Appendix (A1)

The best fitted regression model for maize production is

$$\text{Ln (PROD)} = 6 + 0.079\text{Ln (PESTI)} - 0.201\text{Ln (WR)} + 0.579\text{Ln (QUALI)} \quad (4.2)$$

The pesticides (2.849), weir (4.223) and quality seeds are statistically significant at 1% level. The multiple linear regression's F-test (68.378) is highly significant at 1% level. Adjusted R-square is 0.923 and the R squared is 0.910 which means that 91% of variation in maize production is explained by pesticides, weir and quality seeds and the remaining percentage 9% due to other factors that are not included in the model. The standard error of estimate (0.188) is small.

The value of Durbin-Watson statistic DW = 1.272 exists between the table value of the lower limit d_L = 1.026 and upper limit d_U = 1.669 at 95 % confidence limit with the number of parameter 3. Therefore, there is no autocorrelation. Moreover, since VIF values of pesticides, weir and quality seeds 1.755, 2.369 and 2.220 which is not greater than 5 and total VIF is not greater than 10 .So, there is no multicollinearity. This model satisfies the linear in parameter.

4.3 Detecting and Remedy of Normality Assumption between the Disturbances

Based on the time series data of wheat production, pesticides and substitution sown during the period of 1998 to 2018, the proposed multiple regression model of wheat production in Myanmar on pesticides and substitution sown are fitted. The regression equation results are shown the following Table (4.9). The dependent variable (Y_i) is production (PROD) and the independent variables (X_{ij}) are pesticides (PESTI) and substitution sown (SUBSOWN). The regression equation results are shown in the following Table (4.7).

Table (4.7)

Results for Original Data of Wheat Production

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF	Cov(xij,ui)
Constant	106.430	18.199	5.848	0.000***			
PESTI	0.031	0.032	0.978	0.341	0.944	1.059	-0.0001
SUBSOWN	0.004	0.002	1.747	0.098*	0.944	1.059	-0.002
R-squared		0.221	E (u_i)			0.00	
Adjusted R-squared		0.134	F-statistic			2.55	
Std. Error of the Estimate		32.246	Prob (F-statistic)			0.106	
Durbin-Watson		0.828					

*** denotes significant at 1% level and * denotes significant at 10% level

Source: Appendix (A2)

The estimated regression equation for wheat production is

PROD = 106.430 + 0.031PESTI – 0.004SUBSOWN. (4.3)

From the estimated regression equation for wheat production, it is found that wheat production is positively related to pesticides and substitution sown. It was found that holding the substitution sown is constant; a 1 gallon increase in pesticides led on the average to about 0.031 ton increase in production. Similarly, holding the pesticides is constant; a 1 acreage increase in yield led on the average to about 0.004 thousand acreages increase in production.

The substitution sown (1.747) is statistically significant at 10% level but pesticides (0.978) is not significant. The F-test of multiple linear regression (2.55) is highly significant at 1% level. Adjusted R-square is 0.134 and the R squared is 0.221 which means that 22.1% of variation in wheat production is explained by pesticides, substitution sown and the remaining percentage 77.9% due to other factors that are not included in the model. The standard error of estimate is 32.246.

The value of Durbin-Watson statistic DW = 0.828 is less than the table value of the lower limit for Durbin-Watson statistic d_L = 1.125 at 95% confidence limit with the number of parameter 2. Therefore, there is autocorrelation. The mean of error is zero and the covariance values of error term and independent variables are -0.0001 and -0.002 which can be assumed to zero. Moreover, since both VIF values of pesticides an substitution sown are 1.059 which is not greater than 5 and total VIF is not greater than 10 and tolerance values are 0.944 which are closely to 1. So, there is no multicollinearity. In this section, normality assumption between the disturbances is detected.

4.3.1 Detection of Normality of Disturbance Terms

One of the basic assumptions is that the stochastic disturbance terms are normally distributed. The violation of normality of the disturbance terms are detected by using Kolmogorov-Smirnov test and box plot. The Kolmogorov-Smirnov test and box plot are shown in Table (4.8) and Figure (4.5).

Table (4.8)

Kolmogorov-Smirnov Test for Original Data

	Statistic	Degree of freedom	Sig.
Standardized Residual	0.202	21	0.026**

** denote significant at 5% level.

Source: Appendix (A2)

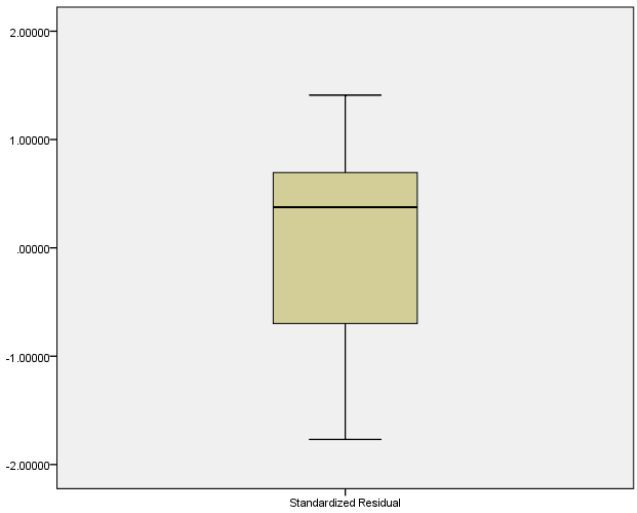


Figure (4.5): Box Plot of Standardized Residual for Original Data

Source: Appendix (A2)

In Table (4.8) based on the Kolmogrov-Smirnov test, the computed value of Kolmogrov-Smirnov test for original data is 0.202. Since the p-value 0.026 is less than alpha value (α) 0.05, reject the null hypothesis that the stochastic disturbance terms are normally distributed. In Figure (4.5), box plot is skewed. Therefore, normality of the stochastic disturbance terms cannot be assumed in these data.

4.3.2 Remedy Violation of Normality of the Disturbance Terms

The redefining variables are needed to remedy the violation of normality of the disturbances. The variables are transformed by taking reflect and square root. The Kolmogrov-Smirnov test and histogram are described as the following Table (4.9) and Figure (4.6).

Table (4.9)

Kolmogorov-Smirnov Test for Redefining Variables

	Statistic	Degree of freedom	Sig.
Standardized Residual	0.132	21	0.200

Source: Appendix (A2)

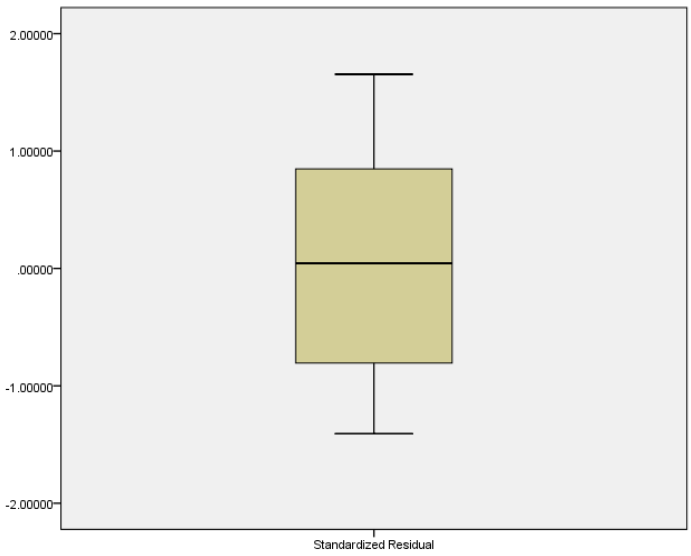


Figure (4.6): Box Plot of Standardized Residual for Redefining Data

Source: Appendix (A2)

In Table (4.9) based on the Kolmogrov-Smirnov test, the computed significance level for transforming data 0.2 is greater than alpha value (α) 0.05 and in Figure (4.6), box plot is symmetric. Therefore, normality of the standardized residual can be assumed in these data.

4.3.3 Fitted Regression Model for Wheat Production

The variables are transformed by taking reflect and square root. Table (4.10) shows the results of these transformation variables for Wheat production as follows.

Table (4.10)

Results for Transforming Data of Wheat Production

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	0.339	2.709	0.125	0.902		
New (PESTI)	0.041	0.090	0.455	0.655	0.966	1.035
New (SUBSOWN)	0.054	0.025	2.132	0.047**	0.966	1.035
R-squared		0.227	F-statistic		2.642	
Adjusted R-squared		0.141	Prob (F-statistic)		0.099*	
Std. Error of the Estimate		2.850				
Durbin-Watson		0.760				

*denote significant at 5% level and *denote significant at 10% level

Source: Appendix (A2)

New (PROD) = Sqrt (K-PROD)

New (PESTI) = Sqrt (K-PESTI)

New (SUBSOWN) = Sqrt (K-SUBSOWN)

K = reflect = 1+ max (respective value)

The fitted regression model for wheat production is

New (PROD) = 0.041 New (PESTI) + 0.054 New (SUBSOWN) (4.4)

The F-test of multiple linear regression (2.642) is significant at 10% level. Adjusted R-square is 0.141 and the R squared is 0.227 which means that 22.7% of variation in wheat production is explained by pesticides, substitution sown and the remaining percentage 77.3% due to other factors that are not included in the model. The standard error of estimate decrease from 32.246 to 2.850.

The value of Durbin-Watson statistic DW = 0.760 is less than the table value of the lower limit for Durbin-Watson statistic d_L = 1.125 at 95% confidence limit with the number of parameter 2. Therefore, there is autocorrelation. The mean of error is zero. Moreover, since both VIF values of pesticides an substitution sown are 1.035 which is not greater than 5 and total VIF is not greater than 10 and tolerance values are 0.966 which are closely to 1. Hence, there is no multicollinearity. In this section, normality assumption between the disturbances is detected.

4.4 Detection and Remedy of Homoscedasticity Assumption

Based on the time series data of rice production, sown acreage and yield per harvested acre during the period of 1966 to 2018, the proposed multiple regression model of rice production in Myanmar on the sown acreage and yield per harvested are fitted. The regression equation results are shown the following Table (4.11).

Table (4.11)

Results for Original Data of Rice Production

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	-20282.5	430.34	-47.13	0.00***		
SN	1.662	0.046	35.846	0.00***	0.383	2.613
YIE	235.276	8.169	28.801	0.00***	0.383	2.613
R-squared		0.995	$E(u_i)$		0.000	
Adjusted R-squared		0.995	F-statistic		4881.215	
Std. Error of the Estimate		575.07	Prob (F-statistic)		0.000***	
Durbin-Waston		0.920				

*** denote significant at 1% level.

Source: Appendix (A3)

The estimated regression equation for rice production is

PROD = -20282.5 + 1.662SN + 235.276YIE. (4.5)

From the estimated regression equation for rice production, it is found that rice production is positively related to sown and yield. It can be found that holding the yield of rice is constant; a 1 thousand acres increase in sown led on the average to about 1.662 ton increase in production. Similarly, holding the sown of rice is constant, a 1 thousand acres increase in yield led on the average to about 235.276 thousand acreages increase in production.

The sown (35.846) and yield (28.801) is statistically significant at 1% level. The F-test of multiple linear regression (2.55) is highly significant at 1% level. Adjusted R-square is 0.995 and the R squared is 0.995 which means that 99.5% of variation in rice production is explained by sown and yield and the remaining percentage 0.5% due to other factors that are not included in the model. The standard error of estimate is 575.07. The value of Durbin-Watson statistic DW = 0.920 is less than the table value of the lower limit for Durbin-Watson statistic d_L = 1.490 at 95% confidence limit with the number of parameter 3. Therefore, there is autocorrelation. The mean of error is zero. Moreover, since both VIF values of sown and yield are 2.613 which is not greater than 5 and total VIF is not greater than 10. Hence, there is no multicollinearity. In this section, homoscedasticity assumption are diagnosed.

4.4.1 Detection of Homoscedasticity Assumption

One of the assumptions of classical linear regression model is that the variance of the disturbance term is constant or homoscedasticity, that is, there is no heteroscedasticity.is In this section, White's General Heteroscedasticity test are used to diagnosis the homoscedasticity assumption. The following Table (4.12) shows the White's General Heteroscedasticity test result for the original data.

Table (4.12)

White's General Heteroscedasticity Test Result for Original Data

Variable	Coefficient	Std. Error	t-statistic	Sig.
Constant	6.943	12.865	0.540	0.592
SN	0.000	0.001	-0.127	0.899
YIE	-0.220	0.255	-0.864	0.392
SN^2	2.552E-008	0.000	0.309	0.759
YIE^2	0.005	0.004	1.361	0.180
SN*YIE	-1.754	0.000	-0.422	0.675
R Squared	0.237	Adjusted R Squared		0.156
Chi Square	11.07	Probability		0.05
Obs*R-squared	12.561	Probability		0.05

Source: Appendix (A3)

In this problem, the number of observations is 53 and number of parameters is 5. The R-squared equals 0.237 which means that 23.7% of variation in rice production is explained by sown and yield. Since observation * R squared (12.561) is greater than the Chi squared distribution with 5 degree of freedom at 5% level (11.07), there reject the null hypothesis that the disturbance terms have equal variances. Hence, heteroscedasticity exists.

4.4.2 Remedy of Homoscedasticity Assumption

The variables which make the violation of the homoscedasticity assumption are transformed to the double-log equation (log - log model) to remedy the heteroscedasticity. The redefining results are shown in following Table (4.13).

Table (4.13)

Redefining Results for Rice Production

Variable	Coefficient	Std. Error	t-statistic	Sig.	TOL	VIF
Constant	-4.923	0.306	-16.133	0.000***		
Ln (SN)	1.114	0.038	29.375	0.000***	0.480	2.085
Ln (YIE)	0.978	0.022	44.258	0.000***	0.480	2.085
R-squared		0.995	F-statistic		4896.133	
Adjusted R-squared		0.995	Sig. (F-statistic)		0.000***	
Std. Error of the Estimate		0.034				
Durbin-Waston		1.472				

***denote significant at 1% level.

Source: Appendix (A3)

Table (4.13) shows redefining result by redefining the variables which make the violation of homoscedasticity such that production, sown and yield are transformed to double - log equation.

The ln of sown (29.375) and ln of yield (44.258) are statistically significant at 1% level. The F-test of multiple linear regression (4896.133) is highly significant at 1% level. Adjusted R-square is 0.995 and the R squared is 0.995 which means that 99.5% of variation in rice production is explained by sown and yield and the remaining percentage 0.5% due to other factors that are not included in the model. The standard error of estimate is 575.07. The Durbin-Watson d statistic is 1.472. In addition, White's General Heteroscedasticity test is used to diagnose the problem of heteroscedasticity for redefining data. The following Table (4.14) shows the White's General Heteroscedasticity result for redefining data.

Table (4.14)

White's General Heteroscedasticity Test Results for Redefining Data

Chi Square	11.07	Probability	0.05
Obs*R-squared	8.162	Probability	0.05
R Squared	0.154	Adjusted R Squared	0.102

Source: Appendix (A3)

The R-squared equals 0.154 which means that 15.4% of variation in redefining rice production is explained by redefining sown and yield. Since Obs * R-squared (8.162) is greater than the Chi squared (11.07) with 5 degree of freedom, do not evidence to reject the null hypothesis that the error term have equal variances. Hence, heteroscedasticity does not exist. Thus, the redefining method is successfully solved the violation in constant variance in the original data.

The best fitted multiple regression model for rice production is

$$\text{Ln (PROD)} = -4.923+1.114 \text{ Ln (SN)} +0.978 \text{ Ln (YIE)} \quad (4.5)$$

4.5 Detecting and Remedy of the Assumptions (Micronumerosity, Multicollinearity, Nature of Independent Variables and Autocorrelation)

Based on the time series data of sesame production, sown acreage, and harvested, irrigation, loan, quality seeds and pesticides during the period of 1989 to 2018, the proposed multiple regression model of sesame production in Myanmar on the sown acreage, and yield per harvested acre, irrigation, loan, quality seeds and pesticides are fitted. The regression equation results are shown the following Table (4.15).

Table (4.15)

Results for Original Data of Sesame Production

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF	Cov(ui,Xij)
Constant	-580.23	299.987	-1.934	0.065*			
SN	0.098	0.143	0.687	0.499	0.114	8.758	0.000
HAR	0.234	0.076	3.105	0.005***	0.123	8.161	-0.0002
IRRI	-0.005	0.503	-0.011	0.991	0.585	1.709	0.0000
LA	0.008	0.004	2.028	0.054*	0.142	7.030	-0.009
QUALI	0.001	0.018	0.067	0.947	0.768	1.301	-0.0011
PESTI	0.000	0.000	-1.264	0.219	0.170	5.885	0.127
R-squared		0.877	E (u_i)		0.000		
Adjusted R-squared		0.846	F-statistic		27.453		
Std. Error of the Estimate		102.99	Prob (F-statistic)		0.000***		
Durbin-Watson		0.900					

**** denote significant at 1% level and * denote significant at 10% level.

Source: Appendix (A4)

The estimated regression equation for sesame production is

PROD = -580.23+0.098SN+0.234HAR-0.005IRRI+0.008LA+0.001QUALI + 0.000PESTI (4.7)

From the estimated regression equation for sesame production, it is found that sesame production is positively related to sown, harvested, loan, quality seeds and pesticides and is negatively related to irrigation. It is found that holding the harvested, irrigation, loan, quality seeds and pesticides are constant, a 1 thousand acre increase in sown led on the average to about 0.098 ton increase in production and holding the sown, irrigation, loan, quality seeds and pesticides are constant, a 1 thousand acre increase in harvested led on the average to about 0.234 thousand acreage increase in production. Holding the sown, harvested, loan, quality seeds and pesticides are

constant, a 1 thousand acre increase in irrigation led on the average to about 0.005 thousand acreage decrease in production. Holding the sown, harvested, irrigation, quality seeds and pesticides are constant, a 1 kyat millions increase in loan led on the average to about 0.008 thousand acreage increase in production. Holding the sown, harvested, irrigation, loan and pesticides are constant, a 1 basket increase in quality seeds led on the average to about 0.001 thousand acreages increase in production and since the coefficient of pesticides is 0, any gallon increase in pesticides led about 580.23 thousand acreages decrease in production holding the other variables are constant.

The harvested (3.105) is statistically significant at 1% level and loan (2.028) is also statistically significant at 10% level but other variables (sown, irrigation, quality seeds pesticides) are not significant. The F-test of multiple linear regression (27.453) is highly significant at 1% level. Adjusted R-square is0.846 and the R squared is 0.877 which means that 87.7% of variation in sesame production is explained sown, harvested, irrigation, loan, quality seeds and pesticides and the remaining percentage 12.3% due to other factors that are not included in the model. The standard error of estimate (102.99) is very large.

The Durbin-Watson d is 0.499 which is less than lower limit (1.026) at the number of observation (21) and parameter (3). The mean of error is zero and the covariance values of error term and independent variables are 0.000, -0.0002, 0.0000, -0.009, -0.0011 and 0.127 which can be assumed to zero. Moreover, since VIF values of sown, harvested, irrigation, loan, quality seeds and pesticides are 8.758, 8.161, 1.709, 7.030, 1.301 and 5.885 which is greater than 5 and total VIF is not greater than 10 and tolerance values are 0.114, 0.123, 0.585, 0.142, 0.768 and 0.170 which are closely to 0.

4.5.1 Detection Micronumerosity Assumption

One of the assumptions of classical linear regression analysis is that there is no micronumerosity. The problem of micronumerosity simply means small sample size. In this problem, a regression model with Ordinary Least Squares (OLS) method cannot be estimated when sesame data set for 5 years (2014-2017) with six parameters are estimated. See Appendix (B1-B4). Sesame data set for 6 years (2013-2017) with six parameters are estimated. See Appendix (B5-B10).

4.5.2 Remedy Micronumerosity Assumption

If the micronumerosity assumption is violated, a regression model with Ordinary Least Squares (OLS) method cannot be estimated and a precise (unbiased) estimate with relatively least standard errors cannot be obtained. This problem is remedied by adding the observations as shown the following Table (4.16).

Table (4.16)

Results for Sesame Production Adding the Observations

Variable	Coefficient	Std. Error	t-Statistic	Sig.
Constant	-580.23	299.987	-1.934	0.065*
SN	0.098	0.143	0.687	0.499
HAR	0.234	0.076	3.105	0.005***
IRRI	-0.005	0.503	-0.011	0.991
LA	0.008	0.004	2.028	0.054*
QUALI	0.001	0.018	0.067	0.947
PESTI	0.000	0.000	-1.264	0.219
R-squared	0.877	**F-statistic**		27.453
Adjusted R-squared	0.846	**Prob (F-statistic)**		0.000***

**** denote significant at 1% level and * denote significant at 10% level.

Source: Appendix (A4)

In Table (4.16), although the F-statistic (27.453) is significant at 1% level of significance, most of the variables (sown, irrigation, quality seeds and pesticides) are insignificant at 10% level of significance. Hence, other assumptions are needed to detect and remedy.

4.5.3 Detection of Multicollinearity Assumption

One of the assumptions of classical linear regression analysis is that there is no exact multicollinearity between the independent variables. To test the assumption of multicollinearity, Variance Inflation Factor (VIF) and tolerance (TOL) can be used, especially in regression analyses. The sesame production data set from the remedy

micronumerosity assumption are used to detect the multicollinearity and the results are shown in following Table (4.17).

Table (4.17)

Results for Original Data by TOL and VIF

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	-580.23	299.987	-1.934	0.065*		
SN	0.098	0.143	0.687	0.499	0.114	8.758
HAR	0.234	0.076	3.105	0.005***	0.123	8.161
IRRI	-0.005	0.503	-0.011	0.991	0.585	1.709
LA	0.008	0.004	2.028	0.054*	0.142	7.030
QUALI	0.001	0.018	0.067	0.947	0.768	1.301
PESTI	0.000	0.000	-1.264	0.219	0.170	5.885
R-squared		0.877	F-statistic		27.453	
Adjusted R-squared		0.846	Prob (F-statistic)		0.000***	

*** denote significant at 1% level and * denote significant at 10% level.

Source: SPSS Output

From Table (4.17), the following are noticeable. The F-statistic is highly significant (P-value 0.000), implying the variables chosen and value of valid independent variables and most of the regression coefficients are insignificant at 5% level of significance. The value of R^2 is 0.877. Variance inflation factor (VIF) and tolerance (TOL) are the measure of multicollinearity. The VIF values of each variable are 8.758, 8.161, 1.709, 7.030, 1.301 and 5.885 and the VIF value of sown, harvested, quality seeds and pesticides are greater than 5. The total VIF value is 32.844 that mean greater than 10. TOL values are 0.114, 0.123, 0.585, 0.142, 0.768 and 0.170 and TOL values are closely to zero. Therefore, these VIF and TOL values are not acceptable.

If in Pearson Correlation Matrix, any of the dependent variables included have a high correlation with any other dependent variable or the value of correlation is

significant, one of these independent variables are removed. The values of correlation among the independent variables are shown in below Table (4.18).

Table (4.18)

Correlation Matrix

		SOWN	HAR	IRRI	LOAN	QUALI	PESTI
SN	Pearson Correlation	1	.911**	.258	.521**	-.396*	.302
	Sig. (2-tailed)		.000	.168	.003	.030	.105
HAR	Pearson Correlation	.911**	1	.076	.537**	-.323	.335
	Sig. (2-tailed)	.000		.691	.002	.082	.070
IRRI	Pearson Correlation	.258	.076	1	.406*	-.081	.438*
	Sig. (2-tailed)	.168	.691		.026	.669	.015
LOAN	Pearson Correlation	.521**	.537**	.406*	1	.019	.885**
	Sig. (2-tailed)	.003	.002	.026		.921	.000
QUALI	Pearson Correlation	-.396*	-.323	-.081	.019	1	.105
	Sig. (2-tailed)	.030	.082	.669	.921		.581
PESTI	Pearson Correlation	.302	.335	.438*	.885**	.105	1
	Sig. (2-tailed)	.105	.070	.015	.000	.581	

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

*. Correlation is significant at the 0.05 level (2-tailed).

Source: Appendix (A4)

In Table (4.18), the correlation between sown and harvested, sown and loan, sown and quality seeds are significant and the correlation between loan and harvested, loan and irrigation, loan and pesticides are significant. The correlation between

pesticides and irrigation is also significant. Therefore, sown, loan and pesticides are highly correlated with other independent variables.

4.5.4 Remedy of Multicollinearity Assumption

In this study, sown, loan and pesticides are needed to remove to remedy the violation of multicollinearity assumption. When sown, loan and pesticides are removed, multicollinearity assumption is again detected by the remained independent variables. The results are given in Table (4.19) as follows.

Table (4.19)

Results for Removing the Variables by TOL and VIF

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	-587.109	129.062	-4.549	0.000***		
HAR	0.333	0.029	11.286	0.000***	0.893	1.119
IRRI	0.420	0.407	1.032	0.311	0.991	1.009
QUALI	0.005	0.018	0.304	0.764	0.893	1.120
R-squared		0.846	F-statistic		47.699	
Adjusted R-squared		0.829	Prob (F-statistic)		0.000***	

*** denote significant at 1% level.

Source: Appendix (A4)

According to Table (4.19), the VIF values of each variable are 1.119, 1.009 and 1.120 and the total VIF value is 3.248 which is less than 10 and TOL values are 0.893, 0.991 and 0.893 which are closely to one. Therefore, this study gives an acceptable level of collinearity and assumes that there is no mutlicollinearity.

4.5.5 Detection of the Nature of Independent Variables Assumption

One of the assumptions of classical linear regression analysis is the nature of independent variables. According to the result of VIF from Table (4.19), there is no evidence of significant problem in multicollinearity. Next step is to check whether violate the nature of independent variables assumption. In the nature of independent variables, there has two types of outliers and they are outliers of the independent

variables and influence observations. The outliers of the independent variables are observed with each box plot as follows Figure (4.7).

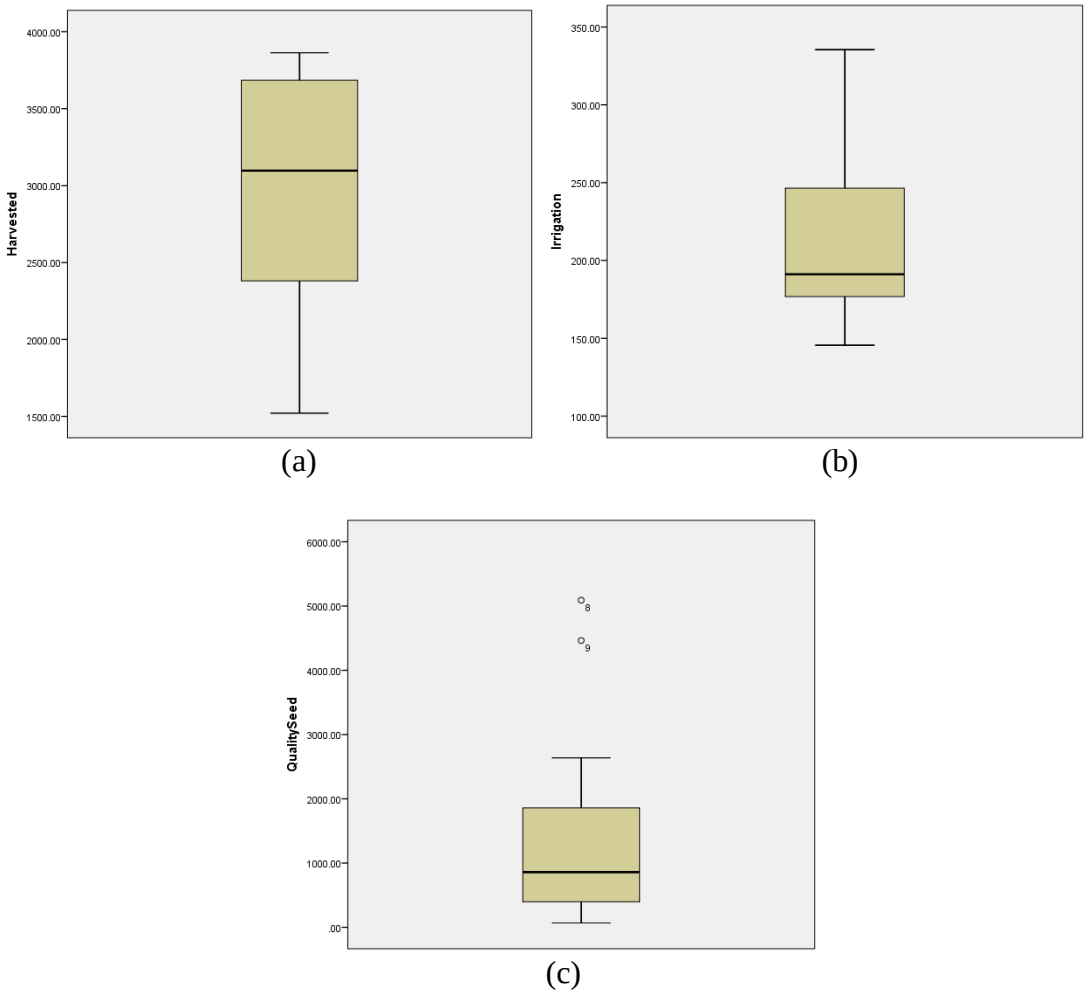


Figure (4.7) Box Plots of Independent Variables for Original Data

Source: Appendix (A4)

According to Figure (4.5), there have no outliers in harvested and irrigation but have the outliers in quality seeds. The quality seeds has the outliers in the case number 8 and 9. The influence observations are recognized by using Cook's distance as the following table.

Table (4.20)

Cook's Distance for Original Data

Minimum	Maximum	Mean	Standard Deviation
0	0.832	0.053	0.151

Source: Appendix (A4)

In Table (4.20), the maximum Cook's Distance is 0.832 and the standard level based on the 50^{th} percentile of the F distribution with 4 and 30 degree of freedom is 0.809. The maximum value of Cook's Distance is greater than the 50^{th} percentile of the F distribution. Therefore, there has influence observation in these data. To show the influence observation, a scree plot of the Cook's distance for original data is shown in Figure (4.8).

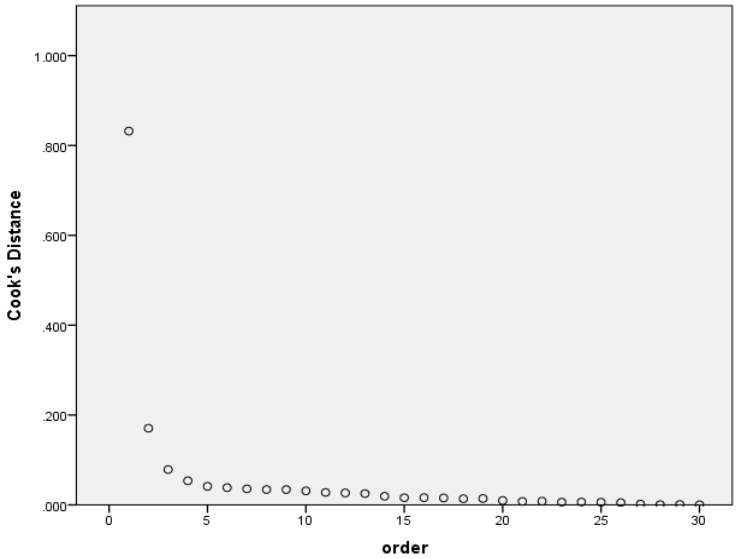


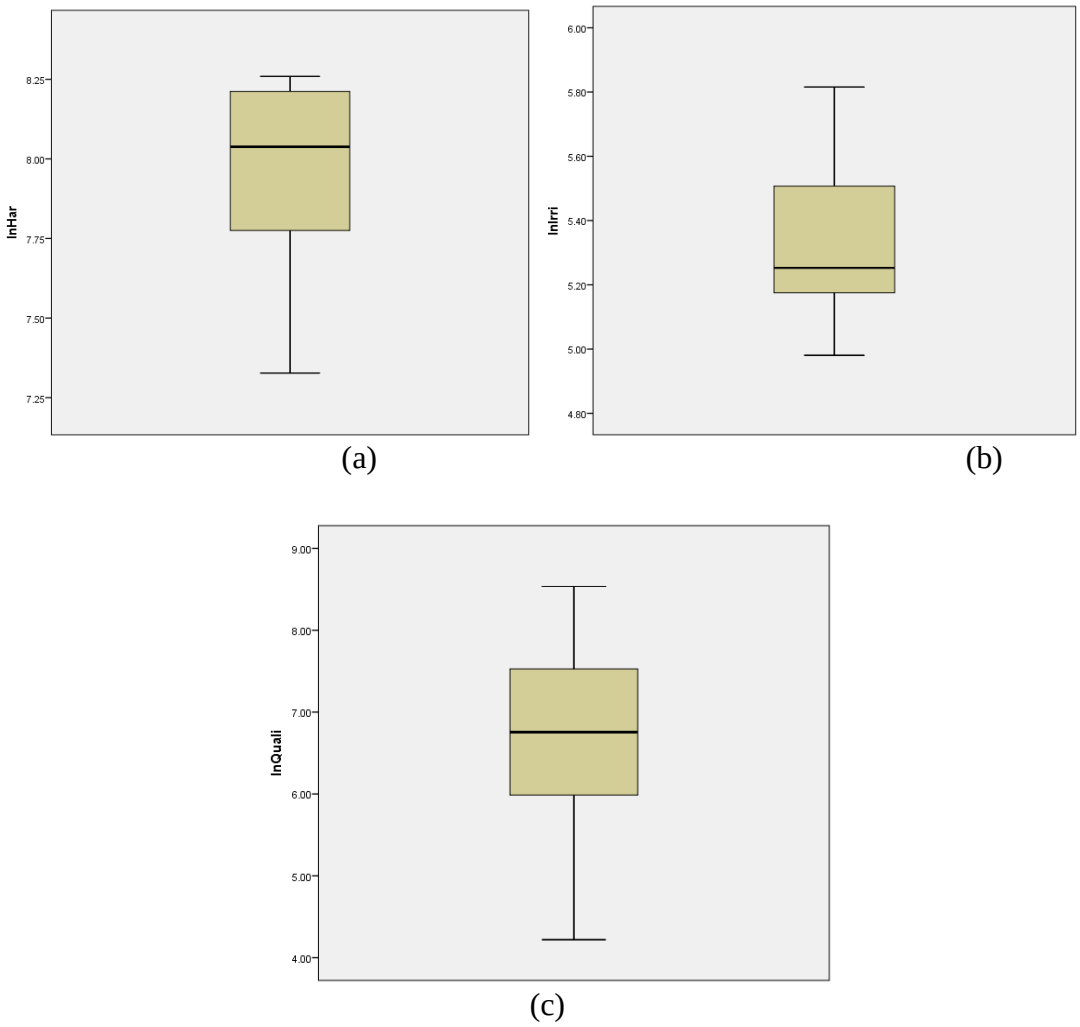
Figure (4.8) Scree plot of Cook's Distance for Original Data

Source: Appendix (A4)

According to Figure (4.8), the scree plot of the Cook's distance suggests that one observation which case number 11 is comparatively more influential than most others and could be subject to sensitivity analysis.

4.5.6 Remedy of the Nature of Independent Variables

The quality seeds has the outliers in the case number 8 and 9 and the case number 11 is the influence observation. If the cross sectional data are used, these case number from the data cut off to detect the outlier. But since the panel data or time series data are used in this study, it is not possible to cut because of the gap of the time lag. In panel data or time series data, if the data has few outliers, these outlier can be neglected and if more, the data are needed to transform. Hence, the variables are transformed by taking the double-log in this subsection. The outliers of the logarithm of independent variables are observed with each box plot as follows in Figure (4.8).



(a) (b)

(c)

Figure (4.9) Box Plots of Independent Variables for Transformation Data

Source: Appendix (A4)

According to Figure (4.9), there have no outliers in the logarithm of independent variables. The influence observations are recognized by using Cook's distance as the following table.

Table (4.21)

Cook's Distance for Transformation Data

Minimum	Maximum	Mean	Standard Deviation
0	1.631	.083	0.297

Source: Appendix (A4)

In Table (4.23), the maximum Cook's Distance is 1.631 and the standard level based on the 50^{th} percentile of the F distribution with 4 and 30 degree of freedom is 0.809. The maximum value of Cook's Distance is greater than the 50^{th} percentile of the F distribution. Therefore, there has influence observation in these transformation

data. To show the influence observation, a scree plot of the Cook's distance for transformation data is shown in Figure (4.10).

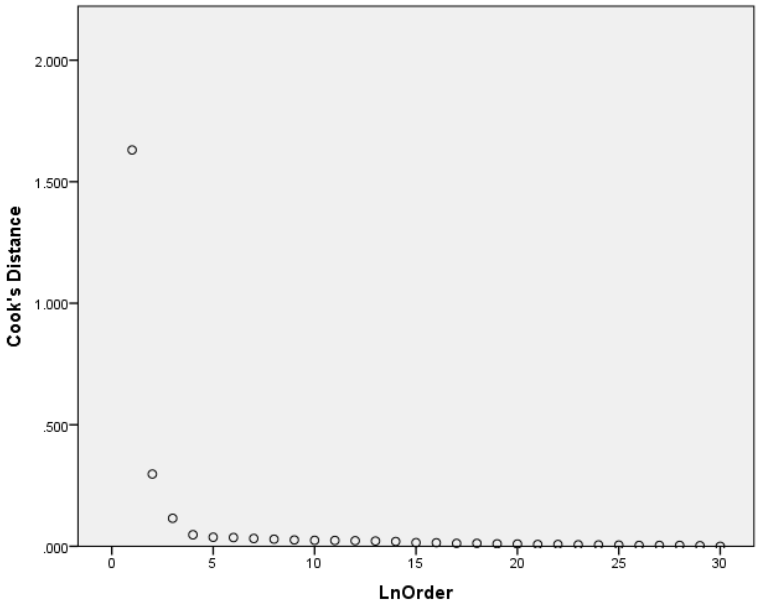


Figure (4.10) Scree plot of Cook's distance for Transformation Data
Source: Appendix (A4)

According to Figure (4.10), the scree plot of the Cook's distance suggests that one observation which case number 11 is comparatively more influential than most others and could be subject to sensitivity analysis. There is no outliers in the independent variables and only one influence observation remain by taking double log. Hence, one influence observation is neglected.

4.5.7 Detection of Autocorrelation between the Disturbances Assumption

One of the assumptions of classical linear regression analysis is that there is no autocorrelation or serial correlation between the disturbances. The Durbin-Watson result is shown to detect the autocorrelation between the disturbances in below Table (4.22) and Figure (4.11).

Table (4.22)

Durbin-Watson Result for Double Log Equation

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	-10.467	1.920	-5.451	0.000***		
Ln (HAR)	1.906	0.181	10.550	0.000***	0.949	1.053
Ln (IRRI)	0.259	0.216	1.198	0.242	0.980	1.021
Ln (QUALI)	-0.005	0.047	-0.101	0.920	0.935	1.070
R-squared		0.823	F-statistic		40.288	
Adjusted R-squared		0.803	Prob (F-statistic)		0.000***	
Durbin-Watson		0.872				

*** denote significant at 1% level

Source: Appendix (A4)

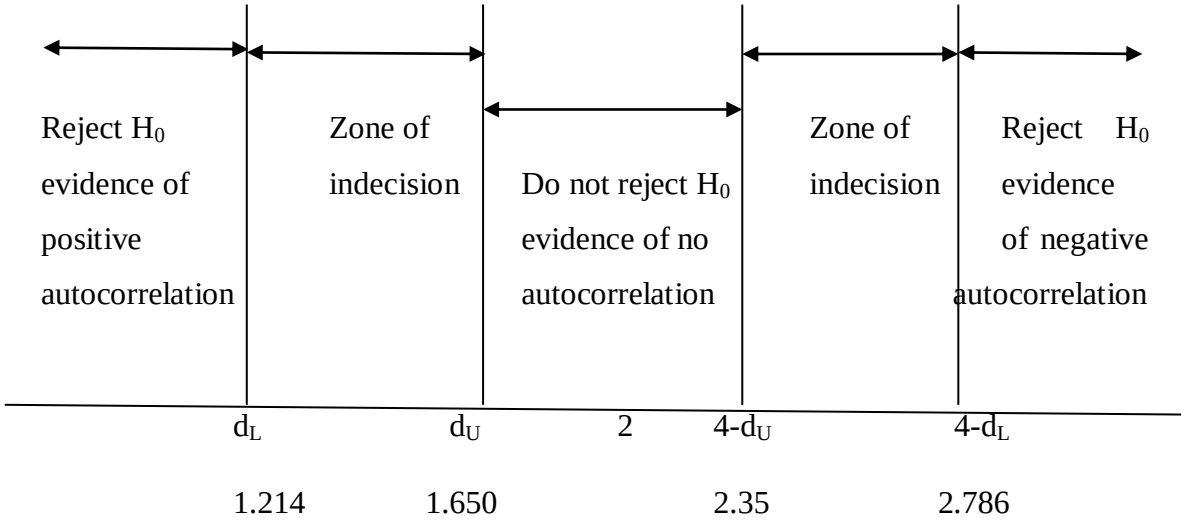


Figure (4.11): Durbin-Watson Statistic for Double Log Equation

Source: Appendix (A4)

In Table (4.22), the computed Durbin-Watson d statistic is 0.872. Since the number of observations is 30 and parameters is 3, there is d_L=1.214 and d_U =1.65 in these data. Since the computed d value is less than 1.214, there is evidence of positive first-order serial correlation.

4.5.8 Remedy of Autocorrelation between the Disturbances

Since ρ value can be estimated based on Durbin-Watson d statistic, the generalized difference equation is used to remedy the autocorrelation between the disturbances. The ρ value based on Durbin-Watson d statistic is 0.564. The analysis results for generalized difference equation are shown in Table (4.23) as follow.

Table (4.23)

Durbin-Watson Result for Generalized Difference Equation

Variable	Coefficient	Std. Error	t-Statistic	Sig.	TOL	VIF
Constant	0.038	0.024	1.583	0.125		
New HAR	0.647	0.091	7.075	0.000***	0.791	1.264
New IRRI	0.037	0.089	0.411	0.685	0.871	1.148
New QUALI	0.004	0.023	0.192	0.849	0.898	1.113
R-squared		0.720	F-statistic		22.293	
Adjusted R-squared		0.688	Prob (F-statistic)		0.000***	
Std. Error of Estimate		0.126				
Durbin-Watson		2.605				

Source: Appendix (A4)

New Production = Ln (PROD) – ρ Δ(Ln (PROD))

New HAR = Ln (HAR) – ρ Δ (Ln (HAR))

New IRRI = Ln (IRRI) – ρ Δ (Ln (IRRI))

New QUALI = Ln (QUALI) - ρ Δ (Ln (QUALI))

Δ = first order differences

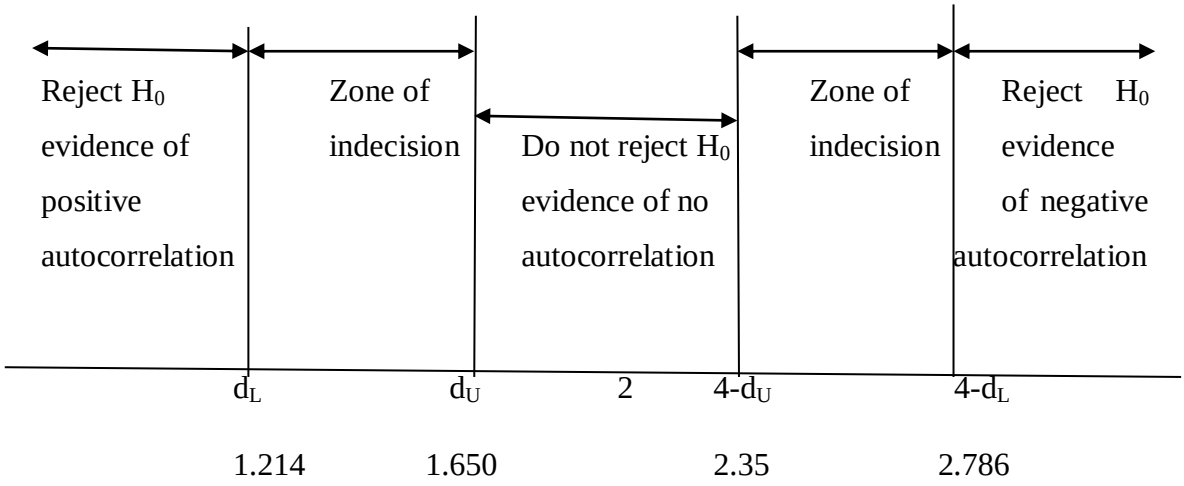


Figure (4.12): Durbin-Watson Statistic for Generalized Difference Equation

Source: Appendix (A4)

In Table (4.23), the computed Durbin-Watson d statistic is 2.605. Hence, since the computed d value lies between the $4 - d_U$ and $4 – d_L$, there is inconclusive evidence regarding the presence of positive first-order serial correlation.

The final regression model for sesame production is

New PROD = 0.038 + 0.647New HAR -0.037New IRRI + 0.004New LA (4.8)

CHAPTER V

CONCLUSION

Multiple regression examines the relationship between a single outcome measure and several predictor or independent variables. The correct use of the multiple regression model requires that several critical assumptions can be satisfied in order to apply the model and establish validity. Inferences and generalizations about the theory are only valid if the assumptions in an analysis have been tested and fulfilled.

Before a complete regression analysis could be performed, the assumptions concerning the original data must be tested. Ignoring the regression assumptions contribute to wrong validity estimates. When the assumptions were not met, the results could be obtained in Type I or Type II errors, or over- or under-estimation of significance of effect size. Hence, meaningful data analysis were necessary to test of the assumptions and the consequences of violations.

In this study, the maize data, wheat data, rice data and sesame data of Myanmar were applied. These various data series were used to diagnosis the individual assumption and to remedy the violation assumption.

Firstly, the maize data of Myanmar was used to examine the linearity assumption. When linearity assumption was exposed by using scatter plots and residual plot, assume that these data violated the linearity assumption. The scatter plots for double log transformation variables was closely to straight line and residual plot for double log transformation was no pattern. Hence, the double log transforming variables satisfied the linearity assumption.

Secondly, wheat data of Myanmar was used to check the normality assumption of the disturbance terms. The normality of the disturbance terms was detected by using Kolmogorov-Smirnov test and boxplot. According to Kolmogorov-Smirnov test, there was significant at 5% level, the disturbance terms of these data did not satisfy the normality assumption. Since box plot was skewed to the above, the variables were transformed by taking reflect and square root. The Kolmogrov-Smirnov test for redefining variables were satisfied the normality assumption. Therefore, reflect and square root transforming variables were satisfied the violation of the normality assumption.

Thirdly, rice data of Myanmar was used to analyses the homoscedasticity assumption. When the homoscedasticity assumption was tested with White's test, the disturbance terms have equal variances, that is, heteroscedasticity exits. The redefining method was used to remedy the violation of homoscedasticity assumption. By redefining method, the variables were redefined to the double log and when the redefining variables were diagnosis by using White's test, heteroscedasticity did not exist. Hence, the redefining method was successfully solved the violation in constant variance in original data.

Fourthly, sesame data of Myanmar was used to explore the micronumerosity, multicollinearity, the nature of independent variables, and autocorrelation between the disturbances assumptions. In micronumerosity problem, a regression model with OLS method could not be estimated when sesame data set from 2014 to 2018 with six parameters (SOWN, HAR, IRRI, LOAN, QUALI AND PESTI) were used to estimate. Besides that when both observations and parameters were the same, a regression model with OLS method could not be estimated. This problem was remedied by adding the observations. When the observations were added from 1989 to 2018, the variables were even insignificant at 10% level. Hence, other assumptions are needed to detect.

Multicollinearity assumption was detected by using these adding data. The VIF and TOL was one of the detecting ways in multicollinearity assumption. The VIF values of SOWN, HAR, LOAN and PESTI were greater than 5 and total VIF was greater than 10. TOL values were closely to zero. Since these suffered the multicollinearity, Pearson correlation matrix were used to know strong correlation between independent variables. According to Pearson correlation, SOWN, LOAN and PESTI were strongly correlated with HAR, IRRI and QUALI. So, when SOWN, LOAN and PESTI were removed, the VIF and TOL values of HAR, IRRI and QUALI gave an acceptable level of multicollinearity. Because of that these data set could be assumed no multicollinearity.

After that, the remaining independent variables were used to be diagnosis the nature of independent variables assumption. In the nature of independent variables, outliers of the independent variables and influence observations were observed with box plot and Cook's distance. By using box plots, quality seeds had two outliers and the maximum value of Cook's Distance was greater than the 50^{th} percentile of F distribution. There had influence observation in these data. To show the influence

observation, a scree plot of Cook's distance was used. By scree plot, one observation was more influential than most others. Although the outliers were cut off in cross section data, these were not possible to cut off because of the gap of the time lag. When the data were transformed by taking the double log, the independent variables had no outlier but there had one influence observation. These influence observation were neglected.

And then, the double log transforming data were used for autocorrelation between the disturbances. When the autocorrelation of the disturbances was tested by Durbin-Watson test, the transforming data was positive first order serial correlation. Since ρ value could be estimated based on the Durbin-Watson d statistic, the generalized difference equation was used to remedy the autocorrelation. By using generalized difference equation, there was inconclusive evidence regarding the presence of negative first order serial correlation.

In this thesis, the unnecessary variables were added in multicollinearity problem using sesame data. The presence of unnecessary variables in the regression model was caused the specification error of inclusion of unnecessary variables and autocorrelation was caused wrong function form. The violation of normality of the disturbances was one of the specification errors.

To sum up, when the classical multiple linear regression analysis were used, the assumptions were needed to detect and when the assumptions were violated, the various remedial ways were used to satisfy the assumptions. If the remedial ways were used for the violation of the assumptions, the data type must be noticed. In cross sectional data, there were many ways to make the reducing but it was not possible in time series data or panel data.

REFERENCES

Andreea Halunga, C. D. (2011). Heteroscedasticity Robust Breusch-Pagan Test for Contemporaneous Correlation in Dynamic Panel Data Models.

Antonakis, J. &. (2011). Looking for Validity or Testing It? The Perils of Stepwise Regression, Extreme-Score Analysis,Heteroscedasticity, and Measurement Error. *Pesonality and Individual Differences*, pp. 409-415.

Barbakh, M. M. (2012, June). A Study on the violation of Homoskedasticity Assumption in Linear Regressin Models. (A.-A. U.-G. Statistics, Ed.)

Baser, O. (2007). Modeling Transformed Health Care Cost with Unknown Heteroscedasticity.

Damodar N. Gujarati, D. C. (n.d.). *Basic Economics* (Fifth Edition ed.).

Osborne, J.,& Wasters,E.(2002). *Four Assumptions of Multiple Regression that Researchers should always test.* Partical Assessment, Research & Evaluation.

Guggenberger, D. W. (2011). Conditional Heteroskedasticity -Robust Confidence Interval for the Autoregressive Parameter.

Hausman, N. W. (2009). Intrumental Variable Estimation with Heteroscedasticity and Many Instruments.

Hayes, A. F. (2009). Using Heteroskedasticity-Consistent Standard Error Estimators in OLS Regression: An Introduction and Software.

Hayes, A. F. (2009). Using Heteroskedasticity-Consistent Standard Error Estimators in OLS Regression: An Introduction and Software.

Hoyt, W. L. (2006). Analysis and Interpretation of Findings Using Multiple Regression Techniques. *Rehabilitation Counseling Bulletin*, pp. 223-233.

Jaccard, J. G.-R. (2006). Multiple Regression Analyses in Clinical Child and Adolescent Psychology. *Journal of Counseling Bulletin*, 456-479.

Keith, T. (2006). Multiple Regression and Beyond. *PEARSON Allyn & Bacon*.

Mario Francuso, J. M. (2007). Two tests for Heteroscedasticity in Nonparametric Regressiom.

Mason, C. &. (1991). Collinearity, Power and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research*, 268-280.

Neale, M. H. (1994). *Multiple Regression with Data Colleect from Relatves: Testing Assumptions of the Model.* Multivarite Behavioral Research.

Oyeyemi, G. B. (n.d.). University of Ilorin, Department of Statistics, Ilorin.

P.Marshall, T. S. (1995). *Testing the distributional assumptions of least squares linear regression.* University of British Columbia 2357 Main Mall, Forest Resources Management.

Pasha, M. A. (2000). Adaptive Estimation of Heteroscedastic Linear Regression Models Using Heteroscedasticity Consistent Covariance Matrix .

Pinkse, J. (2006). Heteroskedasticity Correlation and Dimension Reduction.

Poole, M. &. (1971). The Assumption of the Linear Regression Model. *Transactions of the Institute of British Geographers*, 145-158.

Saez, G. C. (2000). *Collinearity, Heteroscedasticity and Outlier Diagnostics in Regression. Do They Always Offer What They Claim?* University of Giroma, Department of Economics, Girona.

Schreiber-Gregory, D. (2018). Logistic and Linear Regression Assumptions: Violation Recognition and control. *Henry M Jackson Foundation*.

Sevier, F. (1957). Testing Assumptions underlying Multiple Regression. *The Journal of Experimental Education*, 323-330.

Shieh, G. (2010). *On the Misconception of Multicollinearity in detecting of Moderating Effects: Multicollinearity is not Always Detrimental.* Multivariate Behavioral Research.

Terasvirta, T. (2011). Nonlinear Models for Autoregressive Conditional Hetroscedasticity.

Webster, A. L. (1995). *Applied Statistics for Business and Economics* (Second Edition ed.).

Zheng, X. (2009). Testing Heteroscedasticity in Nonlinear and Nonparametric Regression.

APPENDICES

APPENDIX A

Appendix (A1)

Year	Production	Pesticides	Quality Seeds	Weir
1998	303.4	1452	19732	10
1999	297.9	440	19863	9
2000	343.6	111	19650	7
2001	358.9	521	60641	18
2002	524	1467	6974	17
2003	593.4	3872	4273	23
2004	692.9	1293	9167	28
2005	771.1	555	10968	27
2006	903.5	607	10282	30
2007	1015.8	744	1668	25
2008	1128.1	3607	1687	32
2009	1184.7	3651	1390	32
2010	1225.7	3685	1485	33
2011	1354.4	25774	3220	35
2012	1461.5	26484	3910	28
2013	1502	86305	2115	29
2014	1601	4838	1192	31
2015	1693	18314	2323	30
2016	1748	175803	2890	31
2017	1831	379918	1350	31
2018	1909	910543	1995	33

Source: Statistical Year Book (CSO), Agricultural Statistics

Appendix (A2)

Year	Production	Pesticides	Substitution Sown
1998	90.7	242	4863
1999	92	26	1568
2000	115.3	22	5408
2001	92.1	318	6206
2002	94.4	222	5035
2003	105.7	111	6865
2004	122.4	135	8522
2005	150	222	9181
2006	156.2	78	6836
2007	140.2	44	5533
2008	155.3	361	3762
2009	170.4	363	7787
2010	179.2	366	8372
2011	181	369	8301
2012	169.8	344	5701
2013	177.6	863	7741
2014	182.9	366	9117
2015	182.4	221	12616
2016	179	159	13814
2017	166	364	665
2018	123	894	11331

Source: Statistical Year Book (CSO), Agricultural Statistics

Appendix (A3)

Year	Production	Sown	Yield	Year	Production	Sown	Yield
1966	8055.5	12,390	32.22	**2002**	21,914	15,940	66
1967	6636.76	12,328	28.5	**2003**	21,804	16,032	66
1968	7769.66	12,193	32.02	**2004**	23,135	16,168	68
1969	8022.12	12,402	32.67	**2005**	24,751	16,946	70
1970	7984.56	12,243	33.15	**2006**	27,682	18,259	71
1971	8161.9	12,294	32.92	**2007**	30,922	20,076	73
1972	8174.42	12,299	33.28	**2008**	30954	19,989	76.14
1973	7356.56	12,014	31.51	**2009**	32,058	20,001	78.21
1974	8602.13	12,575	34.19	**2010**	32164	19933	78.91
1975	8583.35	12,793	34.09	**2011**	32064	19885	78.6
1976	9207.18	12,858	35.91	**2012**	28221	18762	74.36
1977	9317.76	12,547	36.8	**2013**	26952	17893	74.55
1978	9461.72	12,690	37.73	**2014**	27545	17999	75.63
1979	10527.9	12,957	40.75	**2015**	26423	17722	76.45
1980	10446.5	12,420	45.62	**2016**	26210	17821	77.08
1981	13317.3	12,668	53.8	**2017**	25673	17695	75
1982	14145.6	12,610	57.06	**2018**	25624	17930	75
1983	14,146	12,064	61.1				
1984	14,372	11,938	59.84				
1985	14,253	12,151	60.09				
1986	14,464	12,076	60.09				
1987	14,126	12,193	58.72				
1988	11,548	11,530	59.8				
1989	13,164	11,807	56				
1990	13,803	12,057	57				
1991	13,968	12,220	57				
1992	13,201	11,935	56				
1993	14,837	12,684	57				
1994	16,759	14,021	59				
1995	18,194	14,643	61				
1996	17,952	15,166	58				
1997	17,675	14,518	59				
1998	16,654	14,294	59				
1999	17,077	14,230	61				
2000	20,125	15,528	63				
2001	21,323	15,713	66				

Source: Statistical Year Book (CSO), Agricultural Statistics

Appendix (A4)

Year	Production	Sown	Harvested	Irrigation	Loan	Quality Seeds	Pesticides
1989	142.7	2994	1592	198.72	65.13	2153	326
1990	203.5	3158	2285	196.05	59.72	555	1901
1991	212.4	3271	2454	174.64	60.54	1694	1460
1992	167.9	3184	1984	186.69	61.52	1849	680
1993	233.4	3379	2451	184.98	98.76	2407	889
1994	219.8	3212	2338	145.56	93.04	2638	865
1995	299.3	3288	2797	175.97	96.27	1256	1978
1996	298.8	3153	2234	264.32	850.89	5091	3682
1997	340.3	2830	2746	160.75	851.11	4464	2933
1998	258.7	2557	1789	147.76	315.55	1858	2734
1999	260	2963	1521	285.57	437.71	228	2332
2000	253.2	3352	2381	293.94	523.18	750	158
2001	375.8	3517	3064	232.19	524.83	660	240
2002	339	3416	2865	236.94	590.64	68	3111
2003	405.9	3501	3130	184.98	569.72	306	3386
2004	436.2	3619	3281	208.18	963.22	230	577
2005	473.9	3696	3306	246.46	1241.14	209	333
2006	438.5	3306	2934	188.24	1562.13	154	289
2007	680	3565	3378	186.08	1945.74	769	111
2008	768	3725	3536	194.04	3360.32	398	7581
2009	840	3880	3685	176.83	3581.21	252	7627
2010	854	4038	3863	184.48	5665.1	899	7678
2011	787.4	3918	3754	178.61	10059.67	484	29796
2012	832.1	3941	3785	160.78	11400.85	2150	35324
2013	794.6	3838	3688	156.73	11906.29	443	86324
2014	817.1	4007	3767	209.92	24962.32	825	18654
2015	801.6	3906	3600	281.61	25205.4	893	86218
2016	828	4052	3732	285.66	22639.9	1186	243707
2017	813	4042	3694	335.49	23320.86	1644	447822
2018	764	3685	3652	260.57	58018.6	2000	928447

Source: Statistical Year Book (CSO), Agricultural Statistics

APPENDIX B

Appendix (B1)

Descriptive Statistics

	Mean	Std. Deviation	N
Production	804.7400	24.65579	5
Sown	3938.4000	152.97483	5
Harvested	3689.0000	65.66582	5
Irrigation	274.6500	45.44058	5
Loan	30829.4160	15237.70557	5
QualitySeed	1309.6000	503.02714	5
Pesticides	344969.6000	365620.31808	5

Appendix (B2)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2431.632	4	607.908	.	.[b]
	Residual	.000	0	.		
	Total	2431.632	4			

a. Dependent Variable: Production

b. Predictors: (Constant), Pesticides, Irrigation, Harvested, QualitySeed

Appendix (B3)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-378.043	.000		.	.
	Harvested	.310	.000	.825	.	.
	Irrigation	.524	.000	.965	.	.
	QualitySeed	-.100	.000	-2.042	.	.
	Pesticides	8.088E-005	.000	1.199	.	.

a. Dependent Variable: Production

Appendix (B4)

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Sown	.[b]	.	.	.	.000
	Loan	.[b]	.	.	.	.000

a. Dependent Variable: Production

b. Predictors in the Model: (Constant), Pesticides, Irrigation, Harvested, QualitySeed

Appendix (B5)

Descriptive Statistics

	Mean	Std. Deviation	N
Production	803.0500	22.43798	6
Sown	3921.6667	142.83230	6
Harvested	3688.8333	58.73471	6
Irrigation	254.9967	63.00316	6
Loan	27675.5617	15666.23502	6
QualitySeed	1165.1667	572.35912	6
Pesticides	301862.0000	343645.40541	6

Appendix (B6)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	.	.

a. Predictors: (Constant), Pesticides, Harvested, Irrigation, Loan, QualitySeed

Appendix (B7)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2517.315	5	503.463	.	.[b]
	Residual	.000	0	.		
	Total	2517.315	5			

a. Dependent Variable: Production

b. Predictors: (Constant), Pesticides, Harvested, Irrigation, Loan, QualitySeed

Appendix (B8)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-500.086	.000		.	.
	Harvested	.335	.000	.876	.	.
	Irrigation	.597	.000	1.676	.	.
	Loan	.000	.000	.308	.	.
	QualitySeed	-.098	.000	-2.493	.	.
	Pesticides	6.086E-005	.000	.932	.	.

a. Dependent Variable: Production

Appendix (B9)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-500.086	.000		.	.
	Harvested	.335	.000	.876	.	.
	Irrigation	.597	.000	1.676	.	.
	Loan	.000	.000	.308	.	.
	QualitySeed	-.098	.000	-2.493	.	.
	Pesticides	6.086E-005	.000	.932	.	.

a. Dependent Variable: Production

Appendix (B10)

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Sown	.[b]	.	.	.	.000

a. Dependent Variable: Production

b. Predictors in the Model: (Constant), Pesticides, Harvested, Irrigation, Loan, QualitySeed

Appendix (B11)

Descriptive Statistics

	Mean	Std. Deviation	N
Production	1068.7095	542.16870	21
Pesticide	78570.6667	210229.42327	21
Weir	25.6667	8.46955	21
QualitySeed	8894.0476	13461.00095	21

Appendix (B12)

Descriptive Statistics

	Mean	Std. Deviation	N
Production	144.0762	34.65546	21
Pesticides	290.0000	230.97662	21
SubstitutionSown	7105.9048	3227.93624	21

Appendix (B13)

Descriptive Statistics

	Mean	Std. Deviation	N
Production	17311.4701	7899.67923	53
Sown	14567.9245	2780.28885	53
Yield	56.8828	15.77927	53

Appendix (B14)

Descriptive Statistics

	Mean	Std. Deviation	N
Production	497.9700	262.04522	30
Sown	3499.7667	396.57740	30
Harvested	2976.2000	723.63076	30
Irrigation	210.7580	49.71418	30
Loan	7034.3787	12645.54763	30
QualitySeed	1283.7667	1216.95858	30
Pesticides	64238.7667	187153.36100	30

Appendix (B15)

Descriptives

			Statistic	Std. Error
Standardized Residual	Mean		0E-7	.20701967
	95% Confidence Interval for Mean	Lower Bound	-.4318355	
		Upper Bound	.4318355	
	5% Trimmed Mean		.0205583	
	Median		.3752123	
	Variance		.900	
	Std. Deviation		.94868330	
	Minimum		-1.76818	
	Maximum		1.40998	
	Range		3.17816	
	Interquartile Range		1.61812	
	Skewness		-.521	.501
	Kurtosis		-1.075	.972

Appendix (B16)

Descriptives

			Statistic	Std. Error
Standardized Residual	Mean		0E-7	.20701967
	95% Confidence Interval for Mean	Lower Bound	-.4318355	
		Upper Bound	.4318355	
	5% Trimmed Mean		-.0129864	
	Median		.0441381	
	Variance		.900	
	Std. Deviation		.94868330	
	Minimum		-1.40694	
	Maximum		1.65417	
	Range		3.06111	
	Interquartile Range		1.75090	
	Skewness		.091	.501
	Kurtosis		-1.384	.972