YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO SOLVE MULTICOLLINEARITY IN SESAME PRODUCTION OF MYANMAR (1995-2020)

KAY THI HTUN

M.Econ (Statistics)

Roll No. 3

SEPTEMBER, 2022

YANGON UNIVERSITY OF ECONOMICS

DEPARTMENT OF STATISTICS

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO SOLVE MULTICOLLINEARITY IN SESAME PRODUCTION OF MYANMAR (1995-2020)

This thesis is submitted as a partial fulfillment towards
the Degree of Master of Economics

BY

KAY THI HTUN
M.Econ (Statistics)
Roll No.3

SEPTEMBER, 2022

YANGON UNIVERSITY OF ECONOMICS
DEPARTMENT OF STATISTICS

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO SOLVE MULTICOLLINEARITY IN SESAME PRODUCTION OF MYANMAR (1995-2020)

This thesis is submitted as a partial fulfillment towards
the Degree of Master of Economics

Approved by the Board of Examiners

Supervised by:	**Submitted by:**
Daw Thida Win	Kay Thi Htun
Associate Professor	M.Econ (Statistics)
Department of Statistics	Roll No. 3
Yangon University of Economics	

SEPTEMBER, 2022

YANGON UNIVERSITY OF ECONOMICS
DEPARTMENT OF STATISTICS

This thesis certify that this thesis entitled "**APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO SOLVE MULTICOLLINEARITY IN SESAME PRODUCTION OF MYANMAR (1995-2020)**" submitted as a partial fulfillment towards the requirements to Master of Economics (Statistics) has been accepted by the Board of Examiners.

BOARD OF EXAMINERS

(Chair Person)
Professor Dr. Mya Thandar
Pro-Rector
Yangon University of Economics

(Internal Examiner)
Dr. Hlaing Hlaing Moe
Professor/ Head
Department of Applied Statistics
Yangon University of Economics

(External Examiner)
Daw Khin Nu Win
Associate Professor (Retired)
Department of Statistics
Yangon University of Economics

(Supervisor)
Daw Thida Win
Associate Professor
Department of Statistics
Yangon University of Economics

(Co-Supervisor)
Dr. Nyunt Nyunt Win
Professor
Department of Statistics
Yangon University of Economics

(Chief Examiner)
Dr. Aye Thida
Professor / Head
Department of Statistics
Yangon University of Economics

SEPTEMBER, 2022

ABSTRACT

The aim of this study is to apply the principal component analysis in solving the problem of multicollinearity for sesame production of Myanmar (1995-2020) with five explanatory variables which are sown acreage, harvested acreage, area of crops under irrigation, agricultural loan and quality seeds. The situations of these explanatory variables were described by using descriptive statistics. According to the correlation matrix of explanatory variables, it was found that these are highly correlated themselves. Since the symptoms of collinearity occur between sown acreage and harvested acreage. Thus, principal component analysis was applied to overcome multicollinearity problem. It was found that the two principal components. The first component consisted of sown acreage and harvested acreage, then quality seeds, agricultural loan and irrigation were included in the second component. According to the results of sesame production after removal of multicollinearity, sown acreage and harvested acreage are positively effect on sesame production and the explanatory variables can explain 84.2 % of variation in sesame production of Myanmar.

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to Professor Dr. Tin Tin Htwe, Rector of Yangon University of Economics, for his permission on this thesis.

I would like to express my deep and sincere gratitude to Professor Dr. Mya Thandar, Pro-Rector of Yangon University of Economics, for giving me the valuable comments and suggestions to improve this thesis.

I would like to express my indebtedness to Professor Dr. Aye Thida, Head of the Department of Statistics, Yangon University of Economics, for her helpful advice and gentle encouragement.

I would like to acknowledge and give my warmest thanks to my external examiner, Daw Khin Nu Win, Associate Professor (Retired), Department of Statistics, Yangon University of Economics and examiner Dr. Hlaing Hlaing Moe, Professor/ Head, Department of Applied Statistics, Yangon University of Economics for your brilliant comments and suggestions.

In addition, I am profoundly gratefully to all teachers in Department of Statistics, especially to my co-supervisor, Professor Dr. Nyunt Nyunt Win and supervisor, Daw Thida Win, Associate Professor, Department of Statistics, Yangon University of Economics, for their valuable guidances, helpful advice and suggestions throughout the time writing this thesis.

Furthermore, I would like to special thanks my parents for supporting and encouraging me not to be worried strongly during this study period.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
BLUE	Best Linear Unbiased Estimator
CLRM	Classical Linear Regression Model
DW	Durbin-Watson
HAR	Harvested
IRRI	Irrigation
MOAI	Ministry of Agriculture and Irrigation
MSE	Mean Squared Error
NAPA	National Action Plan for Agriculture
OLS	Ordinary Least Squares
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PROD	Production
QTY	Quality
SE	Standard Error
SSE	Sum of Squared Error
SSR	Sum of Squared Regression
SST	Sum of Squared Total
TOL	Tolerance
VIF	Variance Inflation Factor

CHAPTER I
INTRODUCTION

Among all statistical tools, multiple regression analysis is one of the most frequently employed. According to Myers (1990), regression analysis refers to a statistical technique for studying the relationship among variables and influence of one variable over the others. Thus the multiple linear regression model is a statistical technique which is applicable when someone needs to study the relationship between dependent variable and at least two independent variables. The link between several independent or predictor variables and one dependent or criterion variable is typically explained using the multiple regression analysis. A more complex model, containing additional independent variables, typically is more helpful in providing precise predictions of the response variable. Because it requires two or more predictor variables, multiple linear regression is also known as multiple regression.

1.1 Rationale of the Study

According to International Journal of Statistics and Applications (2018), when one of the independent variables is linearly correlated with one or more of the other independent variables, a multicollinearity problem occurs. One of the prerequisites for numerous regressions is broken in such a circumstance. There is a correlation between two or more independent variables in multiple linear regression analysis since there are numerous independent variables. The independent variable that correlates with each other is called multicollinearity (Zhou and Huang, 2018). Specifically, multicollinearity occurs if there is a high correlation between two independent variables, X_i and X_j. If the correlation coefficient r_{ij} between X_i and X_j in the multiple regression model is high, multicollinearity exists. Any time two or more independent variables are linearly related, some degree of multicollinearity exists.

Willis and Perlack (1978) investigated the multicollinearity is one of several problems confronting researchers using regression analysis. This study examines the regression model when the assumption of independence among the independent variables is violated. The basic properties of the least squares approach are examined, the concept of multicollinearity and its consequences on the least squares estimators are explained. The detection of multicollinearity and alternatives for handling the problem are then discussed.

The presence of multicollinearity creates many problems in use of multiple regression model. The most direct way of testing for multicollinearity is to produce a correlation matrix for all variables in the model. Another way to detect multicollinearity is to use the value of Tolerance. If the value of Tolerance is not less than 0.1, it can be said that there is no multicollinearity problem in this study.

The next way to detect multicollinearity is using the variance inflation factor (VIF). VIF measures the severity of multicollinearity in the regression analysis. VIF is another commonly used tool to detect whether multicollinearity exists in a regression model. If the VIF values are less than 10, so there is no multicollinearity (Alauddin and Nghiemb, 2010).

In any case, it is important to understand that most regression models with two or more independent variables exhibit some degree of multicollinearity. Because multicollinearity poses a significant challenge when attempting to draw conclusions for predictive models. Therefore, it is crucial that discover a better approach to handle multicollinearity. The main objective in this study is to introduce different models of principal component regression to solve multicollinearity problem.

In this study, sesame production data of Myanmar was used to explore the multicollinearity problem and the nature of independent variables. To detect multicollinearity assumption, the sesame production of Myanmar from 1995 to 2020 with five explanatory variables (sown acreage, harvested acreage, area of crops under irrigation, agricultural loan and quality seeds) were used to estimate. If there has multicollinearity problem in sesame production of Myanmar, principal component analysis are applied to solve this problem because of the principal component regression is suitable method for the multicollinearity problem.

1.2 Objectives of the Study

The objectives of the study are:

(i) To study the situation of sesame production in Myanmar.

(ii) To apply the principal component analysis to solve the multicollinearity problem in sesame production of Myanmar.

1.3 Method of Study

In this study descriptive analysis was used to explore information about the sesame production with five explanatory variables (sown acreage, harvested acreage, area of crops under irrigation, agricultural loan and quality seeds). And then, the effect of multicollinearity has been removed from the estimate of the regression coefficients by using principal component analysis are employed.

1.4 Scope and Limitations of the Study

In this study, the sesame production in Myanmar is analyzed by using principle component analysis based on the secondary data from the Statistical Yearbooks and Myanmar Agricultural Statistics. The study period is from year 1995 to year 2020.

1.5 Organization of the Study

This study is divided into five chapters. Chapter I consists of the introduction, rationale of the study, objectives of the study, method of study, scope and limitations of the study and organization of the study. Chapter II is the literature review. The theoretical background of multicollinearity, principal component analysis and multiple linear regression is described in Chapter III. Chapter IV presents the data analysis for production of sesame cultivated in Myanmar. Chapter V includes the conclusion of the study.

CHAPTER II
LITERATURE REVIEW

The researchers are used special terms that help describe these activities, when researchers are manipulating an environmental condition to determine its effect on behavior. A variable is any condition that can vary or change in quantity or quality. The independent variable, or treatment, is under the control and administered by the experimenter. The behavior that is potentially affected by the treatment and that it measure is called the dependent variable. The dependent variable is always a measure of behavior that it record after first manipulating the independent variable. It is referred to as dependent because changes in it depend on the effects of the independent variable. If a systematic relationship is found between the independent and dependent variables, then have established an empirical or causal relationship. It is also sometimes called a functional relationship because changes in the dependent variable are a function of values (different amounts) of the independent variable. From these lawful or functional relationships, it can construct theories and make predictions regarding future behavior. As discuss independent and dependent variables, will notice that they are always defined in precise and measurable terms.

Kaur (2013) studied variable is a term frequently used in research projects. It is pertinent to define and identify the variables while designing quantitative research projects. A variable incites excitement in any research than constants. It is therefore critical for beginners in research to have clarity about this term and the related concepts. Variable to put in layman statement is something that can change and or can have more than one value. It is pertinent for a researcher to know as how certain variables within a study are related to each other. It is thus important to define the variables to facilitate accurate explanation of the relationship between the variables. There is no limit to the number of variables that can be measured, although the more variables, the more complex the study and the more complex the statistical analysis. Moreover the longer the list of variables, the longer the time required for data collection.

Alibuhtto and Peiris (2015) focused on investigating the multicollinearity often causes a huge explanatory problem in multiple linear regression analysis. In presence of multicollinearity the (OLS) estimators are inaccurately estimated. The multicollinearity was detected by using observing correlation matrix, variance influence factor (VIF), and eigenvalues of the correlation matrix. The simulation multicollinearity data were generated using MINITAB software and make comparison between methods of principal component regression (PCR) and the OLS methods. According to the results of this study, it was found that PCR method facilitates to solve the multicollinearity problem.

Ayinde, Alabi and Nwosu (2021) presented the multicollinearity has remained a major problem in regression analysis and should be sustainably addressed. Problems associated with multicollinearity are worse when it occurs at high level among regressors. This review revealed that studies on the subject have focused on developing estimators regardless of effect of differences in levels of multicollinearity among regressors. In this studies have considered single-estimator and combined-estimator approaches without sustainable solution to multicollinearity problems. The possible influence of partitioning the regressors according to multicollinearity levels and extracting from each group to develop estimators that will estimate the parameters of a linear regression model when multicollinearity occurs is a new econometrics idea and therefore requires attention. The results of new studies should be compared with existing methods namely principal components estimator, partial least squares estimator, ridge regression estimator and the ordinary least squares estimators using wide range of criteria by ranking their performances at each level of multicollinearity parameter and sample size. Based on a recent clue in literature, it is possible to develop innovative estimator that will sustainably solve the problem of multicollinearity through partitioning and extraction of explanatory variables approaches and identify situations where the innovative estimator will produce most efficient result of the model parameters. The new estimator should be applied to real data and popularized for use.

Ghorbani (2020) investigated on the multicollinearity is a common problem in linear regression models when two or more regressors are highly correlated, which yields some serious problems for the ordinary least square estimates of the parameters as well as model validation and interpretation. In this paper, the problem of multicollinearity and its subsequent effects on the linear regression along with some important measures for detecting multicollinearity is reviewed, then the role of eigenvalues and eigenvectors in detecting multicollinearity are presented. At the end a real data set is evaluated for which the fitted linear regression model is investigated for multicollinearity diagnostics.

Perez (2017) studied the multiple linear regression models, covariates are sometimes correlated with one another. Multicollinearity can cause parameter estimates to be inaccurate, among many other statistical analysis problems. When these problems arise, there are various remedial measures it can take. Principal component analysis is one of these measures, and uses the manipulation and analyzation of data matrices to reduce covariate dimensions, while maximizing the amount of variation.

Mason and Perreault (1991) defined multiple regression analysis is one of the most widely used statistical procedures for both scholarly and applied marketing research. Yet, correlated predictor variables and potential collinearity effects are a common concern in interpretation of regression estimates. Though the literature on ways of coping with collinearity is extensive, relatively little effort has been made to clarify the conditions under which collineariy affects estimates developed with multiple regression analysis or how pronounced those effects are. The authors report research designed to address these issues. The results show, in many situations typical of published cross-sectional marketing research, that fears about the harmful effects of collinear predictors often are exaggerated.

Vatcheva and Rahbar (2016) investigated the adverse impact of ignoring multicollinearity on findings and data interpretation in regression analysis. The failure to identify and report multicollinearity could result in misleading interpretations of the results. We used simulated datasets and real life data from the Cameron County Hispanic Cohort to demonstrate the adverse effects of multicollinearity in the regression analysis and encourage researchers to consider the diagnostic for multicollinearity as one of the steps in regression analysis.

Joshi (2012) presented regression modeling is one of the most widely used statistical techniques in clinical trials. For instance, the model may fit the data well, even though none of the predictors has a statistically significant impact on explaining the outcome variable. This happens when multicollinearity exists between two or more predictor variables. If the problem of multicollinearity is not addressed properly, it can have a significant impact on the quality and stability of the fitted regression model. The aim of this thesis is to explain the issue of multicollinearity, effects of multicollinearity, various techniques to detect multicollinearity and the remedial measures one should take to deal with it.

Abdi and Williams (2010) investigated the most scientific areas employ Principal Component Analysis (PCA), which is arguably the most well-liked multivariate statistical method. It is also most likely the earliest multivariate method. When applied effectively, Principal Component Regression (PCR), a technique for overcoming multicollinearity, produces better estimation and prediction than conventional least squares. Principal components of the correlation matrix are a new set of orthogonal or uncorrelated variables created using this method from the original k climatic variables. Following this transformation, some of the principal components are eliminated in order to reduce variance. The new orthogonal variables are then ranked according to their significance. Ordinary least squares are used to perform a multiple regression analysis of the response variable against the smaller set of principal components after some of the principal components have been eliminated (OLS). After calculating the regression coefficients for the smaller set of orthogonal variables, a new set of coefficients that correspond to the original or starting set of correlated variables is created mathematically from the reduced set of coefficients. These new coefficients are principal component estimators.

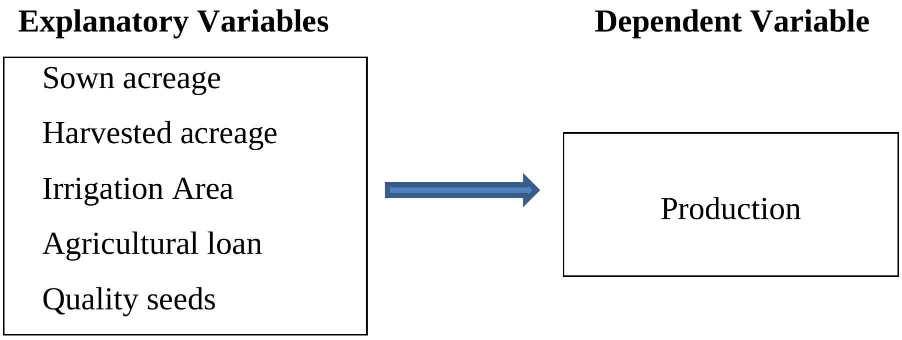
The independent variable is the antecedent while the dependent variable is the consequent. If the independent variable is an active variable then we manipulate the values of the variable to study its effect on another variable. When the explore completely new areas, little information is available to provide guidelines in selecting the independent variable. When dealing with quantitative independent variables, will faced with the additional problem of selecting appropriate values of the variable. This decision is important because too low a dosage may be insufficient to produce an effect, whereas too much may be harmful or even lethal. In fact, most researchers choose what they consider proper values of a treatment condition based on their own experiments and the published experiments of others. Some obvious guidelines should be considered when choosing values of an independent variable. A minimum of two groups is necessary to determine whether the independent variable has an effect. One of these groups would receive the treatment (experimental group), and the other group would either not receive the treatment (control group) or receive a different level of the treatment. If use more than two groups in an effort to ascertain whether increasing levels of the independent variable systematically influence behavior, the choice of values for the independent variable requires more thought. To counter this risk, the first thought that often comes to mind is to select extreme values of the independent variable. The logic is straight forward and simple. The greater difference in value between the experimental conditions, the greater the probability of showing that the independent variable has an effect. Therefore, choose two points along the continuum that are extreme values. Under certain circumstances, the logic is good and would provide an efficient way of determining whether the independent variable is powerful and worthy of additional investigation. Unfortunately, this simple logic could also result in coming to a wrong conclusion, depending on the relationship between the independent and dependent variable. Dependent variable is the variable that is affected by the independent variable. The dependent variable is dependent on the independent variable. The selection of a dependent variable is not in the least a casual matter. Indeed, it is immensely important. It reflects our underlying assumption that the study of behavior is the doorway toward measuring psychological states. Moreover, it is the measure use to ascertain whether the independent variable has an effect. Generally, choose a dependent measure because we judge that it will reveal unobservable but inferable processes that affect it and other behavioral measures. Often assume that our dependent variable reflects some underlying psychological state.

Although the main irrigation infrastructure has been completed, distribution canals and water courses to farmers' field are still under construction. Renovation of the distribution canals of completed dams and reservoirs has also been delayed due to limited maintenance budget. Extension and education activities on efficient utilization of irrigation water by water users are also inefficient due to suboptimal on-farm research and demonstration. There is a great potential for the expansion of irrigated areas by improving irrigation efficiency (FAO, 2016).

In such situations water use is highly inefficient and relatively few farmers benefit. According to Favre and Myint (2009), about 16% of Myanmar's cultivated area is sown with oilseeds, the third most important crop group in the country after cereals and pulses. Cultivation of pulses overtook oilseeds soon after the liberalization of the pulses trade. Important variations on production in main producing countries are because of the heavy dependence of rainfall for sesame crops. Myanmar's water resources are considerable and are centered on four major rivers and their related systems by Raitzer and Wong (2015). And then, less than 10 percent of the total water resources are utilized annually.

FAO (2016) presented that there is some involvement of the private sector in seed production, but this is still relatively limited. In addition, the monitoring of the health status and quality of certified seed, even for rice, does not comply with required technical standards. Farmers might not use certified seed because there is no incentive to do so. It is always assumed that the issue is supply: just produce more seeds and farmers will buy them. But seed availability without fertilizer, water, plant protection or a market is not enough to attract a farmer. The MOAI is responsible for all aspects of agriculture and irrigation as well as water resources with its mission to develop agriculture and irrigation nationally. At the state/regional level, agriculture is organized under a state/regional minister for agriculture who reports directly to the MOAI at the Union level in Nay Pyi Taw where most budget decisions are currently made. The main objective of the MOAI is stated as being to increase crop production. Among several strategies identified by the MOAI for meeting agriculture sector objectives are: the provision of irrigation, the application of modern agrotechnologies including improved seed, fertilizer and crop protection, the development and utilization of new crop varieties, and the development of new agricultural land.

In this study, the conceptual framework is illustrated to find the multicollinearity problem of sesame production in Myanmar. Thus, the dependent variable is the production (PROD) (in thousand ton). The explanatory variables that can be influenced on the dependent variable are sown acreage (SOWN) (in thousand acre), harvested acreage (HAR) (in thousand acre), area of crops under irrigation (IRRI) (in thousand acre), agricultural loan (LOAN) (in kyat millions) and quality seeds (QTY) (basket).



Source: Own compilation

Figure (2.1) Conceptual Framework for Sesame Production of Myanmar

The variables used in this study are defined as follows:

Crop production is the process of growing crops for domestic and commercial purposes. Some of the crops produced on a large scale include rice, wheat, maize, sesame, etc.

Sown acreage is plowed in order to grow a crop and according to the method of planting, if the seed is put into the ground or if it is a crop grown as a seedling or after the seedlings are taken from the nursery and transplanted into the plantation, these areas are called the sown area.

Harvested acreage is a subset of total agricultural acreage that does not include planted acreage that is not harvested.

The irrigated area is assumed to be irrigated for cultivation through such sources as canals (government & private), tanks, tube-wells, other wells and other sources. It is divided into two categories: (i) net irrigated area and (ii) total net un-irrigated area.

Agricultural loan means a loan made by a lending institution or by the authority to any person for the purpose of financing or refinancing land acquisition or improvement, irrigation, fertilizers, pesticides, machinery, containers or supplies or any other products employed in the production, cultivation, harvesting, marketing, distribution or export of agricultural products.

The use of quality seeds are considered as an important factor for increasing crop production. Quality seed is defined as varietally pure with a high germination percentage, free from disease and disease organisms and with a proper moisture content and weight. The use of quality seeds helps greatly in higher production per unit area to attain security of the country. Quality seeds have the ability of efficient utilization of the inputs such as fertilizers and irrigation.

CHAPTER III
METHODOLOGY

In this chapter, multiple linear regression model, multicollinearity and its effects and principal component analysis are the main presented.

3.1 Multiple Linear Regression Model

When two or more independent variables are to be utilized to estimate the dependent variable, multiple regression analysis is a technique for accounting for the relationship between all the variables at once. Extensions of the fundamental concepts used in two-variable regression analysis are applied in regression analysis with two or more independent variables. For the average relationship between the variables, an equation must be found.

In the linear equation that represents the multiple regressions model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad (3.1)$$

where Y_i = value of the dependent variable in the ith trial or observation

β_0 = constant in the regression equation, which indicates the value of Y when all $X_{ik} = 0$

$\beta_1,\ldots,\beta_k$= regression coefficients associated with each of the X_k independent variable

X_{ij} = value of the jth independent variable in the ith trial or observation, associated with the process of sampling

ε_i = the random error in the ith trial or observation, associate with the process of sampling, is assumed that normally distributed ($\varepsilon_i \sim$ N (0, σ^2).

Equation (3.1) can be written in matrix form

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

The least squares estimator $\hat{\beta}$ = $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ of the regression coefficients for the independent variable is

$$\hat{\beta} = (X'X)^{-1} X'y$$

and it's variance is

$$V(\hat{\beta}) = (X'X)^{-1} \sigma^2$$

Each column of X represents measurements for a particular independent variables.

3.1.1 The Standard Error of the Estimate

The standard error of the estimate, S_e, is found much as it was in the case of simple regression. The mean square error (MSE) is found by dividing the sum of the squared errors (SSE) by the degrees of freedom.

$$\text{MSE} = \frac{\sum(Y_i - \hat{Y})^2}{n-k-1}$$

Then,

$$S_e = \sqrt{\frac{\sum(Y_i - \hat{Y})^2}{n-k-1}}$$

This formula require that the predicted value of Y ($\hat{Y}$) be calculated for every observation. The error, the difference between this predicted value and the observed Y value (Y_i), is then squared and summed for all observation.

3.1.2 Evaluating the Model as a Whole

A multiple regression model can be evaluated using a number of tests. Calculate and interpret the standard error of the estimate for this study, analyze the model as a whole using ANOVA and the F-test, and assess the role of each independent variable using t-tests.

The overall F-test is used to test for the significance of overall multiple regression model. The ANOVA procedure tests the null hypothesis that all the β values are zero against the alternative that at least one β is not zero. The multiple regression models is defined as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

The hypothesis for F-test takes the following form

Null Hypothesis : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

There is no linear relationship between the dependent variable and the independent variables.

Alternative Hypothesis : At least one $\beta_j \neq 0$

There is a linear relationship between the dependent variable and at least one of the independent variables.

If the null hypothesis is rejected, it can be concluded that one or more of the parameters in the model is not equal to zero. Thus, the overall relationship between the dependent variable Y and the independent variables $X_1, X_2, \ldots, X_k$ is significant. However, if the null hypothesis is not rejected, its concluded that there is an overall significant relationship and the regression does not significantly to explain the variation in the independent variable.

This ratio of mean square regression to mean square error follows the F-distribution when the assumption that the residual are normally distributed is valid and the null hypothesis is true. The ratio of F-statistic;

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

where; the MSR is the mean square due to the regression which is equal to

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{\mathrm{k}}$$

where; the MSE is the mean square of error which is equal to

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{\mathrm{n-k-1}}$$

where; n-k-1 is the residual degrees of freedom and k is the number of independent variables. The decision rule for the F-test takes the following form;

Reject the null hypothesis if : $F > F_{\alpha,k,n-k-1}$

Do not reject the null hypothesis if : $F \leq F_{\alpha,k,n-k-1}$

where; $F_{\alpha,k,n-k-1}$ is based on the F distribution with k degrees of freedom in the numerator, n-k-1 degrees of freedom in the denominator and the probability of α in the upper-tail of the probability distribution.

3.1.3 Testing Individual Partial Regression Coefficient $\boldsymbol{\beta_j}$

An individual partial regression coefficient, β_j in the multiple regression model is tested to determine the significance of the relationship between x_i's and y. For any parameter β_j the hypothesis take the form.

Null Hypothesis : $\beta_j = 0$

Alternative Hypothesis : $\beta_j \neq 0$

The t statistic for $\hat{\beta}_j$ is simple to compute given $\hat{\beta}_j$ and its standard error:

$$t = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

The decision rule for this test takes the following form:

Reject the null hypothesis if : $|t| > t_{\alpha/2,n-k-1}$

Do not reject the null hypothesis if : $|t| \leq t_{\alpha/2,n-k-1}$

3.1.4 The Coefficient of Multiple Determination R^2

The coefficient of multiple determinations is defined as:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

The numerator of the middle term is the explained sum of squares or the sum of squares due to regression, SSR, as it is sometimes called. The denominator is the total sum of squares SST.

Therefore, it can be written as:

$$R^2 = \frac{SSR}{SST}$$

The coefficient of multiple determination demonstrates the percentage of Y's overall variability that can be accounted for by the independent variables. These are the percentages of the dependent variable's overall variation that the explanatory variables can account for. The value of R^2 will be between zero and one, where $R^2 = 0$, the regression model cannot explain anything about the variation in the dependent variable or the estimated model does not fit the data. The case of $R^2 = 1$ represents a perfect fit of the estimated model of the data. A high value of R^2 shows good fit and a low value of R^2 shows a poor fit.

3.1.5 The Adjusted Coefficient of Multiple Determination ($\bar{R}^2$)

A measure that recognized the number of independent variables in the regression model is called the adjusted coefficient of multiple determinations and is denoted by $\bar{R}^2$.

$$\bar{R}^2 = \frac{\frac{\sum(Y_i - \hat{Y})^2}{(n-k-1)}}{\frac{\sum(Y_i - \bar{Y})^2}{(n-1)}}$$

Reporting the adjusted R^2 is extremely important in comparing two or more regression models that predict the same dependent variable but have a different number of independent variables.

3.1.6 Residual Analysis

The Durbin-Watson statistic is used to test the hypothesis of no autocorrelation between error terms.

Null hypothesis : $\rho_{e_t,e_{t-1}} = 0$

There is no autocorrelation.

Alternative hypothesis : $\rho_{e_t,e_{t-1}} \neq 0$

There is an autocorrelation.

It can be calculated

$$d = \frac{\sum(e_t - e_{t-1})^2}{\sum(e_t)^2}$$

As a general rule, if d is close to 2, assume that autocorrelation is not a problem.

3.2 Multicollinearity and Its Effects

Gujarati (2004), presented the multicollinearity is a statistical phenomenon in which there exists a perfect or exact relationship between the predictor variables. When there is a perfect or exact relationship between the predictor variables, it is difficult to come up with reliable estimates of their individual coefficients. It will result in incorrect conclusions about the relationship between outcome variable and predictor variables.

A number of different techniques for solving the multicollinearity problem have been developed. These range from simple methods based on principal components to more specialized techniques for regularization, (Naes and Indahl, 1998).

3.2.1 The Nature of Multicollinearity

Gujarati (2009) presented the multicollinearity refers to the existence of more than one exact linear relationships and collinearity refers to existence of a single linear relationship. But this distinction is rarely maintained in practice and multicollinearity refers to both cases. That is it meant the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of a regression model. For the *k*-variable regression involving explanatory variables X_1, X_2,…, X_k (where $X_1 = 1$ for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following condition is satisfied.

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdot \cdot \cdot + \lambda_k X_k = 0 \qquad (3.2)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are constants such that not all of them are zero simultaneously. Assume that $\lambda_2 \neq 0$ Equation (3.2) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} \qquad (3.3)$$

Which shows how X_2 is exactly linearly related to other variables or how it can be derived from a linear combination of other *X* variables. In this situation, the coefficient of correlation between the variable X_2 and the linear combination on the right side of Equation (3.3) is bound to be unity.

Similarly, an imperfect linear relationship is said that X variables are intercorrelated but not perfectly, so, as follows:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdot \cdot \cdot + \lambda_k X_k + v_i = 0 \qquad (3.4)$$

Where v_i is a stochastic error term.

If $\lambda_2 \neq 0$ Equation (3.4) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda 2} v_i \qquad (3.5)$$

Which shows that X_2 is not an exact linear combination of the other X's because it is also determined by the stochastic error term v_i.

3.2.2 Estimation in the Presence of Perfect Multicollinearity

It was said that the regression coefficients remain indeterminate and their standard errors are limitless in the case of perfect multicollinearity. The absence of multicollinearity among the regressors included in the regression model is one of the assumptions made by the Classical Linear Regression Model (CLRM), (Gujarati, 1995). If three variables regression model is as follows:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \qquad (3.6)$$

where y is the dependent variable, x_2 and x_3 are the explanatory variables (or regressors), u is the stochastic disturbance term and i is the ith observation. The coefficients β_2 and β_3 are called the partial regression coefficients.

Using the deviation form, where all the variables are expressed as deviations from their sample means, it can express as

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \qquad (3.7)$$

where, $y_i = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} x_{21} & x_{31} \\ x_{22} & x_{32} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{2n} & x_{3n} \end{bmatrix}$, $\hat{\beta} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$, $u_i = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix}$

By using the Ordinary Least Squares (OLS) method, the parameter β can be estimated as follows:

$$\hat{\beta} = (x'\,x)^{-1}\,x'y$$

$$(x'\,x) = \begin{bmatrix} x_{21}\; x_{22} & \dots\dots\dots\, x_{2n} \\ x_{31}\;\; x_{32} & \dots\dots\dots\, x_{3n} \end{bmatrix} \begin{bmatrix} x_{21} & x_{31} \\ x_{22} & x_{32} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{2n} & x_{3n} \end{bmatrix} = \begin{bmatrix} \sum x_{2i}^2 & \sum x_{2i}\,x_{3i} \\ \sum x_{2i}\,x_{3i} & \sum x_{3i}^2 \end{bmatrix}$$

$$(x'\,x)^{-1} = \frac{1}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\,x_{3i})^2} \begin{bmatrix} \sum x_{3i}^2 & -\sum x_{2i}\;\, x_{3i} \\ -\sum x_{2i}\;\, x_{3i} & \sum x_{2i}^2 \end{bmatrix}$$

$$(x'y) = \begin{bmatrix} x_{21}\; x_{22} & \dots\dots\dots\, x_{2n} \\ x_{31}\;\; x_{32} & \dots\dots\dots\, x_{3n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i x_{2i} \\ \sum y_i x_{3i} \end{bmatrix}$$

$$\hat{\beta} = (x'\,x)^{-1}\,x'y$$

$$\hat{\beta} = \frac{1}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\,x_{3i})^2} \begin{bmatrix} \sum x_{3i}^2 & -\sum x_{2i}\;\, x_{3i} \\ -\sum x_{2i}\;\, x_{3i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} \sum y_i x_{2i} \\ \sum y_i x_{3i} \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \dfrac{(\sum x_{3i}^2)\,(\sum y_i x_{2i}) - (\sum y_i x_{3i})(\sum x_{2i}\,x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\,x_{3i})^2} \\ \dfrac{(\sum x_{2i}^2)(\sum y_i x_{3i}) - (\sum y_i x_{2i})(\sum y_i x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\,x_{3i})^2} \end{bmatrix}$$

$$\therefore \hat{\beta}_2 = \frac{\sum x_{3i}^2\,(\sum y_i x_{2i}) - (\sum y_i x_{3i})(\sum x_{2i}\,x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\,x_{3i})^2} \tag{3.8}$$

$$\therefore \hat{\beta}_3 = \frac{(\sum x_{2i}^2)(\sum y_i x_{3i}) - (\sum y_i x_{2i})(\sum y_i x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\, x_{3i})^2} \tag{3.9}$$

Assume that $x_{3i} = \lambda\, x_{2i}$, where λ is a nonzero constant substituting this into Equation (3.8), the estimator can be obtained as:

$$\therefore\ \hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} = \frac{0}{0} \tag{3.10}$$

Similarly, Equation (3.9) becomes

$$\therefore\ \hat{\beta}_3 = \frac{(\lambda \sum y_i x_{2i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\lambda \sum y_i x_{2i})}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} = \frac{0}{0} \tag{3.11}$$

which is an indeterminate expression.

The variance of $\hat{\beta}$ is

$$V(\hat{\beta}) = (x'\,x)^{-1}\,\sigma^2$$

$$V(\hat{\beta}) = \frac{\sigma^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\, x_{3i})^2} \begin{bmatrix} \sum x_{3i}^2 & -\sum x_{2i}\ x_{3i} \\ -\sum x_{2i}\ x_{3i} & \sum x_{2i}^2 \end{bmatrix}$$

$$V(\hat{\beta}_2) = \left[\frac{\sigma^2 \sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\, x_{3i})^2}\right] \tag{3.12}$$

$$V(\hat{\beta}_3) = \left[\frac{\sigma^2 \sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\, x_{3i})^2}\right] \tag{3.13}$$

Assume that $x_{3i} = \lambda\, x_{2i}$, where λ is a nonzero constant. Substituting this into Equations (3.12) and (3.13),

$$V(\hat{\beta}_2) = \left[\frac{\sigma^2 (\lambda^2 \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - (\lambda \sum x_{2i}^2)^2}\right] \approx \infty \tag{3.14}$$

$$V(\hat{\beta}_3) = \left[\frac{\sigma^2 (\sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - (\lambda \sum x_{2i}^2)^2}\right] \approx \infty \tag{3.15}$$

Thus, it can be said that if multicollinearity is perfect, the regression coefficients of the X variables are indeterminate and their standard errors are infinite.

3.2.3 Estimation in the Presence of "High" but Imperfect Multicollinearity

The perfect multicollinearity situation is a pathological extreme. Generally, there is no exact linear relationship among the X variables, especially in data involving economic time series. Thus, turning to the three variables model in the deviation from given in Equation (3.7), instead of exact multicollinearity, it may be expressed as

$$x_{3i} = \lambda\, x_{2i} + v_i \tag{3.16}$$

Where $\lambda \neq 0$ and v_i is a stochastic error term such that $\sum x_{2i}\, v_i = 0$

In this case, the estimation of regression coefficients β_2 and β_3 may be possible. Substituting Equation (3.16) in Equations (3.8) and (3.9), the estimators $\hat{\beta}_2$ and $\hat{\beta}_3$ become

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2} \tag{3.17}$$

$$\hat{\beta}_3 = \frac{(\sum x_{2i}^2)(\lambda \sum y_i x_{2i} + \sum y_i v_i) - (\sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2} \tag{3.18}$$

Where $\sum x_{2i} v_i = 0$.

There is no reason to believe a prior that Equation (3.17) cannot be estimated. Of course, if v_i is sufficiently small, say, very close to zero, Equation (3.16) will indicate almost perfect collinearity and it shall be back to the indeterminate case of Equation (3.10).

The variance of $\hat{\beta}$ is

$V(\hat{\beta}) = (x'\,x)^{-1}\,\sigma^2$

$$V(\hat{\beta}_2) = \left[\frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\, x_{3i})^2}\right] \sigma^2$$

$$V(\hat{\beta}_3) = \left[\frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}\, x_{3i})^2}\right] \sigma^2$$

$$V(\hat{\beta}_2) = \left[\frac{\sum(\lambda x_{2i}+V_i)^2}{(\sum x_{2i}^2)\sum(\lambda x_{2i}+v_i)^2 - (\sum x_{2i}(\lambda x_{2i}+v_i))^2}\right] \sigma^2$$

$$= \left[\frac{\sum(\lambda^2 \sum x_{2i}^2 + v_i^2 + 2\lambda\, x_{2i}\, v_i)}{(\sum x_{2i}^2)\sum(\lambda^2 x_{2i}^2 + v_i^2 + 2\lambda\, x_{2i}\, v_i) - \left(\lambda \sum x_{2i}^2 + \sum x_{2i}\, v_i\right)^2}\right] \sigma^2$$

$$= \left[\frac{(\lambda^2 \sum x_{2i}^2 + \sum v_i^2\ )}{(\sum x_{2i}^2 \sum v_i^2)}\right] \sigma^2\ , [\sum x_{2i}\, v_i\ = 0] \tag{3.19}$$

$$V(\hat{\beta}_3) = \left[\frac{\sum x_{2i}^2}{(\sum x_{2i}^2)\sum(\lambda x_{2i}+v_i)^2 - (\sum x_{2i}(\lambda x_{2i}+v_i))^2}\right] \sigma^2$$

$$= \left[\frac{\sum x_{2i}^2}{(\sum x_{2i}^2)\sum(\lambda^2 x_{2i}^2 + V_i^2 + 2\lambda\, x_{2i}\, v_i) - \left(\lambda \sum x_{2i}^2 + \sum x_{2i}\, v_i\right)^2}\right] \sigma^2$$

$$= \left[\frac{\sum x_{2i}^2}{(\sum x_{2i}^2 \sum v_i^2)}\right] \sigma^2\ , [\sum x_{2i}\, v_i\ = 0] \tag{3.20}$$

If v_i is sufficiently small, say, very close to zero, Equations (3.19) and (3.20) will indicate to infinite.

Therefore, it can be seen that if multicollinearity is less than perfect, as in Equation (3.3) the regression coefficients, although determinate, possess large standard errors, which means the coefficients cannot be estimated with great precision or accuracy.

3.2.4 Consequences of Multicollinearity

The consequences of multicollinearity are the theoretical consequences and the practical consequences.

(i) Theoretical Consequences of Multicollinearity

The Ordinary Least Squares (OLS) estimators of the regression estimators are Best Linear Unbiased Estimator (BLUE) if the classical model's presumptions are met. Now it can be demonstrated that the OLS estimators still maintain the property of BLUE even when multicollinearity is very strong, as in the situation of near multicollinearity.

Goldberger coined the term micronumerosity, to counter the exotic name multicollinearity. According to Goldberger, exact micronumerosity (the counter part of exact multicollinearity) arises when n, the sample size is zero, in which case any kind of estimation is impossible. Near micronumerosity, like near multicollinearity, arises when the number of observations barely exceeds the number of parameters to be estimated.

First, it is accurate that the OLS estimators are unbiased even in the presence of near multicollinearity. However, unbiasedness is a feature of repeated or multi-sample sampling. The average sample values will converge to the estimator's real population values as the number of samples rises, if one gets repeated samples and computes OLS estimators for each of these samples while maintaining the value of X variables constant. This says nothing about the characteristics of estimators in any given sample.

Second, it is also true that the property of minimum variance is not destroyed by collinearity. The OLS estimators are effective because they have minimum variances in the class of all linear unbiased estimators. But it does not mean that the variance of OLS estimator will necessarily be small.

Third, multicollinearity is essentially a sample (regression) phenomenon in the sense that even if the X variables are not linearly related in the population they may be so related in the particular sample. All of these factors make it less comforting in practice that the OLS estimator is BLUE despite multicollinearity.

(ii) Practical Consequences of Multicollinearity

In cases of near of high multicollinearity, one is likely to encounter the following consequences.

1. Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult.
2. Because of consequence 1, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” (i.e., the true population coefficient is zero) more readily.
3. Also because of consequence 1, the *t* ratio of one or more coefficients tends to be statistically insignificant.
4. Although the *t* ratios of one or more coefficients tend to be statistically insignificant, R^2, the overall measure of goodness of fit, can be very high.

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

(a) Large Variance and Covariances of OLS Estimator

The variances and covariances of $\hat{\beta}_2$ and $\hat{\beta}_3$ are given by

$$V(\hat{\beta}_2)= \frac{\sigma^2}{\sum x_{2i}^2\,(1-r_{23}^2)} \tag{3.21}$$

$$V(\hat{\beta}_3)= \frac{\sigma^2}{\sum x_{3i}^2\,(1-r_{23}^2)} \tag{3.22}$$

$$Cov(\hat{\beta}_2,\hat{\beta}_3)= \frac{-r_{23}\,\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2\,\sum x_{3i}^2}} \tag{3.23}$$

where r_{23} is the coefficient of correlation between X_2 and X_3.

It is apparent from Equations (3.21) and (3.22) that as r_{23} tends toward 1. That is, as collinearity increases, the variances of the two estimators increase and in the limit when r_{23}=1 , they are infinite. It is equally clear from Equation (3.23) that as r_{23} increases toward 1, the covariance of the two estimators also increases in absolute value.

The speed with which variances and covariances increase can be seen with variance-inflating factor (VIF), which is defined as

$$\text{VIF}= \frac{1}{(1-r_{23}^2)}$$

VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. As r_{23}^2 approaches 1, the VIF approaches infinity. As can be readily seen, if there is no collinearity between X_2 and X_3, VIF will be 1, using this definition, we can express as

$$V(\hat{\beta}_2)= \frac{\sigma^2}{\sum x_{2i}^2}\ \text{VIF}$$

$$V(\hat{\beta}_3)= \frac{\sigma^2}{\sum x_{3i}^2}\ \text{VIF}$$

Which show that the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$ are directly proportional to VIF.

(b) Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger. Thus, increase of high multicollinearity, the sample data may be compatible with a diverse set of hypotheses. Hence, the probability of accepting a false hypothesis (i.e., type II error) increases.

(c) Insignificant t Ratios

Recall that to test the null hypothesis that $\beta_2 = 0$, use the t ratio and compare the estimated t value with the critical t value from the t table. But in case of high collinearity the estimated standard errors increase dramatically, thereby making the t value smaller. Therefore, one will increasingly accept the null hypothesis that the relevant true population value is zero.

(d) High R^2 but Few Significant t Ratios

Consider the k-variable linear regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

In case of high collinearity, it is possible to find, that one or more of the partial slope coefficients are individually statistically insignificant on the basis of the t test. Yet the R^2 in such situations may be so high, say, in excess of 0.9, that on the basis of the F test one can convincingly reject the hypothesis that $\beta_2 = \beta_3 = \ldots = \beta_k = 0$. This one of the signals of multicollinearity insignificant t values but a high overall R^2 (and a significant F value).

3.2.5 Detection of Multicollinearity

After studied the nature of multicollinearity, the detection of multicollinearity will be discussed in this section.

Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity but between its various degrees. Since multicollinearity refers to the degree of relationship between explanatory variables that are assumed to be non-stochastic, it is a feature of the sample and not of the population. Since multicollinearity is a sample phenomenon do not have one unique method of detecting it for measuring its strength. But, it has some rules of thumb which all the same. Some of them are;

1. **High R^2 but few significant t ratios.** This is the classic symptom of multicollinearity. If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero.

2. **High pair-wise correlations among regressors.** Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess 0.8, then multicollinearity is a serious problem. The problem with this criterion is that, although high zero-order correlations may suggest collinearity, it is not necessary that may be high to have collinearity in any specific case.

3. **Eigenvalues and condition index.** Eigenvalues and the condition index may be used to diagnose multicollinearity. Montgomery and Peck (2021), presented the condition number k defined as

$$k = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}$$

And the condition index defined as

$$\text{CI} = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

If k is between 100 and 1000 there is moderate to strong multicollinearity and if its exceeds 1000 there is severe multicollinearity.

4. **Tolerance and Variance Inflation Factor.** For the k variable regression model, the variance of a partial regression coefficient can be expressed as

$$V(\hat{\beta}_j) = \frac{\delta^2}{\sum x_j^2} \cdot \left(\frac{1}{1 - R_j^2}\right)$$

$$= \frac{\delta^2}{\sum x_j^2} \cdot VIF_j$$

The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity. As per practical experience, if any of the VIF values exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity, Montgomery (2001).

Tolerance can also be used to detect multicollinearity. It is defined as,

$\text{TOL}_j = (1 - R_j^2)$

$= 1\text{- VIF}_j$

$\text{TOL}_j = 1$, if X_j is not correlated with other regressors

$\text{TOL}_j = 0$, if X_j is perfectly related to other regressors.

3.2.6 Remedy of Multicollinearity

Paul, R. K. (2006) focused on multicollinearity is often caused by the choice of model, such as when two highly correlated regressors used in the regression equation. In these situations some respecification of the regression equation may lessen the impact of multicollinearity. One approach to model respecification is to redefine the regressors. If X_1, X_2 and X_3 are linearly dependent, it may be possible to find some function such as $X = (X_1 + X_2) / X_3$ or $X = X_1 X_2 X_3$ that preserves the information content in the original regressors but reduces the ill conditioning.

Another widely used approach to model respecification is variable elimination. That is, if X_1, X_2 and X_3 are nearly linearly dependent, eliminating one regressor may be helpful in combating multicollinearity. Variable elimination is often a highly effective technique. However, it may not provide a satisfactory solution if the regressors dropped from the model have significant explanatory power relative to the response Y, that is eliminating regressors to reduce multicollinearity damage the predictive power of the model. Care must be exercised in variables selection because many of the selection procedures are seriously distorted by the multicollinearity and there is no assurance that final model will exhibit any lesser degree of multicollinearity than was present in the original data.

Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may not be serious as in the first sample. Sometimes simply increasing the size of the sample may attenuate the collinearity problem. If one uses more data, or increase the sample size, the effects of multicollinearity on the standard errors (SE) will decrease. This is because the standard errors are based on both the correlation between the sample size. The larger the sample size, the smaller is the SE.

Biased estimators of regression coefficients can also be obtained by using a procedure known as principal component regression. A small eigenvalues of $\boldsymbol{X'X}$ means that the variance of the corresponding regression coefficient will be large. The principal components regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal components estimator, assume that the regressors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ suppose that the last of these eigenvalues are approximately equal to zero. In principal components regression the principal components corresponding to near zero eigenvalues are removed from the analysis and least squares applied to the remaining components.

3.2.7 The Effects of Multicollinearity

A rule of thumb is that the sample correlation coefficient between two explanatory variables is greater than 0.8 or 0.9, then one have to say that there is a serious problem of multicolinearity. However, some use the determinant of $\boldsymbol{X'X}$ as a measure of multicollinearity. If the explanatory variables are standardized, then $\boldsymbol{X'X}$ become the matrix of simple correlation coefficient for X.

An important feature of note that the regression coefficient β_1 is the same where only X_1 is included in the model or both independent variables are included. The same holds for β_2. Thus is a result of the two independent variables being uncorrelated. The erroneous sum of squares is connected to another crucial property. In general, when two or more independent variables are uncorrelated, the marginal contribution of one independent variable in reducing the error sum of squares when the other independent variables are the model is exactly the same as when this independent variable is in the model alone. When the independent variables are uncorrelated, the effects ascribed to them by the regression model are the same no matter which other independent variables are included in the model.

Regression of Y on X_1 and X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Regression of Y on X_1

$$Y = \beta_0 + \beta_1 X_1$$

Regression of Y on X_2

$$Y = \beta_0 + \beta_2 X_2$$

The regression coefficient β_1 is the same whether only X_1 is included in the model or both independent variables are included.

Multicollinearity among the independent variables may have significant effects on how a fitted regression model is interpreted and applied. The pairwise coefficients of simple correlation between the independent variables, which are the diagnostic tool taken into consideration here for detecting multicollinearity, are frequently useful. However, there are situations when significant multicollinearity is present but not revealed by the pairwise correlation coefficients. Also take into account a variety of corrective actions for reducing the impact of multicollinearity.

It was noted that a near-zero determinant of $\boldsymbol{X'X}$ is a potential source of serious round off errors in least square results. This determinant approaches zero when there is severe multicollinearity. As a result, in cases of extreme multicollinearity, the regression coefficients may be affected by significant rounding errors and sampling variances. Therefore, when multicollinearity is present, it is especially advised to use the correlation transformation when constructing the regression model.

When the independent variables are highly correlated, the partial correlation coefficients between the dependent variable and each of the independent variables also tend to become erratic from sample to sample, making the estimated regression coefficient less precise. When the model's variables are changed using the correlation transformation, the impact of intercorrelations between the independent variables on the standard deviations of the calculated regression coefficients can be quickly observed.

3.3 Principal Component Analysis

Johnson & Wichern (2002) presented through a few linear combinations of these variables, a principle component analysis attempts to describe the variance-covariance structure of a set of variables. Data interpretation and minimization are its main goals.

Although p components are necessary to replicate the overall system variability, k of the primary components often account for the majority of this variability. If the k components contain the same amount of data that the original p variables did. When the initial p variables are replaced by the k principal components, the original data set of n measurements on p variables is transformed into a data set of n measurements on k principal components.

Principal component analysis frequently exposes relationships that were not previously known, allowing for interpretations that would not typically be possible. Principal components are specific linear combinations of the p random variables X_1, X_2, ... , X_p in algebra. These linear combinations, which have variables X_1, X_2, ... , X_p as the coordinate axes, describe the choice of a new coordinate system in terms of geometry. The new axes show the directions with the highest degree of variability and offer a more concise and straight forward explanation of the covariance structure.

The only factor that influences principal components is the covariance matrix $\sum$ (or correlation matrix $\boldsymbol{\rho}$) of variables X_1, X_2, ... , X_p. It is not necessary to use a multivariate normal assumption for their development. On the other hand, the constant density ellipsoids provide appropriate interpretations for main components determined for multivariate normal populations.

Let the random vector $\mathbf{X}'$ = [X_1, X_2 ,…, X_p] have the covariance matrix $\sum$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Consider the linear combinations

$$Y_1 = \boldsymbol{a_1'}\,\boldsymbol{X} = a_{11}\,X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

$$Y_2 = \boldsymbol{a_2'}\,\boldsymbol{X} = a_{21}\,X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = \boldsymbol{a_p'}\,\boldsymbol{X} = a_{p1}\,X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances in (3.1) are as large as possible.

The first principal component is the linear combination with maximum variance. That is, it maximizes Var (Y_1) = $\boldsymbol{a}_1' \sum \boldsymbol{a}_1$. It is clear that Var ($Y_1$) = $\boldsymbol{a}_1' \sum \boldsymbol{a}_1$ can be increased by multiplying any $\boldsymbol{a}_1$ by some constant. It is convenient to focus attention just on coefficient vectors of unit length in order to eliminate this uncertainty. Consequently, it can define

First principal component = linear combination $\boldsymbol{a}_1'\,\boldsymbol{X}$ that maximizes
Var ($\boldsymbol{a}_1'\,\boldsymbol{X}$) subject to $\boldsymbol{a}_1'\,\boldsymbol{a}_1 = 1$

Second principal component = linear combination $\boldsymbol{a}_2'\,\boldsymbol{X}$ that maximizes
Var ($\boldsymbol{a}_2'\,\boldsymbol{X}$) subject to $\boldsymbol{a}_2'\,\boldsymbol{a}_2 = 1$ and
Cov($\boldsymbol{a}_1'\,\boldsymbol{X}$, $\boldsymbol{a}_2'\,\boldsymbol{X}$) = 0

At the ith step

ith principal component = linear combination $\boldsymbol{a}_i' \boldsymbol{X}$ that maximizes

Var ($\boldsymbol{a}_i' \boldsymbol{X}$) subject to $\boldsymbol{a}_i' \boldsymbol{a}_i$= 1 and

Cov ($\boldsymbol{a}_i' \boldsymbol{X}$, $\boldsymbol{a}_k' \boldsymbol{X}$) = 0 for k < i

Let ∑ is the diagonal matrix

$$\sum = \begin{bmatrix} \sigma_{11} & 0 \ldots\ldots\ldots & 0 \\ 0 & \sigma_{22} \ldots.\ldots & 0 \\ . & . & . \\ . & . & . \\ 0 & 0 \ldots\ldots\ldots & \sigma_{pp} \end{bmatrix}$$

A regression variable's multicollinearity can come from a variety of causes. Small eigenvalues $\boldsymbol{X'X}$ are an indication that one or more of these issues are present. It is obvious that the number of input variables can be decreased if there are zero eigenvalues. The situation is less evident if the small eigenvalues are close to zero. Departures from zero could be the result of measurement near reliance, and they might indicate true linear dependence. It's unclear what to do in this situation.

The multiple regression model makes the assumption that there isn't a precise linear relationship between the explanatory variables. If a relationship of this nature does exist, it indicates that the explanatory variables are perfectly collinear. The coefficient vector is not estimable in this situation. This is referred to as a situation where the explanatory variables are multicollinear.

Prediction is generally regarded as adequate when increases in the predictor variable are strongly correlated with increases or decreases in the response variable. Predictor variables contribute to prediction of a response variable only to the extent that variation in the predictor variables helps to account for an explain variation in the response variable prediction equations.

The predictor variable usually has limited usefulness in forecasting the response if it stays largely constant while the responder variable changes. Multicollinearities are generally constant for all responses linear combinations of predictor variables. When the response variable varies but the multicollinearities remain largely constant, they are typically in a data base, which dramatically increases least squares estimator variances and causes other issues with coefficient estimates. This is similar to the case with single-variable least squares. Principal component coefficient estimators aim to add as little bias as possible while significantly reducing estimator variances by removing multicollinearities from the least squares estimator.

3.3.1 Objectives of Principal Component Analysis

According to statistical theory, the set of main components produces a practical set of coordinates, and the component's corresponding variances describe their statistical characteristics. In statistical practice, the linear combination with a high variance is found using the principal component approach. The number of variables considered in many exploratory investigations is too great to handle. Discarding the linear combinations with low variances is one method of lowering the number of variables to be treated because the deviations in these studies are what are important.

Abdi and Williams (2010) investigated the main goal of PCA is to minimize the dimensionality of a data set made up of many interconnected variables while preserving as much of the data set's variance as feasible. The goals of PCA are:

(1) to extract the most important information from the data;

(2) to compress the size of the data set by keeping only important information;

(3) to simplify the description of the data set and

(4) to analyze the structure of the observations and the variables.

PCA computes new variables—called principal components—that are generated as linear combinations of the original variables in order to accomplish these objectives. It is necessary for the first primary component to be as large as possible. The second component must have the biggest variance and be orthogonal to the first component in order to be computed. The order components are calculated in a similar way. Factor scores are the values of these new variables for the observations, and these factor scores can be geometrically understood as the observations projected onto the principal components.

The PCR method may be broadly divided into three major steps:

(1) To acquire the principal components, perform PCA on the observed data matrix for the explanatory variables. Then, choose a subset of the obtained principal components for further use based on some approximative criteria.

(2) Next, perform an ordinary least squares regression on the observe vector of results to determine the predicted regression coefficients for the principle components that were chosen as variables (with dimension equal to the number of selected principal components).

(3) The final PCR estimator (with dimension equal to the total number of covariates) is obtained by transforming this vector back to the scale of the

actual covariates using the chosen PCA loadings (the eigenvectors corresponding to the chosen principal components). This estimator is then used to estimate the regression coefficients that define the original model.

3.3.2 Determining the Number of Principal Components

There is always the question of how many components to retain. There is no definitive answer to this question. Things to consider include the amount of total sample variance explained, the relative sizes of the eigenvalues (the variance of the sample components), and the subject-matter interpretations of the components. In addition, a component associated with an eigenvalue near zero and, hence, deemed unimportant, may indicate as unsuspected linear dependency in the data.

A useful visual aid to determining an appropriate number of principal components is a scree plot. With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus *i*—the magnitude of an eigenvalue versus its number. To determine the appropriate number of components, look for an elbow in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.

3.3.3 Standardized Principal Components

Variables should probably be standardized if the variables are measured on scales with widely differing ranges or if the units of measurement are not commensurate. Principal components may also be obtained for the standardized variables

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

$$\vdots \qquad \vdots$$

$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

In matrix notation,

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu})$$

where the diagonal standard deviation matrix $\mathbf{V}^{1/2}$ is defined in

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Clearly, E ($\mathbf{Z}$) = 0 and

$$\text{Cov (Z)} = (\mathbf{V}^{1/2})^{-1} \sum (\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$$

The principal components of $\mathbf{Z}$ may be obtained from the eigenvectors of the correlations matrix $\boldsymbol{\rho}$ of **X**. All previous results apply, with some simplifications, since the variance of each Z_i is unity and shall continue to use the notation Y_i to refer to the ith principal component and (λ_i, e_i) for the eigenvalue—eigenvector pair from either $\boldsymbol{\rho}$ or $\sum$. However, the (λ_i, e_i) derived from $\sum$ are, in general, not the same as the ones derived from $\boldsymbol{\rho}$.

CHAPTER IV

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO MULTICOLLINEAR DATA

In this chapter, describe the application of principal component analysis by using the sesame production data from the period 1995 to 2020. The firstly describe the descriptive statistics of sesame production data and the secondly describe the correlation matrix, summarizing sample variation and detecting and remedy of multicollinearity. In this study, the dependent variable is the production (PROD) and the explanatory variables considered in the present study are sown acreage (SOWN), harvested acreage (HAR), irrigation (IRRI), agricultural loan (LOAN) and quality seeds (QTY).

4.1 Descriptive Statistics

In this section, mean value, standard deviation, minimum value and maximum value of sesame production in Myanmar are expressed. The descriptive Statistics for production, sown, harvested, irrigation, loan and quality seeds of sesame production in Myanmar is presented as shown in Table (4.1).

Table (4.1)

Descriptive Statistics for Sesame in Myanmar

Variable	Mean	Std. Deviation	Minimum	Maximum
PROD	578.246	229.090	253.2	854
SOWN	3568.46	390.979	2557	4052
HAR	3192.19	632.796	1521	3863
IRRI	225.643	56.942	147.76	335.49
LOAN	17226.410	32623.938	96.27	123518.10
QTY	1195.62	1246.126	68	5091

Source: Own calculation (2022)

In Table (4.1), the sesame production ranges between 253.2 ton and 854 ton with mean 578.246 ton and standard deviation 229.090 ton. Sown acreage ranges between 2557 acres and 4052 acres with mean 3568.46 acres and with standard deviation 390.979 acres and harvested acreage ranges between 1521 acres and 3863 acres with mean 3192.19 acres and with standard deviation 632.796 acres. Irrigation ranges between 147.76 acres and 335.49 acres with mean 225.643 acres and with standard deviation 56.942 acres and agricultural loan ranges between 96.27 kyats in millions and 123518.10 kyats in millions with mean 17226.410 kyats in millions and with standard deviation 32623.938 kyats in millions. Quality seeds ranges between 68 baskets and 5091 baskets with mean 1195.62 baskets and with standard deviation 1246.126 baskets.

4.2 Computation of the Correlation Matrix

Principal Component Analysis (PCA) is based on correlations between measured variables, a correlation matrix containing the inter correlation coefficients for the variables must be computed. The variables should be measured at least the ordinal level, although two-category nominal variables can be used. If all variables are nominal variables, then specialized forms of factor analysis, such as Boolean factor analysis (BMDP, 2004) are more appropriate.

Table (4.2)

Correlation Matrix

		SOWN	HAR	IRRI	LOAN	QTY
Correlation	**SOWN**	1.000	.901	.124	.134	-.360
	HAR	.901	1.000	-.044	.287	-.216
	IRRI	.124	-.044	1.000	.554	.094
	LOAN	.134	.287	.554	1.000	.197
	QTY	-.360	-.216	.094	.197	1.000

Source: Own calculation (2022)

One of these independent factors is eliminated from the correlation matrix if any of the dependent variables included there have a strong correlation with any other dependent variable or if the value of the correlation is significant. The values of correlation among the explanatory variables are shown in above Table (4.2).

In correlation matrix of Table (4.2), these correlation coefficients demonstrate among the explanatory variables have correlations with each other. High correlations between explanatory variables might lead to multicollinearity issues. From the Table (4.2), it can be seen that there are relatively highly correlations among sown acreage and harvested acreage. Likewise, sown acreage and harvested acreage have a weakly correlation among all other explanatory variables.

There are relatively fairly correlations among irrigation and agricultural loan. Then, there are relatively weakly correlation among irrigation and other remaining explanatory variables.

There are relatively weakly correlations among quality seeds and all other explanatory variables. Additionally, given that the pair-wise correlations of the majority of the explanatory variables are quite high, multicollinearity may have developed in the data set.

Table (4.3)

Eigenvalues of Correlation

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Number
1	2.121	42.428	42.428	1.000
2	1.603	32.052	74.479	1.324
3	0.817	16.343	90.823	2.596
4	0.429	8.574	99.397	4.948
5	0.030	0.603	100.000	70.362

Source: Own calculation (2022)

Table (4.3) gives an eigenvalue analysis of the independent variables after they have been centered and scaled. The sum of all eigenvalues of the correlation matrix is 5 (2.121 + 1.603 + … + 0.030) that point out is equal to the number of explanatory variables. Results of eigenvalues in Table (4.3) indicated the multicollinearity problem in the data set because of eigenvalues near zero be a sign of a multicollinearity problem.

The third column of Table (4.3) shows incremental percent. This eigenvalue's proportion of the total is expressed in incremental percentages. This percent approaching 0, according to the same eigenvalue theory, point to a multicollinearity issue in the data.

The last column of eigenvalue of correlation is condition number, largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalue are really variances, the condition number is a ratio of variances. Condition numbers greater than 1000 indicate severe multicollinearity problem and between 100 and 1000 indicate a mild multicollinearity problem (Montgomery, 2001). According to that, it was found the multicollinearity is present.

The report of Table (4.3) displays the eigenvectors associated with each eigenvalue. By rotating the axis from those specified by the variables to a new set defined by the variances of the variables, eigenvalue analysis is based on the idea that the axes can be more accurately described. By using weighted averages of the standardized original variables, rotation is achieved.

4.3 Detecting Multicollinearity

In this section, based on the time series data of sesame production and the explanatory variables are sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds during the period of 1995 to 2020 are calculated. The results are shown the following Table (4.4).

From the below Table (4.4), predictor variable irrigation is a negative correlation with dependent variable production since the estimated regression coefficient β_3 has a negative sign.

Table (4.4)

Least Squares Multicollinearity Results of Sesame Production

Variables	**Unstandardized Coefficients**		**Standardized Coefficients**	**T**	**Sig.**	**Collinearity Statistics**	
	B	**Std. Error**	**Beta**			**TOL**	**VIF**
Constant	-1084.612***	287.309		-3.775	0.001		
SOWN	0.512***	0.196	0.875	2.620	0.016	0.069	14.441
HAR	0.009	0.124	0.025	0.072	0.944	0.065	15.285
IRRI	-1.152**	0.652	-0.286	-1.768	0.092	0.294	3.402
LOAN	0.023	0.019	0.126	1.236	0.231	0.743	1.346
QTY	0.002**	0.001	0.313	1.954	0.065	0.300	3.330
R-squared		0.846	F-statistic			21.923***	
Adjusted R-squared		0.807	Prob (F-statistic)			.000	
Std. Error of the Estimate		100.613					
Durbin-Watson		0.790					

Source: Own calculation

*** denote significant at 1% level.

** denote significant at 10% level.

The estimated regression equation for sesame production is

PROD = -1084.612 + 0.512 SOWN + 0.009 HAR -1.152 IRRI + 0.023 LOAN + 0.002 QTY

From the estimated regression equation for sesame production, it is found that sesame production is positively related to sown acreage, harvested acreage, agricultural loan and quality seeds and is negatively related to irrigation. If holding the harvested acreage, irrigation, agricultural loan and quality seeds are constant, a 1 thousand acre increase in sown acreage led on the average to about 0.512 ton increase in production and holding the sown acreage, irrigation, agricultural loan and quality seeds are constant, a 1 thousand acre increase in harvested acreage led on the average to about 0.009 thousand acreage increase in production. Holding the sown acreage, harvested acreage, agricultural loan and quality seeds are constant, a 1 thousand acre increase in irrigation led on the average to about 1.152 thousand acreage decrease in production. Holding the sown acreage, harvested acreage, irrigation and quality seeds are constant, 1 kyat millions increase in agricultural loan led on the average to about 0.023 thousand acreage increase in production and 1 basket increase in quality seeds led on the average to about 0.002 thousand acreage increase in production holding the other variables are constant.

The sown acreage (2.620) is statistically significant at 1% level and irrigation (1.768) and quality seed (1.954) are also statistically significant at 10% level but other variables (harvested acreage and agricultural loan) are not significant. The result from Table (4.2) indicates that the overall model is significant (F-value is 21.923, p-value, 0.000). Specifically, an assessment of statistical significance of individual predictors indicates that some predictors are significant while others are insignificant. The major reason is presence of collinearity among predictors. It can be concluded that a linear relationship exists between sesame production and at least one of the explanatory variables.

The variance inflation factor (VIF) is a measure of multicollinearity. It is the reciprocal of 1- R_x^2 , where R_x^2 is the R^2 obtained when this variable is regressed on the remaining independent variables. The R_x^2 for sown is 0.931, harvested is 0.935, irrigation is 0.706, loan is 0.257 and quality seeds is 0.7. After that, the VIF is calculated by $\frac{1}{1-R_x^2}$ that is illustrate in Table (4.2). As the result of Table (4.2), the VIF values of sown acreage, irrigation, agricultural loan and quality seeds are 14.441, 15.285, 3.402, 1.346 and 3.330 and VIF values of sown acreage and harvested acreage are greater than 10 and the VIF values of irrigation, agricultural loan and quality seeds are not greater than 10. Tolerance is just 1- R_x^2 , the denominator of the variance inflation factor (VIF). According to the results, tolerance values of each variable are 0.069, 0.065, 0.294, 0.743 and 0.300 which are closely to 0. Therefore, these VIF and TOL values are not acceptable. It indicates that there occurs a multicollinearity problem in this study.

The R value is 0.920, it means that the production of sesame has a positive relationship with sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds.

The coefficient of multiple determination demonstrates the percentage of dependent variable overall variability that can be accounted for by the explanatory variables. The R-Square value is 0.846, 84.6% of the variation in sesame production was explained in terms of sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds. Because of R-Square value is 0.846, in this result the value of R-Square shows good fit.

Adjusted R-Square value is 0.807 and it means that all the explanatory variables can explain 80.7% of variation in sesame production is explained sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds and the remaining percentage 19.3% due to other factors that are not included in the model. It has been found that there is a strongly positive relationship between the sesame production of Myanmar with sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds. The standard error of the estimate (100.613) is very large.

Figure (4.1)

Durbin-Watson Statistic

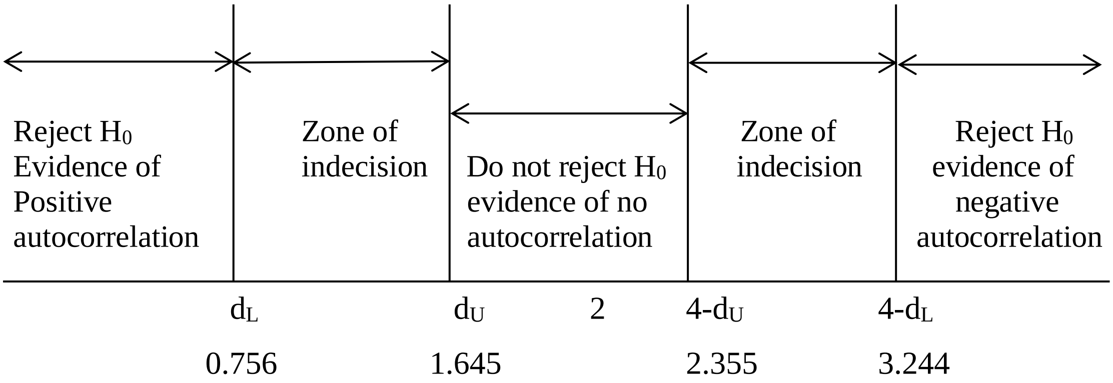

Reject H_0 Evidence of Positive autocorrelation
Zone of indecision
Do not reject H_0 evidence of no autocorrelation
Zone of indecision
Reject H_0 evidence of negative autocorrelation
d_L
d_U
2
4-d_U
4-d_L
0.756
1.645
2.355
3.244

Figure (4.1), represents to determine if the null hypothesis of no autocorrelation is rejected. For α = 0.01 or 1% level of significance, critical values for the Durbin-Watson d statistic are d_L = 0.756 and d_U = 1.645. Since, DW = 0.790, the null hypothesis is not rejected and it is concluded that there is no evidence of autocorrelation. As a general rule, if d is less than 2, assume that the Durbin-Watson d statistic is a positive autocorrelation.

4.4 Summarizing Sample Variation

The sample variation of the components has been described in Table (4.5).

Table (4.5)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	%of Variance	Cumulative %	Total	%of Variance	Cumulative %	Total	%of Variance	Cumulative %
1	2.121	42.428	42.428	2.121	42.428	42.428	2.075	41.500	41.500
2	1.603	32.052	74.479	1.603	32.052	74.479	1.649	32.980	74.479
3	0.817	16.343	90.823						
4	0.429	8.574	99.397						
5	0.030	0.603	100.000						

Source: Own calculation (2022)

Extraction Method: Principal Component Analysis

Table (4.5) presents principal components together with its corresponding eigenvalues and total variance explained. The principal components in this attributes of satisfaction are uncorrelated attributes in the original data set. The eigenvalues were listed in descending order from largest to smallest value.

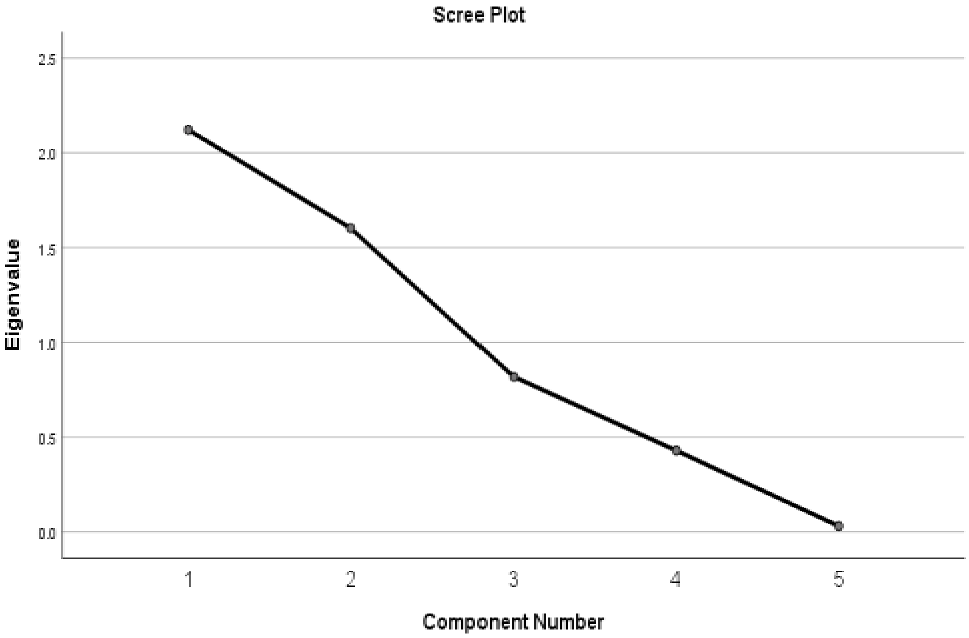
According to Table (4.5), the total variance explained section presents the number of common factors computed, the eigenvalues associated with these factors, the percentage of total variance accounted for by each factor and the cumulative percentage of total variance accounted for by the factors. Although five factors have been computed, it is obvious that not all five factors will be useful in representing the list of five variables. There were two eigenvalues (2.121, 1.603) greater than 1.0. The first principal component has largest variance that account for 42.4% of the total variance. This PC has comparatively largest eigenvalue of 2.1 which is equivalent to the eigenvalues of two variables. The second PC has an eigenvalue of 1.6 that accounts to 32.1% of the variability of the data. As a rule of thumb, the first two components have eigenvalue greater than one and collectively account to 74.5% of variability of the original data set losing only 25.5% of the information. Therefore only two principal components are extracted and retained from five principal components without much loss of information. This implies that the original information was reduced from 5-dimension of data set into a minimum size (2-dimension) while at the same time maximizing the variability of the original data set. The principal component analysis can frequently be used as a solution to multicollinearity.

4.5 The Number of Principal Components

The important useful criterion for deciding the number of principal components to be retained is based on observation of visual appearance of scree plot. The arrangement of eigenvalues are arranged in descending order from largest to smallest values which are presented on y-axis against number of principal components. Determining the maximum eigenvalues based on scree plot is subject to researcher judgment and some time the break point cannot be seen clearly.

The following Figure (4.2) represents the eigenvalues plotted against the corresponding component. A scree plot is observed from the three components onwards indicating that an eigenvalue is less than 1, and only two components have been retained.

Figure (4.2)



Source: SPSS Output

Table (4.6)

Rotated Component Matrix

	Component	
	1	**2**
SOWN	0.955	0.114
HAR	0.925	0.151
QTY	-0.536	0.395
LOAN	0.139	0.882
IRRI	-0.016	0.824

Source: own calculation (2022)

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization

a. Rotation converged in 3 iterations.

In above Table (4.6) indicate the outputs of varimax methods of rotation which is used to smooth the loadings and hence simplify interpretation. The outputs of varimax provide clear interpretation of the principal components in a way that only high loadings are retained to specific components and the low loadings are minimized. This improves impression of output by identifying the variables that are highly related to a corresponding PC.

Thus the two selected principal components are the linear combination of the original variables that contribute much to the total variance. The fitted PCs are:

Z_1 = 0.955 SOWN + 0.925 HAR

Z_2 = 0.395 QTY + 0.882 LOAN + 0.824 IRRI

The first principal component is the linear combination of two variables which are related namely variable SOWN and HAR. The second principal component composed of variable QTY, LOAN and IRRI which are highly related. The variables within components are highly related while the group of variables in a particular components are not related with another group of variables loaded to another component.

After removing the principal components which are less important, the modified linear regression model is now:

$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon$

Where predictors Z_1 and Z_2 are principal components

Table (4.7)

Results of Sesame Production After Removal of Multicollinearity

Variables	**Unstandardized Coefficients**		**Standardized Coefficients**	**T**	**Sig.**	**Collinearity Statistics**	
	B	**Std. Error**	**Beta**			**TOL**	**VIF**
Constant	-817.814***	125.103		-6.537	.000		
Z_1	.215***	.020	.886	10.856	.000	.948	1.055
Z_2	.001	.001	.131	1.608	.121	.948	1.055
R-squared		0.855	F-statistic			67.706	
Adjusted R-squared		0.842	Prob (F-statistic)			.000	
Std. Error of the Estimate		90.1017					
Durbin-Watson		0.689					

Source: Own calculation

*** denote significant at 1% level.

Based on the results shown in above table, the fitted model can be developed as follows:

$$\text{PROD} = -817.814 + 0.215\, Z_1 + 0.001\, Z_2$$

Instead of using OLS method of estimating parameters in linear regression model, principal components regression was used. Table (4.7) indicates results of regressing dependent variable on two explanatory variables (PCs). In comparison this method brought some changes on standard error. In the original regression model where OLS method was employed, standard errors of estimate coefficients were large that weakened the statistical power due to presence of severe multicollinearity. This contrary to the output of regression model after utilizing principal component, where the standard errors were smaller compared to the OLS methods. In assessing whether the collinearity exist between variables, VIF was computed on each of the variables treating as dependent variable and regress on the rest of the variables. The result indicates that multicollinearity problem was eliminated since VIF values for each of the variables were less than 10.

In Table (4.7), the VIF values of Z_1 and Z_2 are 1.055 and the total value is 2.11 which are not greater than 10 and TOL values are 0.948 which is closely to one. As a result, this study assumes that there is no multicollinearity and provides an acceptable amount of near collinearity. The PCR approach indicates the correct sign for the coefficient of sesame production from the above fitted model. It provides the best-fit model for manufacturing as a result.

The coefficient of multiple determination demonstrates the percentage of dependent variable overall variability that can be accounted for by the explanatory variables. The R-Square value is 0.855, 85.5% of the variation in sesame production was explained in terms of explanatory variables. Because of R-Square value is 0.855, in this result the value of R-Square shows good fit.

Adjusted R-Square value is 0.842 and it means that all the explanatory variables can explain 84.2% of variation in sesame production and the remaining percentage 15.8% due to other factors that are not included in the model.

CHAPTER V
CONCLUSION

In this chapter, the findings of sesame production of Myanmar are presented with recommendations and further studies.

5.1 Findings and Discussions

This thesis was analyzed the sesame production data of Myanmar for 25 years from 1995 to 2020. Sesame production of Myanmar has been analyzed in this thesis based on five predictions which are sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds. Multicollinearity is generally defined as the existence of an exact or nearly exact linear relationship between the explanatory variables. In this thesis, the issue of multicollinearity and its effects are investigated, as well as how to recognize collinearity in any given circumstance. Then, in the event of multicollinearity, Principal Component Analysis (PCA) is the most often utilized biased regression technique that may be applied.

The key objective was to demonstrate how principal components method can be used to eliminate multicollinearity problem that may exist when running linear regression model. The real application of the techniques was presented in the problem of predicting factors influencing sesame production where overall was predicted on five variables. The results of linear regression model revealed a large standard error of coefficients, the situation which resulted into biasness of the mean estimates of the coefficients. The major reason is the violation of ordinarily least square assumption that requires the predictors to be independent. The variance inflation factor was used as indicator to detect collinearity among predictors. It was observed that VIF values of two predictors exceed 10 which indicate presence of multicollinearity problem. Thus ignoring this statistical problem can lead to wrong conclusion.

After confirming the presence of high relationship between explanatory variables, the principal components was utilized to find the possible linear combination of variables that can produce large variance without much loss of information. The first component contained the variables which were highly related namely SOWN and HAR. Similarly, the second component contained variables QTY, LOAN and IRRI. These original five set of variables were transformed into two variables (principal components) as a linear combination of related variables, but the new variables are independent to each other.

The last step was to assess the efficiency of principal component methods in solving multicollinearity problem. In order to examine the presence of relationship between predictors, dependent variable was regressed on these two principal components. The results show that VIF values for each predictor which indicate that multicollinearity problem was eliminated. Principal components method helps not only in identifying which variables are highly related, but also providing solution for improving results of the estimated coefficients. The method transforms a set of linearly related variables into artificial variable that are not related with each other. If these new variables can be named meaningfully it may be treated as variables for further analysis and considered as a remedial solution to multicollinearity. Regardless of the strength of principal components in removing multicollinearity, its application is limited to a large sample size specifically a minimum of 300 observations (Comrey and Lee, 1992).

In this study, although five factors have been computed, it is obvious that not all five factors will be useful in representing the list of five variables. There were two eigenvalues (2.121, 1.603) greater than 1.0. The first principal component has largest variance that account for 42.4% of the total variance. This PC has comparatively largest eigenvalue of 2.1 which is equivalent to the eigenvalues of two variables. The second PC has an eigenvalue of 1.6 that accounts to 32.1% of the variability of the data. As a rule of thumb, the first two components have eigenvalue greater than one and collectively account to 74.5% of variability of the original data set losing only 25.5% of the information. Therefore only two principal components are extracted and retained from five principal components without much loss of information. The principal component analysis can frequently be used as a solution to multicollinearity.

From the eigenvalues and component number plot, this trend is observed from the four components onwards indicating that an eigenvalues of less than 1, and hence only two components have been retained. The principal component matrix indicates that the component matrix rotated using the Varimax rotation technique which further provides the rotated component matrix.

The explanatory variables' regression coefficients and standard errors are indeterminate if there is perfect collinearity among them. As a result, it is impossible to determine sesame's exact production values. The impacts of multicollinearity are diverse. High variance coefficients can make estimates less accurate, multicollinearity can make coefficients appear to have the wrong sign and estimates of coefficients might vary depending on the sample data used.

5.2 Recommendations

PCR is a regression analysis technique that is based on PCA. Regression is typically thought of as the result of a set of covariates, which are also known as predictors, explanatory variables, or independent variables, based on a traditional linear regression model. Principal component coefficient estimators aim to add as little bias as possible while significantly reducing estimator variances by removing multicollinearities from the least squares estimator.

In constructing multiple linear regression equation, the estimators are estimated by principal component regression method. 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 were added from 1995 to 2020, the variables were even insignificant at 10 % level. Hence, other assumptions are needed to detect.

In this thesis, before the multicollinearity is remedied, the adjusted R-Square value is 0.807 and it means that all the explanatory variables can explain nearly 81% of total variation of the sesame production. And then, after the multicollinearity was remedied, adjusted R-Square value is 0.842 and it means that all the explanatory variables can explain 84.2% of total variation of the sesame production. It has been found that there is a strongly positive relationship between the sesame production of Myanmar with sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds.

Using these additional data, the assumption of multicollinearity was discovered. One of the methods for detecting the multicollinearity assumption was the VIF and tolerance. The VIF values are greater than 10 and the tolerance values were closely to zero. Since these suffered the multicollinearity, Pearson correlation matrix were used to know strong correlation between explanatory variables. The model is significant and R^2 value shows that the model can explain 85.5 % on dependent variable Y.

5.3 Further Studies

This study is analyzed only sesame production of Myanmar but other oilseeds production such as groundnut and sunflower can be studied. Moreover, sesame and other oilseeds can be compared with that of their production and sown acreage, harvested acreage, irrigation, agricultural loan and quality seeds. The correlation between these predictor variables and production of sesame in Myanmar can be analyzed. The principal objective of this study was solution to multicollinearity when fitting linear regression model. Multicollinearity was detected using VIF, tolerance and principal component analysis as solution to the problem was presented. The study indicated that principal component analysis is one of the appropriate methods of solving this matter. Therefore applying principal components produce better estimation and prediction than ordinary least squares when predictors are related. However, the standardization of the variables could not make in this study. For future studies, if variables are measured on scales with widely differing ranges or if the units of measurement are not commensurate, the variables should be standardized in the principal component analysis.

REFERENCES

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.

Alauddin, M. & Nghiem, H. S. (2010). Do instructional attributes pose multicollinearity problems? An empirical exploration. *Economic Analysis and Policy,* 40(3), 351-361.

Alibuhtto, M. C. & Peiris, T. S. G. (2015). Principal component regression for solving multicollinearity problem.

Ayinde, K., Alabi, O. O. & Nwosu, U. I. (2021). Solving Multicollinearity Problem in Linear Regression Model: The Review Suggests New Idea of Partitioning and Extraction of the Explanatory Variables. *Journal of Mathematics and Statistics Studies*, 2(1), 12-20.

Keprt, A. & Snásel, V. (2004). Binary Factor Analysis with Help of Formal Concepts. In *CLA*, 110, 90-101.

Comrey, A. L. & Lee, H. B. (1992). A first course in factor analysis, Second Edition. hillsdale, nj: L.

Chatterjee, S. & Hadi, A. S. (2006). *Regression analysis by example*. John Wiley & Sons.

Favre, R. & Myint, K. (2009). *An analysis of the Myanmar edible oil crops sub-sector*. Rural Infrastructure and Agro-Industries Division, Food and Agriculture Organization of the United Nations.

FAO (2016). Formulation and Operationalization of National Action Plan for Poverty Alleviation and Rural Development through Agricultural (NAPA), Working Paper 1: Crop Production, Extension and Research, United Nations.

Ghorbani, H. (2020). ILL-CONDITIONING IN LINEAR REGRESSION MODELS AND ITS DIAGNOSTICS. *The Pure and Applied Mathematics*, 27(2), 71-81.

Gujarati, D. N. & Damodar, N. (2009). *Basic Econometrics Fifth Edition*. McGraw-Hill.

Gujarati, D. N. (2004). Basic Econometrics Fourth Edition, Tata McGraw-Hill, New Delhi.

Gujarati, D. N. (1995). Econometrics Third Edition, McGraw-Hill, Inc., New York.

Gwelo, A. S. (2019). Principal components to overcome multicollinearity problem. *Oradea Journal of Business and Economics*, 4(1), 79-91.

Herawati, N., Nisa, K., Setiawan, E., Nusyirwan, N. & Tiryono, T. (2018). Regularized multiple regression methods to deal with severe multicollinearity. *International Journal of Statistics and Applications*, 8(4), 167-172.

Hnin Hnin Aung (2014). Comparison of Principal Component Regression and Ridge Regression in Solving Multicollinearity Problem. Thesis Paper, Yangon University of Economics.

Johnson, R. A. & Wichern, D. W. (2002). *Applied multivariate statistical analysis* 5 (8). Upper Saddle River, NJ: Prentice Hall.

Joshi, H., Kulkarni, H. & Deshpande, S. (2012). Multicollinearity Diagnostics in Statistical Modeling and Remedies to deal with it using SAS. *Pharmaceutical Users Software Exchange*, 1, 1-34.

Kaur, S. P. (2013). Variables in research. *Indian Journal of Research and Reports in Medical Sciences*, 3(4), 36-38.

Mason, C. H. & Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of marketing research*, 28(3), 268-280.

May Thu (2019). The Violation for Assumptions of Multiple Regression Model. Thesis Paper, Yangon University of Economics.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Montgomery, D. C. (2001). Design and analysis of experiments. John Wiley & Sons. Inc., New York, 1997, 200-1.

Myers, R. H. (1990). Classical and modern regression with applications. PWS.

Næs, T. & Indahl, U. (1998). A unified description of classical classification methods for multicollinear data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 12(3), 205-220.

Ohnmar Oo (2017). Principal Component Analysis of Urban Population in Myanmar. Thesis Paper, Yangon University of Economics.

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. *IASRI*, New Delhi, 1(1), 58-65.

Perez, L. V. (2017). Principal component analysis to address multicollinearity. *Walla Walla, WA, 99362*.

Raitzer, D. A., Wong, L. C. & Samson, J. N. G. (2015). Myanmar's Agriculture Sector: Unlocking the Potential for Inclusive Growth| ADB Economics Working Paper No. 470, 2015.

Rivera, J. P. R. & Reyes, P. O. (2011). Remittances as avenue for encouraging household entrepreneurial activities. *Journal of International Business Research*, 10(3), 85-113.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), 4-11.

Thu Thu Han Min (2014). Solving Multicollinearity Problem by Principal Component Regression. Thesis Paper, Yangon University of Economics.

Thuzar Linn (2013). Value Chain Analysis of Sesame in Magway Township. Thesis Paper, Yangon University of Economics.

Vatcheva, K. P., Lee, M., McCormick, J. B. & Rahbar, M. H. (2016). Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, 6(2).

Willis, C. E. & Perlack, R. D. (1978). Multicollinearity: effects, symptoms, and remedies. *Journal of the Northeastern Agricultural Economics Council*, 7(1), 55-61.

Zhou, X. P. & Huang, X. C. (2018). Reliability analysis of slopes using UD-based response surface methods combined with LASSO. *Engineering Geology*, 233, 111-123.

Zougmoré, R., Partey, S., Ouédraogo, M., Omitoyin, B., Thomas, T., Ayantunde, A. & Jalloh, A. (2016). Toward climate-smart agriculture in West Africa: a review of climate change impacts, adaptation strategies and policy developments for the livestock, fishery and crop production sectors. *Agriculture & Food Security*, 5(1), 1-16.

APPENDIX A

Year	PRODUCTION (in thousand ton)	SOWN ACREAGE (in thousand acre)	HARVESTED ACREAGE (in thousand acre)	IRRIGATION (in thousand acre)	Agricultural LOAN (in kyat millions)	Use of QUALITY SEEDS (basket)
1994-1995	299.3	3288	2797	175.97	96.27	1256
1995-1996	298.8	3153	2234	264.32	850.89	5091
1996-1997	340.3	2830	2746	160.75	851.11	4464
1997-1998	258.7	2557	1789	147.76	315.55	1858
1998-1999	260	2963	1521	285.57	437.71	228
1999-2000	253.2	3352	2381	293.94	523.18	750
2000-2001	375.8	3517	3064	232.19	524.83	660
2001-2002	339	3416	2865	236.94	590.64	68
2002-2003	405.9	3501	3130	184.98	569.72	306
2003-2004	436.2	3619	3281	208.18	963.22	230
2004-2005	473.9	3696	3306	246.46	1241.14	209
2005-2006	438.5	3306	2934	188.24	1562.13	154
2006-2007	680	3565	3378	186.08	1945.74	769
2007-2008	768	3725	3536	194.04	3360.32	398
2008-2009	840	3880	3685	176.83	3581.21	252
2009-2010	854	4038	3863	184.48	5665.1	899
2010-2011	787.4	3918	3754	178.61	10059.67	484
2011-2012	832.1	3941	3785	160.78	11400.85	2150
2012-2013	794.6	3838	3688	156.73	11906.29	443
2013-2014	817.1	4007	3767	209.92	24962.32	825
2014-2015	801.6	3906	3600	281.61	25205.4	893
2015-2016	828	4052	3732	285.66	22639.9	1186
2016-2017	813	4042	3694	335.49	23320.86	1644
2017-2018	764	3685	3652	260.57	58018.6	2000
2018-2019	630	3544	3416	300.4	113775.9	2155
2019-2020	645	3441	3399	330.22	123518.1	1714

Source: Statistical Yearbooks and Myanmar Agricultural Statistics

APPENDIX B

Appendix (B1)

Statistics

		Prod	Sown	Harv	Irri	Loan	Qty
N	Valid	26	26	26	26	26	26
	Missing	0	0	0	0	0	0
Mean		578.246	3568.46	3192.19	225.6431	17226.4096	1195.62
Std. Deviation		229.0904	390.979	632.796	56.94184	32623.93278	1246.126
Minimum		253.2	2557	1521	147.76	96.27	68
Maximum		854.0	4052	3863	335.49	123518.10	5091

Appendix (B2)

Correlation Matrix

		Sown	Harv	Irri	Loan	Qty
Correlation	Sown	1.000	.901	.124	.134	-.360
	Harv	.901	1.000	-.044	.287	-.216
	Irri	.124	-.044	1.000	.554	.094
	Loan	.134	.287	.554	1.000	.197
	Qty	-.360	-.216	.094	.197	1.000

Appendix (B3)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1109602.488	5	221920.498	21.923	.000[b]
	Residual	202458.137	20	10122.907		
	Total	1312060.625	25			

a. Dependent Variable: Prod

b. Predictors: (Constant), Loan, Sown, Qty, Irri, Harv

Appendix (B4)

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	T	Sig.	Tolerance	VIF
1	(Constant)	-1084.612	287.309		-3.775	.001		
	Sown	.512	.196	.875	2.620	.016	.069	14.441
	Harv	.009	.124	.025	.072	.944	.065	15.285
	Irri	-1.152	.652	-.286	-1.768	.092	.294	3.402
	Qty	.023	.019	.126	1.236	.231	.743	1.346
	Loan	.002	.001	.313	1.954	.065	.300	3.330

a. Dependent Variable: Prod

Appendix (B5)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.920[a]	.846	.807	100.6127	.790

a. Predictors: (Constant), Loan, Sown, Qty, Irri, Harv

b. Dependent Variable: Prod

Appendix (B6)

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.121	42.428	42.428	2.121	42.428	42.428	2.075	41.500	41.500
2	1.603	32.052	74.479	1.603	32.052	74.479	1.649	32.980	74.479
3	.817	16.343	90.823						
4	.429	8.574	99.397						
5	.030	.603	100.000						

Extraction Method: Principal Component Analysis.

Appendix (B7)

Rotated Component Matrix[a]

	Component	
	1	2
Sown	.955	.114
Harv	.925	.151
Qty	-.536	.395
Loan	.139	.882
Irri	-.016	.824

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 3 iterations.

Appendix (B8)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.925[a]	.855	.842	90.1017	.689

a. Predictors: (Constant), Z2, Z1

b. Dependent Variable: PROD

Appendix (B9)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1099309.906	2	549654.953	67.706	.000[b]
	Residual	186721.098	23	8118.309		
	Total	1286031.005	25			

a. Dependent Variable: PROD

b. Predictors: (Constant), Z2, Z1

Appendix (B10)

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-817.814	125.103		-6.537	.000		
	Z1	.215	.020	.886	10.856	.000	.948	1.055
	Z2	.001	.001	.131	1.608	.121	.948	1.055

a. Dependent Variable: PROD