HANDLING OF MULTICOLLINEARITY PROBLEM IN MODELLING NON-PERFORMING LOANS IN AFRICA'S PORTFOLIO DATA

by

Malebo Tshegofatso Molebatsi

Submitted in accordance with the requirements for the degree of

Master of Science

in the subject of

Statistics

at the UNIVERSITY OF SOUTH AFRICA

Supervisor: Professor Peter M. Njuho

January 2023

DECLARATION

Name: Malebo T Molebatsi

Student No.: 61113697

Degree: Masters in Statistics

Thesis Title: HANDLING OF MULTICOLLINEARITY PROBLEM IN MODELLING NON-PERFORMING LOANS IN AFRICA'S PORTFOLIO DATA

I declare that the above thesis is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Signature mtmolebatsi Date 25 January 2023

ACKNOWLEDGMENT

I wish to express my deepest gratitude to my supervisor Professor Peter M. Njuho, without your persistent guidance and encouragement this would not have materialised. Thank you to my former statistics lecturers, Dr Gretel Crafford and Dr Sollie Millard, who ignited my passion for the subject. Thank you to my mother Nkomeng M. Molebatsi, my brothers Kaelo H. Molebatsi and Thuto Molebatsi for being there throughout this process. Special thanks to my sister Nonjabulo Dladla for all the assistance and support. Last but not least, thank you to my broader family and friends for always being there to support and encourage me.

Abstract

Non-performing loans (NPLs) are detrimental to profits in the banking sector. Predicting the level of NPLs using macroeconomic variables is vital in order to build mitigating actions for such scenarios to safeguard the profitability of the institution. Macroeconomic variables are susceptible to high correlations amongst each other, bringing about the problem of multicollinearity. Predicting in the presence of multicollinearity brings about unreliable and inefficient results. This study aims to find an optimal and efficient way of forecasting NPLs using Ordinary Least Squares (OLS), Ridge Regression (RR) and Principal Component Analysis (PCA) while correcting for multicollinearity. To do this, NPL data from bank X was attained, along with multiple macroeconomic variables, specifically for Kenya and Nigeria. It is critical to assess the determinants of NPLs so that effective and efficient policies can be deployed to prevent the rising trajectory of NPLs. To minimize the risks of using expert judgement, it is necessary to consider effective statistical methods for predicting NPLs. The benefits accrued from such methods include (1) minimum collection costs incurred when a loan defaults, such as less phone calls urging the customers to pay, less litigation costs when trying to recover the assets, less shortfalls incurred when disposing off the assets that have been repossessed and less auction sales if the assets have to be auctioned, to mention a few; (2) correct pricing for the risk; (3) be able to differentiate between high-risk and low-risk accounts based on the macroeconomic factors; and (4) be more prudent in granting credit to minimize losses and maximise profits. This study considers the OLS, RR and PCA in modeling the NPLs data from bank X. The results showed that multicollinearity exists for most variables. Some of the variables did not conform to the assumptions of the OLS. The models for OLS for both countries were significant, while some of the variables displayed illogical outcomes, possibly due to multicollinearity among the predictor variables. RR method solved for multicollinearity and had a relatively predictive power for Nigeria data and not Kenya, whereas PCA solved for

multicollinearity and introduced a positive factor in data reduction and had a relatively better predictive power. The mean square errors (MSEs) for PCA and RR were lower than that of OLS. A key limitation was inadequate data from the banking sector due to sensitivity issue. We conclude that the data can be expanded, and the number of variables reduced so that prediction can be more precise. Further work using other methods such as GARCH can be explored to improve the prediction of the NPLs in the midst of multicollinearity.

Key words: non-performing loans, Financial institutions profitability, Macroeconomic variables, Multicollinearity, Ordinary Least Squares, Ridge Regression, Principal Component Analysis.

1 Introduction 13

1.1 Background 13

1.2 Motivation 15

1.3 Justification 16

1.4 Purpose 16

1.5 Statement of the problem 17

1.6 Objectives 18

2 Background 19

2.1 Macroeconomic influences on non-performing loans 19

2.2 Ridge regression in tackling the multicollinearity problem 26

2.2.1 Choosing the ridge regression parameter K 26

2.2.2 Advantages of the ridge regression over the ordinary least squares 31

2.2.3 Limitations of ridge regression 32

2.3 The principal component regression in tackling the multicollinearity problem 37

2.3.1 Link between principal component analysis and factor analysis 37

2.3.2 Determining the number of components 39

2.3.3 Superiority of principal component regression in relation to other regression models 41

2.3.4 Confines of principal component regression 44

2.4 Other Statistical approaches to solve multicollinearity 45

3 Methodology 49

3.1 Model fitting procedure 49

3.1.1 Variable selection 49

3.1.2 Regression model comparison 50

3.2 Assessing the model competences 59

3.3 Out-of-time Testing 60

4 Data Exploration 61

4.1 Data Structure 61

4.2 Data Management 63

4.3 Issues and types of multicollinearity 64

4.3.1 Diagnosis for Multicollinearity 66

4.4 Analysis of the Nigeria portfolio data 69

4.4.1 Data exploration – Nigeria portfolio data 69

4.4.2 Checking the validity of the assumptions 72

4.4.3 Testing for Correlations – Pearson Correlation Coefficient 74

4.4.4 Ordinary Least Squares using Nigeria portfolio data 76

4.4.5 Variable selection (post multicollinearity) – Nigeria portfolio data 83

4.4.6 Ridge Regression – Nigeria portfolio data 86

4.4.7 Principal Component Analysis – Nigeria portfolio data 90

4.5 Analysis of Kenya portfolio data 96

4.5.1 Descriptive statistics – Kenya portfolio data 96

4.5.2 Testing for Assumptions – Kenya portfolio data 98

4.5.3 Testing for Correlations – Pearson Correlation Coefficient: Kenya portfolio data 101

4.5.4 Ordinary Least Squares – Kenya portfolio data 102

4.5.5 Variable Selection (Post Multicollinearity) – Kenya portfolio data 107

4.5.6 Ridge Regression – Kenya portfolio data 111

4.5.7 Principal Component Analysis – Kenya portfolio data 116

5 Discussion 122

5.1 Case Study – Nigeria portfolio data 122

5.1.1 Descriptive Analysis Discussion – Nigeria portfolio data.....122
5.1.2 Comparative Assessment of Models – Nigeria portfolio data.....123
5.2 Case Study – Kenya portfolio data.....128
5.2.1 Descriptive Analysis Discussion – Kenya portfolio data.....128
5.2.2 Comparative Assessment of Models – Kenya portfolio data.....130
6 Conclusions.....135
BIBLIOGRAPHY.....137
ANNEXURE.....146
Annexure A – SAS Program: Nigeria.....146
Annexure B – SAS Program: Kenya.....151

List of Figures

Figure 2.1 Trace plot in estimation of *K* 28

Figure 2.2 The schematic relationship between the PCA and the FA 39

Figure 2.3 A scree plot of eigenvalues generated from a correlation matrix 40

Figure 2.4 Plot of the proportion of variability accounted for by each principal component .. 41

Figure 3.1 Steps in standardisation of variables 54

Figure 3.2 Demonstrating bias in ridge regression model 56

Figure 4.1 Relationship between perfect and less than perfect multicollinearity 65

Figure 4.2 Boxplots of the variables in the Nigeria portfolio data 70

Figure 4.3 Residual plot of the fitted OLS model for the Nigeria portfolio 81

Figure 4.4 Residual by the individual independent variables of the fitted OLS model for the Nigeria portfolio 82

Figure 4.5 Residual plot of the OLS post variable selection of the Nigeria portfolio data 85

Figure 4.6 Residual plots by the individual independent variables for Nigeria portfolio 86

Figure 4.7 Ridge Trace a – for Nigeria portfolio 87

Figure 4.8 Ridge Trace b – for Nigeria portfolio 87

Figure 4.9 Residual plot of fitted ridge regression for Nigeria portfolio 89

Figure 4.10 Residual plots of individual independent variables for fitted ridge regression for Nigeria portfolio 90

Figure 4.11 Scree Plot for Nigeria portfolio 92

Figure 4.12 Residuals of the predicted values for the 4PC for Nigeria portfolio 95

Figure 4.13 Residuals of the predicted values for the individual principal components for Nigeria portfolio 95

Figure 4.14 Box plot of the variables in the Kenya portfolio data 97

Figure 4.15 Residual plot of the fitted OLS model for the Kenya portfolio 106

Figure 4.16 Residual plot of the fitted OLS model for the individual independent variables for Kenya portfolio 107
Figure 4.17 Residual plot for the OLS post variable selection for Kenya portfolio 110
Figure 4.18 Residual plot for the OLS post variable selection for individual regressors for Kenya portfolio 110
Figure 4.19 Ridge Trace a – for Kenya portfolio 112
Figure 4.20 Ridge Trace b – for Kenya portfolio 112
Figure 4.21 Residual plot for the fitted ridge regression predicted values for Kenya portfolio 115
Figure 4.22 Residual plot for the fitted ridge regression for the individual regressors for Kenya portfolio 116
Figure 4.23 Scree plot for Kenya portfolio 117
Figure 4.24 Residual plots of the principal components regression residuals against the predicted values for Kenya portfolio 120
Figure 4.25 Residual plots of the principal components regression residuals against PCs for Kenya portfolio 121
Figure 5.1 Forecasted NPLs using Ordinary Least Squares (Fitted reduced model) on Nigeria portfolio data 125
Figure 5.2 Forecasted NPLs using Ridge regression (Fitted reduced model) on Nigeria portfolio data 126
Figure 5.3 GDP Rates in Nigeria 127
Figure 5.4 Forecasted NPLs using Ordinary Least Squares (Fitted model) on Kenya portfolio data 131
Figure 5.5 Forecasted NPLs using Ridge Regression (Fitted Model) on Kenya portfolio data 133

List of Tables

Table 4.1 Summary statistics of non-performing loans' data for Nigeria 69

Table 4.2 Variability, skewness and presence of outliers as displayed in the boxplot for each variable.....71

Table 4.3 The Durbin-Watson, stationarity and normality test using Nigeria portfolio data ..72

Table 4.4 Testing for Correlations – Nigeria portfolio data 75

Table 4.5 Ordinary least squares output – Nigeria portfolio data.....77

Table 4.6 Ordinary least squares output – Nigeria portfolio data: fitted model 79

Table 4.7 Ordinary Least Squares post variable selection – Nigeria portfolio data 83

Table 4.8 Ordinary Least Squares post variable selection – Nigeria portfolio data: fitted model 84

Table 4.9 Ridge Regression output – Nigeria portfolio data: fitted model.....88

Table 4.10 Eigenvalues of the Correlation Matrix – Nigeria portfolio data.....91

Table 4.11 Eigenvectors of the Nigeria portfolio data.....92

Table 4.12 Principal Component Regression results – Nigeria portfolio data 94

Table 4.13 Summary statistics on Kenya non-performing loans data 96

Table 4.14 Summary table for skewness and outliers – Kenya portfolio data 98

Table 4.15 Auto-correlation, Stationarity, and normality test 99

Table 4.16 Testing for correlations – Kenya portfolio data.....101

Table 4.17 Ordinary Least Squares output – Kenya portfolio data 103

Table 4.18 Ordinary least squares output – Kenya portfolio data: fitted model.....105

Table 4.19 Ordinary Least Squares output post variable selection – Kenya portfolio data...108

Table 4.20 Ordinary Least Squares post variable selection – Kenya portfolio data: Fitted model 109

Table 4.21 Ridge regression output – Kenya portfolio data 113

Table 4.22 Ridge Regression output – Kenya portfolio data - fitted model 114

Table 4.23 Eigenvalues of the correlation matrix – Kenya portfolio data.....117

Table 4.24 Eigenvectors of the correlation matrix on the Kenya portfolio data.....118

Table 4.25 Principal component regression output – Kenya portfolio data 119

CHAPTER 1

MOTIVATION, JUSTIFICATION, PURPOSE, AND RESEARCH PROBLEM

1 Introduction

1.1 Background

In the banking sector, credit risk accounts for approximately 80% of the risk thereof. Credit risk in financial institutions is defined as the risk of loss when a bank considers that the obligor is unlikely to pay its credit obligations in full or the obligor is more than 90 days past due on any material credit obligation (Risk and Capital Management Report, 2016). The risk typically increases when a loan is 90 days overdue without payment of due instalment. This situation is termed a non-performing loan (NPL). In addition, a loan can be classified as non-performing in the presence of compelling evidence such as cash flow generation challenges and the high likelihood of inability to fulfil the repayment obligations.

The losses incurred within a bank emanate mostly from non-performing loans. From January 2018, a new accounting principle that deals with an expected loss model (International Financial Reporting Standard – IFRS 9) rather than an incurred loss model (International Accounting Standard – IAS 39) has been enforced across all banks in accordance with the International Accounting Standards Board. The IAS 39 focuses only on an incurred loss approach for financial assets. The disadvantage with this approach is that the loss is only recognised when a trigger event occurs. This is deemed to be a ‘too little, too late’ recognition of loan losses within the bank. The IFRS 9 can incorporate a ‘forward-looking model’, where a potential NPL outlook based on the economic forecasts is taken into consideration in determining the losses within the bank.

The economic data used in a forward-looking model is characterised by linearly correlated variables which cause a multicollinearity problem. The multicollinearity problem affects the determinants of the credit risk. A forward-looking model that disregards or ignores the multicollinearity problem fails to predict the NPL ratio accurately. The model violates one of the fundamental principles of linear regression, which requires taking caution on the use of linearly correlated independent variables. For the optimum use of regression analysis, Gujarati (2003) outlines the ten assumptions regarding the independent variables and error terms that are critical for the valid interpretation of regression estimates. One of these assumptions state that there is no perfect multicollinearity amongst the independent variables. This assumption underpins the rationale behind this study. Violation of the assumption leads to regression estimates that are unstable and with low precision. The estimates attain the wrong signs which affect the variable selection process, thus leaving out variables that could be otherwise important (Kumari, 2008).

Multicollinearity, a linear dependency that exists between the explanatory variables can either be perfect or near perfect. The regression estimates are indeterminate and possess infinite standard errors when perfect multicollinearity exist. For near-perfect multicollinearity, the regression coefficients possess large standard errors and cannot be estimated with great precision (Gujarati, 2003). The existing remedial measures in addressing the multicollinearity problem, such as ridge regression and principal component regression techniques, are unpacked including some trade-offs.

In solving the multicollinearity problem, the ridge regression technique operates by adding some degree of biasness to the regression estimates, which leads to smaller standard errors (Duzan and Shariff, 2016). The principal component regression analysis reduces large number

of highly correlated variables to fewer uncorrelated variables (Herawati *et al*., 2018**)**. Each generated variable is a linear combination of the original variables (Freund and Wilson, 2006).

The applied techniques in addressing multicollinearity problem are illustrated using forecasting data on macroeconomics relating to the NPL ratios for IFRS 9. Two data sets from Kenya and Nigeria spanning from 2006 to 2018 are used. The two countries were chosen based on the complexity surrounding their portfolios. Kenya has had interest rate caps, while Nigeria has had a huge drop in oil sovereign revenues, that might be worthwhile to investigate the impact this has had on the banking sector.

1.2 Motivation

Time series data such as macroeconomic indicators are prone to having the problem of multicollinearity, as illustrated by Kumari (2008). The linear dependency between the variables make it impossible to assess the unique influence of each independent variable on the dependent factor, while holding other variables constant. The inherent risk of modelling data in the presence of multicollinearity brings about several challenges. These include wide confidence intervals due to inflated standard errors. This leads to not rejecting the null hypothesis (Gujarati 2003). The presence of multicollinearity in predictive models affects the parameter estimates that are overly sensitive to small changes in the sample. The consequences are insignificant results and tripling effects, leading to wrong conclusions (Kumari, 2008). In the case of credit risk, misinterpreting and making wrong conclusions could have detrimental effects on the organization's profits and ultimately the return on equity for the investors. It is imperative to remove any hindrance (such as the multicollinearity problem) that can affect the validity of the inferences drawn on credit risk data. This study demonstrate how to address the

problem of multicollinearity through the three suggested models namely, Ridge regression, Ordinary Least Squares and Principal Component Analysis.

1.3 Justification

Comprehensive understanding of the multicollinearity problem and how it can be solved can alleviate the consequences brought about by this conundrum. Chatterjee and Hadi (2006) points out that multicollinearity is a condition of deficient data and not a modelling error. It is the user's responsibility to find better methods that can be used to predict the dependent variable and be able to get marginal effects of each independent variable in data that is deficient. The option of not dealing with a deficient that is characterized by multicollinearity by re-collecting needs to be assessed, even though it may be time-consuming and costly to accomplish. A large spectrum of regression techniques that deal with multicollinearity exists. There is a need to explore such techniques, including adjustments on the estimates instead of collecting new data with a hope that the multicollinearity problem will not occur. The review of the techniques dealing with the multicollinearity problem lead to understanding them well and allows for adjustments that result to stable estimates and appropriate conclusions. This study aims to offer a variety of techniques to handle the multicollinearity problem associated with the determinants of credit risk in forecasting the NPL in the Africa credit portfolio space. In addition, the analysis creates awareness of the key macroeconomic factors that are likely to affect the customers' affordability, thereby increasing their propensity to default.

1.4 Purpose

The intention is to assess the ridge regression, ordinary least squares, and principal component analysis in addressing the multicollinearity problem associated with the determinants of the credit risk analysis.

1.5 Statement of the problem

With the inception of the expected loss model (IFRS 9), currently there exists no statistical model that bank X uses to forecast the NPL ratio in the Africa region entities (name of Bank kept anonymous due to data privacy compliance). In the interim, expert judgement is being used across the region due to the problems encountered with the ordinary least squares (OLS) method. The fundamental principle of IFRS 9 is the incorporation of the macroeconomic forecasting model, where the intention is to provide the ability to forecast the NPL ratio based on country-specific macroeconomic factors (The International Accounting Standards Board, 2014). The usage of macroeconomic variables observed in time series involve some linearly related variables, hence giving rise to multicollinearity and its associated consequences (Kumari, 2008). The presence of multicollinearity would not be a problem if the scope was only to forecast the NPL ratio, but it would be if the scope is extended to analysing the marginal effects of each independent variable. Marginal analysis seeks to understand the precise effect of each variable to the dependent variable.

The NPL ratio forecast, particularly in volatile macroeconomic environments, is vital in credit risk because it minimizes the losses incurred. It does so by offering insights into the type of macroeconomic factor that leads to high propensity of defaults. Using expert judgement in forecasting the NPL ratio is not ideal due to the volatility of macroeconomic indicators. To minimize the risks of using expert judgement, it is necessary to consider effective statistical methods for predicting NPLs. The benefits accrued from such methods include (1) minimum collection costs incurred when a loan defaults, (2) correct pricing for the risk, (3) being able to differentiate between high-risk and low-risk accounts based on the macroeconomic factors, and (4) being more prudent in granting credit to minimize losses and maximise profits.

1.6 Objectives

Using the Africa regional data from Kenya and Nigeria, ridge regression and principal component regression techniques were applied in handling the multicollinearity problem to predict the NPL ratio. The performance of these techniques is compared to that of ordinary least squares in predicting the NPL ratio. The review of these techniques is based on their forecasting power, usefulness, effectiveness, efficiency, and the associated trade-offs encountered in solving the multicollinearity problem. The statistical techniques are illustrated using the quarterly macroeconomic data from 2006 to 2016, together with consolidated NPL ratios from Kenya and Nigeria. We effectively predict the NPL ratio for the expected loss model.

Overall objective

To improve the prediction of the NPL ratio by addressing the multicollinearity problem encountered in credit portfolio data using the ridge regression and principal component regression statistical techniques.

The specific objectives

1) To assess the performance of ridge regression and principal component regression in handling the multicollinearity problem in macroeconomic data.
2) To assess the robustness of ridge regression and principal component regression relative to ordinary least squares in relation to the multicollinearity problem.
3) To investigate the predicting power of the ordinary least squares, ridge regression and principal component regression on the non-performing loan ratio using 2016-2018 as the out-of-time testing.

CHAPTER 2

LITERATURE REVIEW

2 Background

The assessment of separate influences of a predictor variable on a dependent variable, while keeping others constant in multiple regression, is solely dependent on the assumption of orthogonality among the predictor variables. The lack of orthogonality is most common in regression applications, resulting in ambiguous regression outcomes (Chatterjee and Hadi, 2006). The concept of multicollinearity has gained more coverage in literature to have meaningful analysis, as illustrated by the International Financial Reporting Standard 9 (IFRS 9), that is the forward-looking model that incorporates the prediction of the level of NPLs using macroeconomic variables.

2.1 Macroeconomic influences on non-performing loans

Credit risk accounts for at least 80% of risk undertaken by banks. Due to the size of the risk associated with credit, it is imperative that this risk be minimized accordingly so that profits can be realised. Credit risk is largely driven by the performance of the portfolio, hence Ugoani (2016) reiterates that NPLs erode the ability of banks to make profits. The NPLs can either be secured or unsecured, where secured loans generally attract lesser losses than unsecured loans. Secured loans are backed by collateral or security that the bank can sell or dispose to settle the remaining debt, while unsecured loans are not backed by any asset (Standard Bank Glossary). The performance of NPLs is either caused by internal aspects within the bank, or external influences that the bank management has no control over. One of the core objectives of the IFRS 9 is to be able to predict the level of NPLs using macroeconomic variables with losses anticipated and mitigating actions put in place.

In the Kenya perspective, the causes of NPLs revolve around the unfavourable economic environment (Waweru, 2009). Internal factors such as the lack of credit risk evaluation officials, insider lending and high interest rates charged have a significant impact, although the impaired loans may be largely driven by macroeconomic factors. The involvement of bank management, particularly in government policy formulation and strategy concerning economic matters is recommended as a mitigating action for the prevention of high NPLs (Waweru 2009). In Kenya, a law was passed in 2016 to cap the interest rates on lending at 4% above the Central Bank Rate, and on the deposits at 70% of the base set (Muriuki, Mathuva and Egondi, 2017). According to Erickson (2018), the interest rate capping hindered credit growth and ultimately the banks' profitability. To cope in such an environment, banks had to retrench some of their staff to reduce operational costs. Erickson (2018) further mentions the Nicaragua and Ecuador example as empirical evidence of the negative effect of interest capping and the associated resultant effect of sluggish credit growth. Some of the recommendations to mitigate against interest capping include introducing tax incentives to encourage savings, credit literacy and incorporating the informal sector into the financial sector (Erickson, 2018).

The effect of macroeconomic variables on the performance of commercial banks in Kenya drew inconclusive results. The relationship between the gross domestic product (GDP) and the return on equity (ROE) exhibited positive correlation but was not significant. For the period 2001 to 2010, inflation was negatively correlated with banks' profitability (Ongore and Kusa, 2013). The 2008 financial crisis that affected the financial sector across the world fell within this period. The profitability of banks in Kenya was affected by factors that were under the control of managers (internal factors, rather than external factors) such as management efficiency and capital adequacy, based on the regression results.

In an IMF working paper titled ‘The Impact of Oil Prices on the Banking System in the GCC’, 42 banks were sampled to show that oil prices and economic activity affected banks’ asset quality. For the gulf cooperation countries (GCC) economies, the NPL ratios increased due to the decline in oil prices and the sluggish economic growth. As a mitigating action, the GCC banks have subsequently set their capital ratios and provisioning levels accordingly to align with business and financial cycles (Khamdelwal, Miyajima and Santos, 2016).

According to Skenderi, Islami and Mulolli (2016), the NPLs ratio increased between 2010 and 2014 in Kosovo because of the influence of macroeconomic variables. Some of the macroeconomic factors investigated included interest rates, nominal gross domestic product (GDP), nominal inflation, maturity, unemployment rate, and exchange rates to mention a few. The impact of these factors on the NPL ratio can either be good or bad depending on the angle one looks at. For instance, in a high inflation environment, the real value of the debt of the borrower can be reduced – which is positive – while on the negative side, the real disposable income of the borrower would decrease, rendering the borrower unable to fulfil their obligations (Skenderi *et al*., 2016; Touny and Shebab, 2015). The researchers conclude by making recommendations to the Central Bank of Kosovo (CBK). The CBK can influence the direction of the economy through monetary policy that uses interest rate variability to impart the changes needed. The primary responsibility of the CBK is to stabilise prices so that inflation and interest rates do not increase. Similarly, recommendations to the government were proposed so that fiscal policies that encompasses economic and financial strategies could be used to drive sound economic growth. The government was further urged to deploy strategies to stimulate business development and offer subsidies where necessary. For the commercial banks in Kosovo, robust credit policies such as extension on the period of the maturity of the

loans and grace periods were recommended as they are likely to have an impact on the reduction of NPL ratios.

There exists a significant positive relationship between interest rates and high NPLs because high loan defaults cause asset erosion and ultimately capital erosion (Farhan, Sattar, Chaudhry and Khalil, 2012). Higher interest rates are likely to trigger increases in the cost of credit, thus making it difficult for borrowers to pay their debts, hence the enlarged NPL ratios. For independent variables such as GDP, the regression analysis discloses a negative relationship with the NPL ratios. For instance, if GDP increases, the NPL ratio is likely to reduce (Skenderi *et al*., 2016).

In the context of Namibia, Sheefeni (2015) explored the effect of macroeconomic factors on the impaired loans portfolio. This study was two-fold as it incorporated the short- and long-run effects. Using the time series econometric techniques like co-integration, the study revealed that in the long run, log of GDP, interest rate and inflation were found to have a significant effect on NPLs, but not necessarily in the short run (Sheefeni, 2015). With several other employed techniques such as granger causality and impulse response function, all the results substantiate the importance of paying attention to the macroeconomic environment so that the impact on the levels of impaired loans can be minimized.

Similarly, Rulyasri, Achsani and Mulyati (2017) determined the impact of macroeconomic factors on NPLs in Indonesia, particularly for small and medium enterprises (SMEs). The methodology also mirrors that of Namibia in assessing the long- and short-run effects of variables such as GDP, currency exchange rate, consumer price index (CPI) and the total of money circulation (M2). In the long run, GDP and CPI have a negative influence against NPL

growth on the retail segment. Exchange rate and M2, on the other hand, exhibited positive relationships with NPLs (Rulyasri *et al*., 2017). This is in line with initial assumptions that support the analysis whereby, purchasing power increases when there is more money circulating in the economy. The downside of having higher purchasing power is that it can trigger a rise in prices and ultimately increase inflation. On the other hand, M2 is found to be the only variable that has a significant positive effect on NPLs (Rulyasri *et al*., 2017). The amount of money in circulation must be diligently and cautiously monitored to prevent the ripple effect of this in economic cycles.

According to Adeola and Ikpesu (2016), macroeconomic factors have an impact on the level of impaired loans. The common macroeconomic variables (GDP, M2, unemployment, inflation, lending rate and exchange rate) have a significant impact as reflected by the high coefficient of determination. The study does not assess the impact of multicollinearity between the explanatory variables. It fails to articulate the regression coefficients of each variable through a forecasting model so that robust models can be built to cushion the impact of these variables in the future.

Using the error correction methodology, Tyona, Tyohemba and Eya (2017) examined the determinants of the impaired loans in Nigeria. The conclusion from the analysis of the macroeconomic variables such as GDP, inflation and money supply are based on two scenarios: short and long run. In the short run, GDP and inflation are negatively related to NPLs, while money supply shows a positive relationship (Tyona *et al*., 2017). In contrary to the priori inferences and the dissimilarity to Rulyasri *et al*. (2017), the co-integration equation results from Tyona *et al*. (2017) confirms that money supply converges to the negative relationship with NPLs in the long run.

In any country, a sound financial sector is needed to drive economic growth as well as attract foreign investment, but if NPL ratios are high, there arises a need to understand the root causes as the inherent risk effect can hamper economic growth. The ripple effect of high NPLs can reduce credit flow into the country, which ultimately affects the efficiency and productivity of business (Farhan *et al*., 2012). The NPLs are not only a problem in conventional banking, but also for Islamic banking. Reference is made to the research done in the Islamic banking sector in Malaysia, where interest rates were found to have a positive significant relationship with NPLs, while the producer price index exhibited a significant negative relationship with NPLs. In the case of Pakistan, the crisis in the energy sector (the biggest driver of government revenues) such as load-shedding of gas and electricity, high costs associated with energy have resulted in several industry closures. The inability of the consumers to service their debt led to an increase in the level of NPLs. Other variables such as unemployment, inflation and exchange rate were found to have a positive relationship with NPLs, while GDP was found to have a negative relationship. It is critical to assess the determinants of NPLs so that effective and efficient policies can be deployed to prevent the rising trajectory of NPLs. The study does not assess the impact of the high correlations observed between the unemployment rate and interest rate (80%), the energy crisis and the interest rate (91%), as well as unemployment and the energy crisis (71%), as these epitomise the problem of multicollinearity. Furthermore, one could also explore the associated effects that arise from modelling data with such a problem.

The theme of non-performing loans has gained coverage in recent periods, particularly after the 2008 global financial crisis (Touny and Shebab, 2015). According to Messai and Jouini (2013), minimisation of NPLs is a necessity in terms of bank profitability and ultimately the growth of the economy. In their study, the authors consider the impact of GDP and the unemployment rate on the level of NPLs. The sample taken to assess this conundrum included

Greece, Spain and Italy. The rationale behind choosing these countries was based on respective characteristics, such as worsening public finances and subprime mortgage crisis post the 2008 global financial crisis. The analysis further supports the negative relationship of GDP with non-performing loans, while the unemployment rate has a positive relationship with NPLs. Banks that are generally profitable are less likely to grant loans to high-risk counterparties, while banks that are considered inefficient are likely to grant credit facilities to high risk clients and ultimately incur high levels of NPLs (Messai and Jouini, 2013). The research acknowledges the correlation between GDP and the unemployment rate at 51%, hence a test for multicollinearity is conducted using variance inflation factors (VIF). The VIF are determined to be less than 4, indicating that there is no problem of multicollinearity. In contrast, the study does not address the impact of other macroeconomic factors such as inflation and exchange rate on NPLs. The other factor that could have been addressed entails the effect of macroeconomic shocks on the impaired loans and the resilience of banks using stress testing models.

The presence of multicollinearity in macroeconomic and market-related variables are detrimental to the predictive power of financial models (Kumari 2008). It is therefore paramount that the root cause of the problem be assessed and removed to remain with the best fit model. Kumari (2008) further explains that multicollinearity is a sample phenomenon and not a population issue; hence, multicollinearity is classified as a measurement of the degree of this problem. Adeboye, Fagoyinbo and Olatayo (2014) agree that multicollinearity is a sample issue, and the degree to which it matters is subjective. For this purpose, it is the researchers' prerogative to assess how harmful the relationship between the explanatory variables is. This then forms the basis of where 'to draw the line' when assessing how harmful multicollinearity

really is. The two solutions being proposed for this conundrum are ridge regression and principal component regression.

2.2 Ridge regression in tackling the multicollinearity problem

Ridge regression is a commonly applied method that was introduced to combat the problem of multicollinearity using biased estimators by allowing modification on the compilation of regression coefficients, as outlined in various literature such as Dorugade (2014). This method assists in alleviating the consequences of multicollinearity without having to increase the sample size, improve the quality of the sample size or eliminate variables from the model (Garcia, Salmeron, Garcia and Martin, 2016).

2.2.1 Choosing the ridge regression parameter K

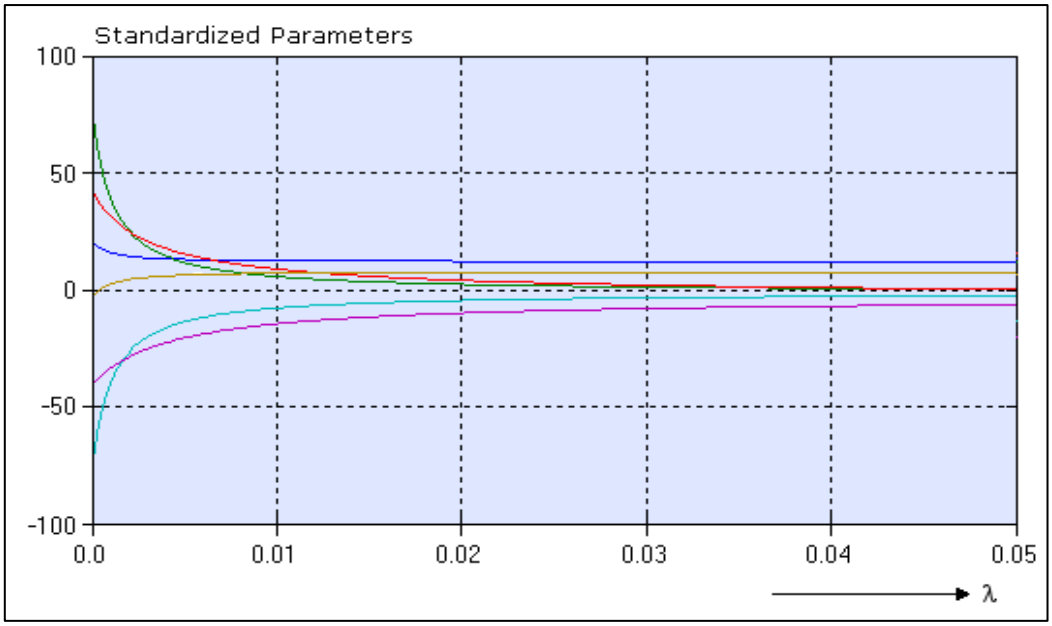
The difference between the ordinary least squares (OLS) estimate and the ridge regression (RR) is in the addition of a small number K in the main diagonal elements of the matrix $\mathbf{X}$ such that the regression coefficients are determined by $\widehat{\boldsymbol{\beta}}(K) = (\mathbf{X}'\mathbf{X} + K\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$, where $K > 0$, $\widehat{\boldsymbol{\beta}}$ is a p x 1 vector of the regression coefficients, $\mathbf{X}$ is a n x p design matrix of the explanatory variables, $\mathbf{I}$ is a p x p identity matrix, and $\mathbf{Y}$ is a n x 1 data vector of the dependent variables. The K almost dissipates the linear association between the explanatory variables. If $K = 0$, then the estimation of the ridge regression coefficients converges to that of OLS. Additionally, $K > 0$ is not a single solution, but rather a variety of solutions, as mentioned by El-Denery and Rashwan (2011).

A simpler mathematical equation of choosing the *K*, suggested by Bager *et al.,* (2017) and Gorgees (2017) has the form:

$$K = \frac{p\hat{\sigma}^2}{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}}$$

where $\hat{\sigma}^2$ and $\widehat{\boldsymbol{\beta}}$ are obtained from the OLS method, and p is the number of explanatory variables in the model. The advantage of this method is that it is easy to calculate since the inputs are computed easily. The choice of optimal *K* using this method is supported by Polat and Turkan (2016). Compared to other methods in determining the value of the ridge regression parameter, the subjectivity involved is minimal.

The optimal K can be chosen visually via a ridge trace plot, which is a graphical representation of the estimated ridge regression parameters at different increasing levels of K. The optimal value of K is chosen where the ridge trace starts to stabilise or does not change rapidly. The stability does not confirm convergence of the regression coefficients, but shows that as K increases, the variance reduces, and the coefficients become more stable (Ambra and Sarnacchiaro, 2010). According to Polat and Gunay (2015), choosing the ridge parameter through the ridge trace is subjective, making it highly probable to choose a higher value of K. The desirable way of choosing the parameter would be through a scientific way. An example of a ridge trace is depicted in Figure 2.1.



Source: http://www.statistics4u.info/fundstat_eng/img/ridge_regression_trace.png

Figure 2.1 Trace plot in estimation of *K*

In Figure 2.1, the y-axis represents the regression coefficients of the independent variables, while the x-axis represents the value of the ridge regression parameter *K* (parameter lambda)

Mardikyan and Cetin (2008) recommend another procedure that makes use of a plot in choosing the ridge regression parameter. In this instance, the ridge regression parameter is determined using the degrees of freedom trace (df-trace), where the degrees of freedom of the model are plotted against the different values of the ridge regression parameter. The ridge regression parameter is then chosen where the degrees of freedom become stable.

An alternative way to choosing *K* is using the variance inflation factor (VIFs). The VIF is a multiplier of the variance of an estimated regression coefficient due to its correlation with other independent variables as compared to its variance. This is subject to the explanatory variables being orthogonal (O'Brien, 2007). For instance, in a model with two independent variables,

the $VIF = \frac{1}{1-r_{1,2}^2}$, where $r_{1,2}^2$ would denote the correlation between two explanatory variables. It suffices to choose the ridge regression parameter K, where the VIF denotes the lowest value. The VIFs find use in the diagnostic phase of multicollinearity and the testing phase. The diagnostics phase entails checking if multicollinearity exists, while the testing phase involves assessing the effectiveness of the ridge regression parameter chosen to alleviate the problem of multicollinearity. An augmented VIF is used in the diagnostic phase while the VIF for the testing phase requires the variables to be standardized (Garcia *et al.*, 2016). The standardisation associated with the augmented VIF, often called the correlation transformation as the $\mathbf{X}'\mathbf{X}$ entries range between -1 and 1. This standardisation is less subject to rounding off errors; hence, even in plotting the ridge trace, it is paramount that standardized coefficients be used. Additionally, Garcia *et al.* (2016) points out that there is danger of naively misinterpreting the meaning of the plot, citing that the appearance of the ridge trace may be fundamentally changed by a simple scale transformation of variables and translation of the origin.

Aside from the standardisation of variables playing a critical role in the testing phase of the ridge regression parameter's effectiveness, there exists certain conditions that a researcher must be aware of. These include the VIFs must be continuous at $K = 0$, monotonically decreasing and higher than 1 for all K (Garcia *et al.*, 2016). Moreover, when $K = 0$, the augment VIFs from the ridge regression would resemble the VIFs from OLS. Failure to satisfy this condition will lead to incorrect conclusions being drawn about the alleviation of multicollinearity. The monotonicity condition, on the other hand, ensures that the parameter K chosen is at the level where augmented VIF is decreasing and not increasing as it can lead the researcher to choose the wrong K (Garcia *et al.*, 2016).

Mardikyan and Cetin (2008) point out that the difficulty encountered is in choosing the optimal value of K and ensuring the regression coefficients from the ridge regression perform better and are optimal when compared to OLS regression coefficients. A mathematical programming model is chosen to determine an efficient ridge regression parameter that minimizes the VIF values and simultaneously maximises the coefficient of determination R^2 (Mardikyan and Cetin, 2008). Through this highly analytical method, the study proves to be beneficial to the researcher as it avoids the trial-and-error re-runs of choosing the ridge regression parameter at different levels. Additionally, the added advantage of the mathematical programming from a researcher's point of view is its efficiency – it can be incorporated in statistical software tools such as SPSS, is readily made available in MS Excel and most importantly it saves time. The study does not cite other statistical software tools that are widely used, like R or statistical analysis software (SAS) which can take large datasets. One drawback noted on the mathematical programming way of choosing the ridge regression parameter is that although it is deemed reliable and most likely accurate than the conventional ridge trace, it might be difficult to compute as it involves complex calculations like using macros (Mardikyan and Cetin, 2008).

Dorugade and Kashid (2010) and Al-Hassan (2010) suggest alternative ways of choosing the ridge regression parameter that outperforms OLS estimators. According to Dorugade (2016), a better method to attain the ridge regression parameters without computing the shrinkage parameter K exists. Attaining the ridge regression estimator through conventional means comprises of complex equations because the ridge regression estimator is a non-linear function of the ridge regression parameter (Dorugade, 2016). Compared to the conventional ridge regression estimator, the adjusted ridge regression estimator uses the information on the correlation coefficient between variables. The correlation coefficient can be used to detect

multicollinearity and as part of the solution for the problem of multicollinearity. According to Dorugade (2016), the new estimators performed better than the ordinary ridge regression, generalised ridge regression estimators and OLS estimators in terms of the mean square error (MSE), particularly under the conditions of extreme multicollinearity. An opportunity exists where enhancements on the ordinary ridge regression improves the performance in terms of efficiency and effectiveness.

2.2.2 Advantages of the ridge regression over the ordinary least squares

Muniz, Kibria and Shukur (2012) established, through the Monte Carlo simulation, the strength of different ridge regression (RR) parameters based on the varying sample sizes, varying number of explanatory variables in the models, different levels of correlation between explanatory variable and the different levels of standard deviation. In certain instances of multi-layered circumstances such as having increased correlations between the independent variables, increased standard deviations and increased number of explanatory variables, the detrimental impact on the mean square error (MSE) might be high. Thus, instead of the MSE decreasing as per the norm, it increases. Increasing the sample size would result in a lower MSE, even when the correlations between the explanatory variables increased. The MSE obtained using the generalised ridge regression approach is relatively lower than that of OLS even when different computations of the RR parameters are used (Muniz *et al.* 2012). This is also substantiated through several literatures such as Ogunjobi *et al.*, (2017).

Ridge regression has been found to outperform OLS, particularly when dealing with macroeconomic data, as outlined in the research study regarding the unemployment rate in Iraq (Bager *et al.* 2017). Using ridge regression, the macroeconomic indicators that affect the unemployment rate were ascertained to be economic output, inflation rate, volume of

investment, public expenditure and size of the labour force. The strength of the ridge regression in the alleviation of multicollinearity tested using VIF was satisfactory even without omitting the highly correlated variables from the data. Some of the explanatory variables were statistically not significant since the Iraqi macroeconomic data used was highly susceptible to various shocks (Bager *et al.,* 2017). Further research into the optimal performance of ridge regression when multicollinearity and outliers caused by shocks are present in the data is necessary.

Ridge regression would not be sufficient if simultaneous problems arise in a dataset, such as having multicollinearity and outliers, (Zahari, Ramli and Mokhtar, 2014). Time series data, particularly macroeconomic data, is vulnerable to having outliers that are normally caused by shocks in the economy. For instance, a sudden change in the government structure may cause the currency to depreciate, triggering all the related variables in the economy to behave in an unusual way. Zahari *et al.* (2014) suggested bootstrapping robust ridge regression estimates with fixed resampling. The bootstrapping technique produces better parameter estimates, with lower standard errors compared to those of OLS in scenarios, where the sample size and outliers increase. The enhancement of ridge regression through the bootstrapping technique has the added advantage of possible extension in alleviating the problem of multicollinearity in logistic and poisson models (Zahari *et al.*, 2014).

2.2.3 Limitations of ridge regression

Ridge regression is not without critics, as illustrated by Shariff and Ferdaos (2017) and Adegoke *et al.* (2016). The Shariff and Ferdaos (2017) criticise ridge regression for its limitations, particularly when dealing with data that has a combination of correlated explanatory variables and consisting of outliers. The presence of outliers tends to render the

OLS estimates meaningless (Bagheri and Midi, 2009). A new proposed robust ridge regression estimator, the generalised m-estimator, outperforms both the ordinary least squares and ridge regression estimators because it encompasses the use of weights to filter out the outliers (Shariff and Ferdaos, 2017). This research prompts the users of data to always check for outliers, particularly in data that is vulnerable to outliers and to compute descriptive analysis to check how far each variable is from its mean or median. The incorrect conclusions because of the presence of the outliers are avoid as a result.

The presence of outliers in a regression model unduly influences the parameter estimates as well as the predictive power of the regression model (Polat and Turkan, 2016). These outliers can occur either in the explanatory variables or the dependent variable. In the case of predicting the NPLs as part of the IFRS 9 methodology using macroeconomic data, it is vital that any limiting influence on the predictive power of the model be assessed for materiality and subsequently be removed as this can have an unfavourable effect on the losses of the bank. Consequently, Polat and Turkan (2016) suggested using a robust ridge regression (RRR). They argue that this method is better than the classical ridge regression as it caters for outliers through the deployment of a 'scale m-estimator'. This estimator is cited as best to ensure robustness and efficiency, as illustrated when the ratio of independent variables (p) and number of observations (n) is large. The measuring of the performances of the models using the trimmed root mean squared error (TRMSE) instead of mean squared error (MSE) is suggested when comparing the predictive ability of RRR with the classical ridge regression (Polat and Turkan, 2016). Using the TRMSE instead of the MSE has advantages, as large errors are penalised or assigned more weight than small errors and it has proven to be effective in improving the performance of a model (Chai and Draxler, 2014). To further illustrate the preference and competitive advantage of RRR, Polat and Turkan (2016), compared six models, i.e. classical

ridge regression (RR), principal component regression (PCR), statistically inspired modification of the partial least squares method (SIMPLS) as well as their robust counterparts. From these simulations, RRR was found to outperform all the other five models as it had the smallest TRMSE. The simulation results could not clearly differentiate the performance of other robust models of PCR and SIMPLS from RRR in situations where n>p or when p>n as well as distinguishing the types of outliers that the data consists of (bad leverage points and vertical outliers).

The conventional way of calculating the ridge regression parameter tends to sway towards the notion that the observations are assumed to be identically and independently distributed (i.i.d). The central limit theorem holds under these assumptions, where for a given population mean and standard deviation, the distribution tends to be approximately normally distributed, particularly for large sample sizes. The use of ridge regression to solve the problem of multicollinearity requires these assumptions to hold. Mansoon and Shukur (2011) offer a modification to the calculation of the ridge regression parameter by introducing a poisson ridge regression estimator, particularly for the count data. The maximum likelihood (ML) estimation method is used in the poisson ridge regression model. The performance of the poisson ridge regression parameters is compared to that of ML in terms of the MSE. Through different simulations, the MSE decreased as the sample size and value of the intercept increased but was higher when the number of independent variables and the correlations between the explanatory variables increased. The proposed poisson ridge regression method is better than the ML in the presence of multicollinearity (Mansoon and Shukur, 2011).

When determining the robustness of the ridge regression (RR), particularly on some probability distributions, Zakari, Yau and Usman (2018) encountered mixed results. The probability

distributions analysed included gamma, beta and chi square distributions, where RR was found to outperform LASSO regression (LR) and partial least squares regression (PLSR) on gamma distribution at most times. Zakari *et al.* (2018) employed performance indicators such as mean absolute error (MAE), r-square (R^2), mean square log error (MSLE) along with variations of several explanatory variables (p = 4 and p =10) and varying sample sizes (n = 60 and n=90). At n = 60, with p = 10, RR possess a higher predictive power on the gamma and chi-square distributions than other models in terms of MAE, MSLE and R^2. When the explanatory variables are reduced to 4, and the sample size increase to 90, the RR only outperforms the other models on the gamma and beta distributions. In all simulations of varying sample sizes and number of explanatory variables, the RR seems to have higher predictability power consistently on the gamma distribution than beta and chi-squared distributions.

Choosing of the bias *K* in ridge regression to alleviate the problem of multicollinearity affects the stability or consistency of the variances (Chandrasekhar, Bagyalakshmi, Srinivasan and Gallo, 2016). An alternative to the ridge regression (RR) is the partial ridge (PR) model, as it selectively alters the ridge constants associated with highly collinear variables to control volatility in the variances of coefficient estimates. The bias is only added when necessary to the variables that experience a high degree of collinearity instead of adding the bias across all variables irrespective of whether they are highly collinear or not. The ridge regression's method for adding the bias across all variables decreases the precision of the parameter estimates (Chandrasekhar *et al*., 2016). To further prove that the partial ridge model outperforms the ridge regression, the MSE criterion, relative efficiency and estimation bias was used for assessment. The results proved that the PR had a significantly lesser MSE than that of RR, particularly at high levels of collinearity. The parameter estimates for the PR model were much

closer to the true β than the RR model (Chandrasekhar *et al*., 2016). A suggestion is for further research to be conducted to assess the consistency in other datasets.

According to Bagya, Gallo and Srinivasan (2018), the partial ridge regression method (PRR) is superior when compared to other enhancements of the ridge regression model. In contrast to ridge regression, PRR selectively adjusts the ridge constants for variables that are highly collinear to control the instability of the coefficient variances. The PRR method utilises the dimension reduction procedure, the singular value decomposition (SVD) in the selection of biasing the correlated variables to obtain the regression coefficients (Bagya *et al*., 2018). The comparison methods used were the ridge regression, generalised ridge regression (GRR) and the directed ridge regression (DRR). The GRR criteria for choosing K allows for a separate biasing parameter for each explanatory variable, while in DRR, the shrinking of the parameter vector is only limited for those coefficients with small eigenvalues. Using the Monte Carlo simulation, the PRR was proved to be a better method in dealing with multicollinearity when compared to RR, GRR and DRR. For lower levels of sample sizes and correlation, RR and PRR performed well, while the MSE values for DRR were less than that of GRR when the sample size and number of explanatory variables was less. Overall, the PRR had the lesser MSE, prompting the conclusion that a single perturbation of the ridge regression estimator K is much efficient and stable when compared to multiple ridge regression estimators (Bagya *et al*., 2018).

The ridge regression simultaneously deals with multicollinearity as well as ensure that predictive power is robust and can be enhanced to achieve robust methods (Lipovetsky, 2010). The enhanced model, such as RE3 adjusted ridge regression, is proven to be efficient, less biased, yield robust solutions, have coefficients that can be interpreted easily and overall

encompasses the good characteristics of the ordinary ridge regression (ORR). In ORR, the main objective is to find a ridge regression parameter, such that the model is robust. When determining the best possible value for this estimator, one finds that normally as the ridge regression parameter increases, the regression coefficients tend to shrink to zero. This is not the case with enhanced models. In all modelling of data, a guiding principle that the researcher must ensure is the practicality of the suggested method in real life.

2.3 The principal component regression in tackling the multicollinearity problem

Principal component regression (PCR) is a method that transforms the correlated variables into a set of uncorrelated variables called principal components, thereby combating the problem of multicollinearity. The characteristic of orthogonality of the new set of variables reduces the severity of the consequences of modelling non-orthogonal data. PCR has several assumptions that it adheres to, so that the model can be robust. These include the linearity nature of the independent variables, and that each variable is postulated to be normally distributed amongst the others. This ideally makes PCR to be one of the advocated methods to combat the problem of multicollinearity when the variables are linearly correlated.

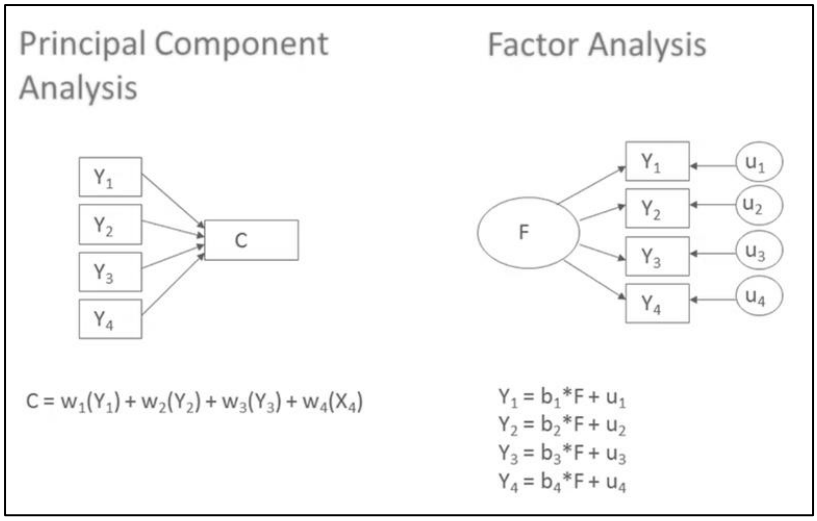
2.3.1 Link between principal component analysis and factor analysis

Often principal component analysis (or regression) can be mistaken for factor analysis (FA) due to the similarities between the two. As Brown (2009) states, the general notion of factor analysis is to represent a set of variables in terms of a smaller number of hypothetical variables. Likewise, factor analysis attempts to do the same, hence it suffices that both methods are data reduction techniques as they approximate data in lower dimensions.

The other similarity between the two methods stems from the usage of eigenvalues and eigenvectors in determining the number of components or factors to retain. Additionally, the interpretation of the results is the same in both PCA and FA and both models can use statistical software such as SAS for analysis. The FA has the added advantage that the factors can be rotated, making it much easier to interpret.

The correlation matrices in both methods are also used in the analysis of the variables that are being studied. Though one can say that PCA is a method that utilises and unpacks the variance structure, while FA uses the covariance structure. PCA seeks to minimize the sum of errors, hence the error variances as well as patterns that emerge in the data are more important for this type of analysis (Brown, 2009). PCA is expressed as a linear combination of the original variables. In FA, the primary objective is to analyse the covariance accounted for in the research that exists between the variables. The structure of FA covariances is such that the diagonal entries are not like that of PCA, instead the diagonal entries are variances of communalities.

Figure 2.2 depicts the difference between the PCA and the FA structures and the way they are computed. For PCA, the arrows depict the weighted contributions of the Y variable to the component C. For FA, the variables are pointing in the opposite direction of the latent factor F. The optimal weights are used to determine F, while the error terms depict the variance in Y that is not explained in the model. In principal component analysis, each of the principal component is express as a linear function of the original variables, whereas in factor analysis, each variable is expressed as a linear function of the selected factors.



Source: https://thecraftofstatisticalanalysis.com/principal-component-analysis/

Figure 2.2 The schematic relationship between the PCA and the FA

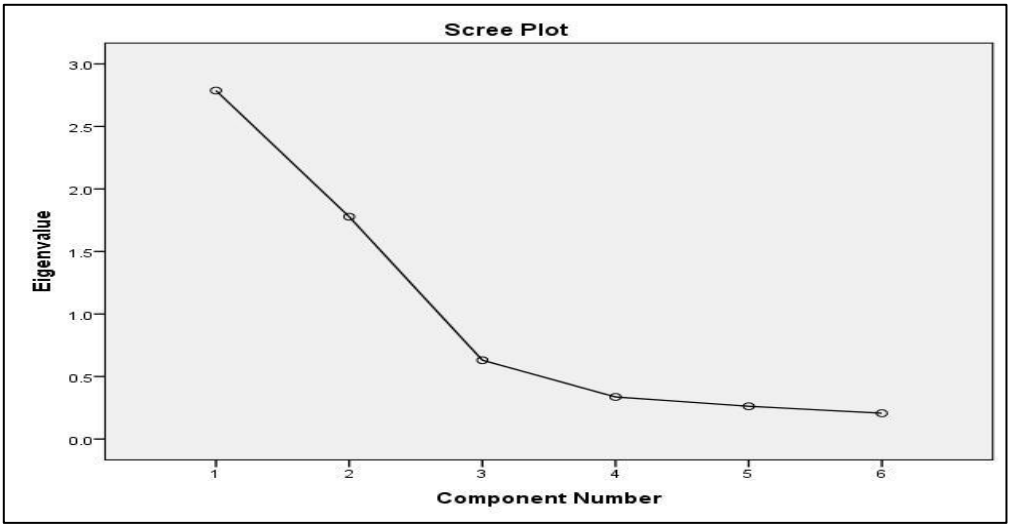
The choice of the appropriate method to apply depends on the desire to generate new variables, ease of interpretation and the attainable assumptions.

2.3.2 Determining the number of components

Three commonly used techniques are the scree plot, eigenvalues greater than unit for correlation matrix, and a set proportion of variation accounted for by components. Consideration of at least two methods is recommended in deciding the appropriate numbers of principal components. A scientific way of choosing the number of components to be included in the principal component regression can be determined using eigenvalues. Through the Kaiser criterion, the eigenvalues that are greater than 1 are included in the model. Moreover, according to Thupeng *et al.,* (2018), the minimum number of principal components that accounts for the most variation in the data are chosen, hence principal component regression is often referred to a data reduction technique. Chen and Tidal (2014) criticises this method as choosing the principal components by variation assumes that the degree of variation is indicative of causation.

A scree plot is a graphical technique (Thupeng *et al*., 2018) that plots the order of the eigenvalues on the x-axis and the magnitude on the y-axis. To select an appropriate number, consider the elbow of the scree plot. Consider those before the flatting of the plot as the appropriate number. For the eigenvalues obtained using the correlation matrix, those above 1 are considered ideal. The eigenvalues are used to indicate the extent of multicollinearity and those that are equal or greater than 1 indicate that the independent variables are orthogonal, while eigenvalues closer to *zero* indicates presence of multicollinearity (Alibuhtto and Peiris, 2015). The retained components are those on the steep part of the trend on the scree plot before it flattens out as depicted in *Source*:https://www.empirical-methods.hslu.ch/decisiontree/interdependency-analysis/reduction-of-variables/3191-2/

Figure 2.3.

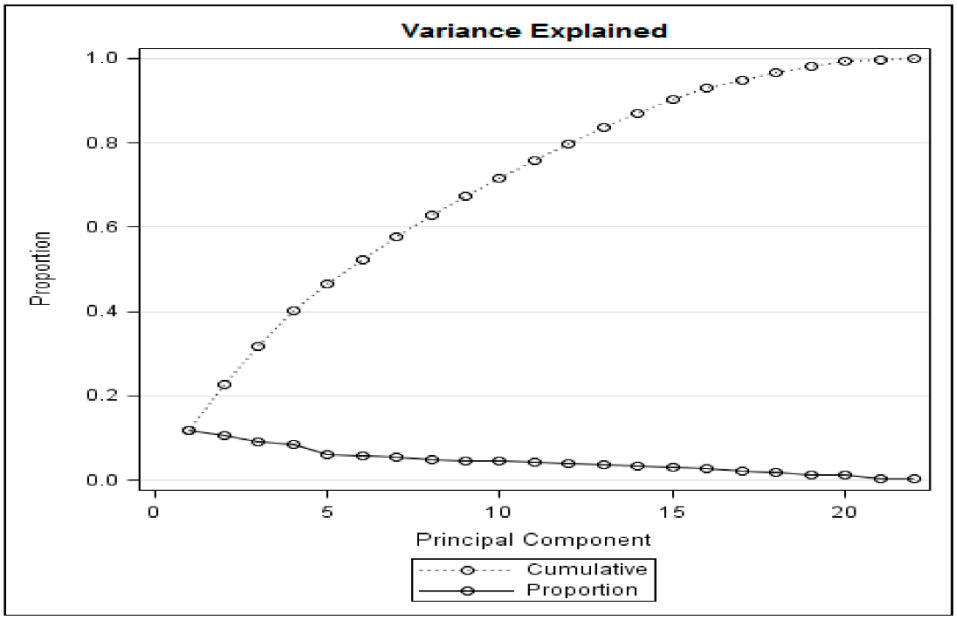


Source:https://www.empirical-methods.hslu.ch/decisiontree/interdependency-analysis/reduction-of-variables/3191-2/

Figure 2.3 A scree plot of eigenvalues generated from a correlation matrix

The graphical presentation of the variance of the principal components depicting the components that account for the maximum variance in the data shows the amount accounted for by each component. Through standardisation, each variable has a mean of zero and a

variance of 1. Standardisation implies that each variable has a variance of 1, and the total variance is equal to the number of variables in the original dataset (Kumar and Goyal, 2011). Figure 2.4 shows the proportion of variance encountered for by each variable.



Source: https://www.mdpi.com/2071-1050/12/14/5537/htm

Figure 2.4 Plot of the proportion of variability accounted for by each principal component

2.3.3 Superiority of principal component regression in relation to other regression models

According to Ma and Dai (2011), the principal component regression (PCR) effectively solves the problem of multicollinearity and is hailed as a better method owing to its computational simplicity. The PCR is used in several existing software packages, because of its low cost and is preferred to other alternative methods. The normality assumption on the data required in performing the PCR, in particular in bioinformatics, may or may not hold. The PCR performed satisfactory despite the lack of normality in the data and with no theoretical justification (Man and Dai, 2011).

Principal component regression (PCR) is widely used particularly in environmental related problems that involve collinearities in the data. This biased regression method is basically used to stabilise the regression estimates in the presence of multicollinearity by introducing a bias that ultimately reduces the variance of the estimated coefficients (Thupeng *et al.*, 2018). PCR method can also be used to solve the problem of multicollinearity by eliminating the model instability and reducing the variances of the regression coefficients (Ambra and Sarnacchiaro, 2010. The reduction of this variance must compensate for the biasness introduced in the data and be able to only minimize the number of estimators included in the model. The Kaiser criterion was deployed to choose the number of components to be included in the model. The PCR model was observed to perform better at forecasting daily peak ambient ground level ozone concentrations as it denoted through the F-value and the coefficient of variation that were statistically significant (Thupeng *et al.*, 2018). The advantage of using the PCR model is that it can be duplicated and implemented in other similar areas to measure air pollution, though it can be limiting if there are other variables present that were not included in the original model.

Kumar and Goyal (2011) illustrated the versatility of PCR in forecasting the air density index to combat the air pollution problem. They demonstrated the superiority of PCR over multiple linear regression (MLR) in tackling the multicollinearity problem. Once multicollinearity is detected, the PCR determines the relevant independent variables to include in the model and reduce the complexity associated with large datasets (Kumar and Goyal, 2011). When compared to the OLS method, PCR is a better method as it tends to only include the variables that account for the most variability in the data (Alibuhtto and Peiris, 2015). In contrast, the statistical results revealed the under-prediction of the PCR using normalized mean square (NMSE) and the coefficient of determination (R^2) in the four seasons that were analysed. When

comparing the overall performance of PCR and MLR, the PCR outperforms the competitor model, particularly because it eliminates collinearity that leads to unreliable predictions if left untreated.

The principal component regression (PCR) is used as a data reduction technique and to reduce the risk of overfitting by using a small number of estimators (Marinoiu, 2017; Shlens, 2014). A subset of the principal components is used which considers only the variables that have higher variability in the regression model (Shlens, 2014). Despite these findings, Marinoiu (2017) argues that even though the objective of orthogonality is achieved through PCR, the correlation degree between the new set of variables and the dependent variable is ignored and not thoroughly analysed.

In a paper by Gorgees (2017), the author demonstrated the advantageous view of principal component regression over different types of ridge regression. The ridge regression types proposed, varied on the evaluation of the ridge parameter including the Bayesian approach and using the concept of the condition number. Since the number of observations in a sample size is deemed important for the performance of the different estimators, Gorgees (2017) demonstrated the analogy by varying the sample sizes as well as standard deviation. The criterion used to rank the performance was the minimum mean square error (MSE) that revealed that principal component regression performed better than its competitors. The author did not further test the performance of principal component regression against other modified versions of ridge regression such as the Jacknife ridge regression method.

2.3.4 Confines of principal component regression

Principal component regression (PCR) as indicated by Maestre and Escudero (2009), Ogah (2011) and Alibuhtto and Peiris (2015) proved it to be a method that can be used to alleviate the problem of multicollinearity by producing meaningful results. The respective papers suggest using eigenvalues with a value of more than 1 for further analysis in the respective analysis. Ciampi and Gordini (2008), concentrated on the effectiveness of using sets of economic-financial ratios for company default prediction statistical modelling. The findings revealed that though multicollinearity could be alleviated, the principal component regression is limited in its usefulness as the first component could only explain 23% of the variation, while the second could only explain 32% of the variation (Ciampi and Gordini, 2008). The authors therefore opted to use discriminant analysis and logistic regression for prediction purposes.

Principal component regression is a useful tool in combating multicollinearity and simultaneously as a data reduction technique (Lokesh, Maurya, Koutu, Singh, Shukla, and Mishra, 2017). Simply altering the correlated independent variables into uncorrelated components is a huge win in evading the consequences that arises from multicollinearity. The criticism of this approach stems from the difficulty encountered when attempting to interpret the estimated regression coefficients. The difficulty ascribed to the basic notion that the components are the linear combinations of the original variables (Pepler, 2014). Nduka and Ijomah (2012) also attest that interpretation of the regression coefficients from this model is a problem as importance of the predictors is being masked by the way the linear combinations were formed.

In a study by Ayinde, Lukman and Arowolo (2015), the performance of the principal component regression is explored to alleviate multicollinearity, particularly when the error

terms are not independent. This issue of dependent error terms, known as autocorrelation, is often encountered in time series data like macroeconomic data. To deal with this multi-layered issue, an enhanced estimator is developed that incorporates the principal component regression with a feasible generalised estimator. This estimator was found to outperform the OLS in terms of the MSE (Ayinde *et al.*, 2015).

Principal component regression has been known to be useful when it comes to reducing the severity of multicollinearity. The principle of using components that account for the maximum variability that this multivariate analysis employs may be problematic. This stems from the fact that the use of limited data comes with the risk of imprecision as important variables may be left out of the model (Junttila and Laine, 2017). In addition, the principal components that are chosen, only explain the independent variables rather than the dependent variables, hence it is not guaranteed that the chosen components are relevant for the inclusion in the model (Nduka and Ijomah, 2012). To counteract this deficiency on principal component regression, prior information regarding the independent variables is important to the overall analysis before making the final decision to exclude the variables with less variability.

2.4 Other Statistical approaches to solve multicollinearity

Despite the advantages of using ridge regression and principal component regression in combating the problem of multicollinearity, there are several methods that can be deployed to solve the problem of multicollinearity. Even though some authors recommend non-statistical ways such as ‘do nothing’, literature is ripe with other methods to solve the problem of multicollinearity. A highlight of these methods follows.

2.4.1 Exclusion of high collinear variables

One of the methods that is not reverberated in literature regarding handling of multicollinearity would be to exclude variables that have a high level of dependency on one another. As illustrated by Thompson, Kim, Aloe and Becker (2017), one of the options would be to just re-construct the model to exclude the variables contributing to multicollinearity.

O'Brien (2017) argues 'why is it typically not a good idea to drop highly collinear variables from the model'. A distinguishing factor between what is deemed as the 'independent variable of interest' (IVOI) and of less interest variable as a 'control variable'. To substantiate the hypothesis that multicollinearity is not a sufficient reason to drop variables from the model, the author argues the criterion of model influence. Model influence is based on whether dropping one variable from the analysis results in the change of direction of the regression coefficient of IVOI, or shifts the p-value substantially (O'Brien, 2017). Having said that, the variables that are often deemed as less of interest (control variables), have a role to play in the model despite their statistically insignificant status. Thus, researchers need to scrutinise the data thoroughly before choosing to drop a variable from the model.

2.4.2 Partial least squares regression

Partial least squares regression (PLSR) can be explained as a method that seeks to predict the dependent variable at the back the explanatory variables that are singular (Ambra and Sarnacchiaro, 2010). This method can be viewed as a combination of principal component regression (PCR) and ordinary least squares (OLS). In contrast to the principal component regression, the PLSR method seeks to find components that not only have a high variance, but rather are relevant for the dependent variable. The chosen components have high collinearity percentages with the dependent variable. The advantageous aspect of this method is its ability

to solve the problem of multicollinearity, without losing the covariance power of the explanatory variables on the dependent variable (Ambra and Sarnacchiaro, 2010). Additionally, since PLSR encompasses PCR, the caveat derived is that it can be viewed as a dimension reduction method (Polat and Gunay, 2015). Using the real-life dataset of air pollution in the Polat and Gunay (2015) research study, the PLSR method proved to have the highest predictive power, while alleviating the problem of multicollinearity when compared to the principal component regression, multiple linear regression, and ridge regression.

2.4.3 Perturbation model

The use of perturbation of eigenvalues can also produce a robust model that outperforms the conventional principal component regression and ridge regression. Nduka and Ijomah (2012), outline the importance of eigenvalues from being used for detection of multicollinearity, but rather more for producing a superior prediction model. An eigenvalue that is close to zero indicates that there multicollinearity exists. If the eigenvalues are re-constructed such that they are 'pushed away from zero', in so doing, collinearity among the explanatory variables is minimized (Nduka and Ijomah, 2012). Thus, employing varying simulation of different sample sizes to prove the predictive power of the Perturbation Model. Using the root mean square error, the perturbed eigenvalue estimator was found to outperform the ridge regression, principal component regression and ordinary least squares. The new model was found to be superior in the predictive power than the other models and also alleviated the problem of multicollinearity.

2.4.4 Inequality constrained least squares and dual estimator models

Gordinsky (2016) suggests a regression approximation of the distribution of the event $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}'\boldsymbol{\beta}$ of the edgeworth series. In this approach, the researcher concentrates on using the

inequality constrained least squares (ICLS) and dual estimator (DE) methods to show that with external information, these methods can greatly reduce the euclidean distance between the estimated regression coefficients and the associated true parameters as well as the confidence intervals (Gordinsky, 2016). The advantageous effect of ICLS is its flexibility as it can consider constraints of the unknown regression coefficients. The method only considers inequalities of priori information such that appropriate sizes are bounded, i.e. $lb \leq \beta \leq ub$, while inequalities such as $\beta_i > \beta_j$ cannot be utilised (Gordinsky, 2016). In using priori information, the researcher must confirm that this information is not related to the dependent variable, otherwise obtaining the inequalities of the regression coefficients might be difficult. In terms of the dual estimator, the model proves to be superior to the ICLS model in situations, where multicollinearity is present and mutual compensation of the explanatory variables is absent. As stated by Gordinsky (2016), in this instance, the dual estimator can produce unbiased and consistent solutions, while the ICLS would not be applicable.

Backward in time selection method (BTS) is often recommended when dealing with time series that has lagged variables. This method evaluates the inclusion of the lagged variables starting with the most recent and going back in time as well as the ones that are correlated with the dependent variable. The logic behind this method stems from the belief that multicollinearity is inherent for variables that are close than the ones that are falling further apart in time (Vlachos and Kugiumtzis, 2013). The BTS model is quite simple in its computation and proved to be conservative as it contains minimal redundant information. When compared with other models, the BTS model demonstrated consistency on the prediction ability and power to correct the problem of multicollinearity given the varying sample sizes, unlike the other methods such as forward stagewise regression, ordinary least squares and least absolute shrinkage and selection operator (LASSO) regression (Vlachos and Kugiumtzis, 2013).

CHAPTER 3

STATISTICAL TECHNIQUES APPLICABLE TO THE MODELLING OF THE NON-PERFORMING LOAN DATA

3 Methodology

The three methods chosen to combat the problem of multicollinearity in the Africa portfolio data are ordinary least squares (OLS), ridge regression (RR) and principal component regression (PCR). An analysis to review the best method that deals with multicollinearity was completed, while also measuring the robustness of each method. In assessing these methods, the ultimate objective remained to build a predictive model that can forecast the non-performing loan (NPL) ratio, using data that has been adjusted for multicollinearity.

3.1 Model fitting procedure

3.1.1 Variable selection

Post detection and diagnosis of multicollinearity, the next step is finding regression models that can be used in predicting NPLs. Descriptive statistics used to gain more knowledge of the variables, include the mean, average, minimum value and maximum values of each variable in the dataset.

Thereafter, the three methods (ordinary least squares, ridge regression and principal component regression) are modelled, and the results compared in terms of the significance of the individual t-tests of each independent variable against an appropriate p-value, the adjusted coefficient of determination, standard deviation of each variable and finally, the mean square error (MSE). The method that performs better than OLS is the method with the lowest value of MSE.

3.1.2 Regression model comparison

Ordinary least squares

Ordinary least squares (OLS) is a method that is mathematically and intuitively appealing, hence, it is a widely popular statistical technique to use in regression analysis (Gujarati, 2003). One of the vital benefits of OLS stems from the Gauss-Markov theorem, whereby the estimated parameter vector $\widehat{\boldsymbol{\beta}}$ is unbiased. This means that if repeated samples of data were taken and sampled, the distribution of $\widehat{\boldsymbol{\beta}}$ would eventually converge to the true population parameter $\boldsymbol{\beta}$. The disadvantage of OLS is its adherence to the ten assumptions outlined by Gujarati (2003), where failure to adhere to the assumptions affects the efficiency of the model. The presence of multicollinearity results in OLS estimators possessing large variances, have larger confidence intervals and are sensitive to small changes within the data.

Ordinary least squares model – structure

The ordinary least squares (OLS) model is denoted in matrix form by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\boldsymbol{y}$ is an n x 1, the data vector

$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$, $\boldsymbol{X}$ is an n x p matrix denoting the independent variables

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$, $\boldsymbol{\beta}$ is a p x 1 parameter vector of the regression coefficients of the independent variables, $\boldsymbol{\varepsilon} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$, $\boldsymbol{\varepsilon}$ is n x 1 vector of random errors that are assumed to be identically,

independently and normally distributed with a mean of zero and a constant variance.

Random errors ε are important as they depict the differences between the predicted values of $\hat{y}$ and the observed values of y.

$$\varepsilon_i = Y_i - \hat{Y}_i \text{ , } i=1,2,\ldots, n$$

Ideally, we want the random errors to be as close to zero as possible. OLS attempts to minimize the sum of squared error terms (SSE) by obtaining parameter estimates $\widehat{\boldsymbol{\beta}}$ such that $SSE = \sum_i^n e_i^2 = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$, where SSE explains the total variation of the predicted variable and the observed values. This then translates to a set of normal equations that are obtained by minimising SSE to find parameter estimates $\widehat{\boldsymbol{\beta}}$ as outlined below:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

The above equation can be determined if $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists.

Properties of the ordinary least squares estimator – $\widehat{\boldsymbol{\beta}}$

One of the vital characteristics of ordinary least squares (OLS) stems from the Gauss-Markov theorem that stipulates that the OLS estimates are the 'best linear unbiased estimates (BLUE)'. This theorem can be described as follows:

a. The parameters of the model are 'linear'. This means that all the variables have exponents equal to 1.

b. The 'unbiased' concept of $\widehat{\boldsymbol{\beta}}$ is defined by the notion that if repeated samples of data were taken and sampled, the distribution of $\widehat{\boldsymbol{\beta}}$ would eventually converge to the true population parameter, $\boldsymbol{\beta}$. We show $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator:

   Let $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

$$E(\hat{\beta}) = E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y})$$

$$= E\left[\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\,(\boldsymbol{X}\boldsymbol{\beta}+\,\boldsymbol{\varepsilon})\right)\right]$$

$$= E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\,)] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'E(\boldsymbol{\varepsilon})$$

$$= E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\,)] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'(\mathbf{0})$$

$$= E(\boldsymbol{\beta}) + \mathbf{0}$$

$$= \boldsymbol{\beta}$$

Using $E(\boldsymbol{\varepsilon}) = 0$ as an assumptions of the ordinary least squares.

Violation of the independent error terms and the independent variables is referred to as endogeneity. When $E[\varepsilon|x] = 0$ the $E(\varepsilon) = 0$ and $Cov(x, \varepsilon) = 0$. The coefficient estimates of the model are affected when the estimates of the model parameters are incorrect, when there is correlation between the explanatory variables and the error terms. The Gauss-Markov theorem encompasses that the variance of the $\widehat{\boldsymbol{\beta}}$ must be the minimum variance across all other estimators (Gujarati, 2003).

The mean square error can be calculated by:

$MSE = \frac{SSE}{n-(p+1)}$, where p+1 includes the constant term.

The MSE of the OLS is normally small but can be inflated when $d_j \neq 0$, in the presence of multicollinearity. This can lead to incorrect estimates. The MSE provides a comparative metric for accuracy to determine the best model between the ridge regression and the principal component regression in combating the problem of multicollinearity.

The consistency property of the OLS estimators states that as the sample size *n* increases, the estimator would converge to the true value of the population parameter, i.e. $\lim_{n\to\infty} E[\hat{\beta}_n] = \beta$.

Residual Analysis

Autocorrelation

One of the assumptions of ordinary least squares is "no autocorrelation between the error terms". For any pair of observations, the error terms e_i and e_j, the $Cov(e_i, e_j) = 0$. Violation of this assumption results in inflated variances of the estimated coefficients. The Durbin-Watson (DW) test significance of this assumption. The DW test is given by,

$$DW = \frac{\sum_{j=2}^{n}(e_j - e_{j-1})^2}{\sum_{j=1}^{n} e_j^2},$$

where for the uncorrelated errors, the DW is close to 2, while for values less than 2, it is indicative of serial correlation between the errors. In SAS, a command "*DWPROB*", that produces the DW values as well as the associated p-values can be used.

Stationarity

In addition to autocorrelation, it is imperative to test for stationarity of the variables, particularly when dealing with time series data (Gidigbi, 2017). A time series would be deemed stationary if its mean, variance as well as autocorrelation do not fluctuate over time (Gujarati, 2003). In simple terms, the mean, variance, and autocorrelation are independent of time. The augmented dickey fuller test is used to test for stationarity. The "*PROC ARIMA*" command in SAS provides the augmented dickey fuller test. Modelling data that is deemed to be non-stationary is likely to produce spurious results.

Correlation Transformation of Variables

Correlation transformation, as outlined in Literature Review, is a necessity when trying to alleviate the problem of multicollinearity, particularly when using ridge regression, principal component regression as well as ordinary least squares. This transformation involves centering and scaling of regression coefficients. The magnitudes of the standardized coefficients are not

affected by the scales of the measurement of the various model variables (Chatterjee and Hadi, 2006). Suppose we have $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, then the dependent and independent variables are standardized as follows:

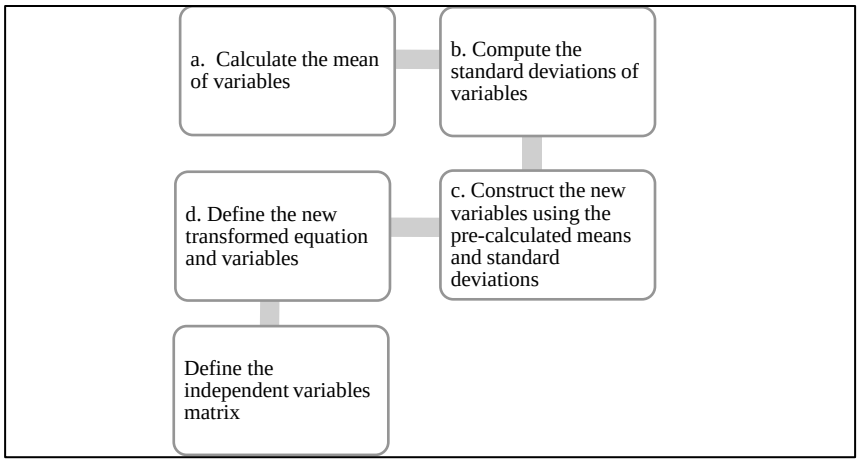


Figure 3.1 Steps in standardisation of variables

Ridge Regression

The strength of ridge regression lies in choosing an ideal small parameter K in such a way that the determinant of the design matrix of explanatory variables is not zero (Panik, 2009). Then it follows that the ridge regression coefficients become $\boldsymbol{\beta}(K)=(\boldsymbol{X}'\boldsymbol{X}+K\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$, where $K>0$. The ridge regression estimate shrinks the OLS estimate as $\boldsymbol{\beta}(K)$ is a coefficient vector with minimum length (El-Dereny and Rashwan, 2011). In addition, ridge regression is best suited to combat multicollinearity as there exists K such that the mean square error (MSE) is less than that of OLS. On that premise, unlike OLS, the ridge regression estimator is biased. This means that with the introduction of the ridge regression parameter K, ridge regression fails to uphold the Gauss-Markov theorem that in repeated samples, there exists biasness on the ridge regression coefficients – meaning that it cannot be equal to the true population parameter unlike the OLS estimates.

As mentioned earlier in the literature section, choosing the ridge regression can either be through a mathematical equation, ridge trace or using VIFs.

Characteristics of Ridge Regression

a. The ridge regression is a linear transformation of the OLS estimates as illustrated below:

   Let $\hat{\beta}^*$ be the ridge regression parameter, while the OLS estimate is

   $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, then since ridge regression is merely an

   $$\widehat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + KI)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + KI)^{-1}(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)\widehat{\boldsymbol{\beta}}$$

b. The length of the ridge regression parameter is a decreasing function of the ridge regression parameter.

c. The ridge regression estimator, unlike the OLS estimator, is biased. This is illustrated in a graphical presentation in Figure 3.2:

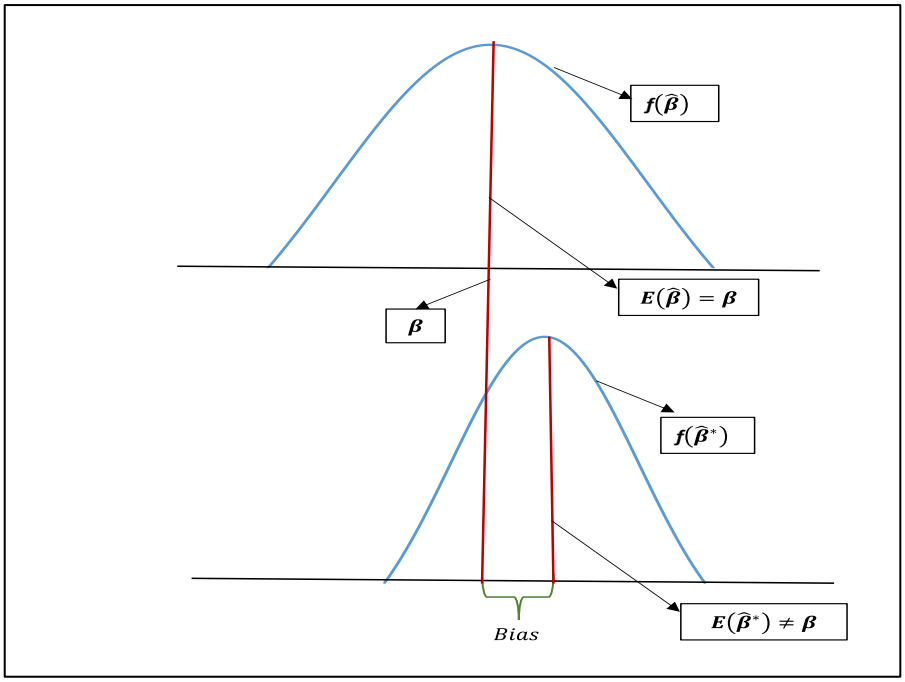


Figure 3.2 Demonstrating bias in ridge regression model

The introduction of the ridge regression parameter K, introduces the bias because it cannot be equal to the true population parameter, unlike the OLS estimates. Consider,

$$E(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + K\boldsymbol{I})^{-1}(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)\boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

In repeated samples, the ridge regression parameter will not converge to the true parameter, unlike the OLS estimate.

d. The mean square error (MSE) of ridge regression is calculated as:

$$MSE(\hat{\boldsymbol{\beta}}^*) = E\left[(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*)^2\right] = V(\hat{\boldsymbol{\beta}}^*) + \left[\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right]^2 = V(\hat{\boldsymbol{\beta}}^*) + \left[bias(\hat{\boldsymbol{\beta}}^*)\right]^2$$

where $V(\hat{\boldsymbol{\beta}}^*)$ measures the precision of $\hat{\beta}^*$ and $bias(\hat{\boldsymbol{\beta}}^*)$ measure its accuracy.

e. For finite $\boldsymbol{\beta}'\boldsymbol{\beta}$ there always exists $K > 0$ such that $MSE(\hat{\boldsymbol{\beta}}^*) < MSE(\hat{\boldsymbol{\beta}})$

Principal Component Regression

Principal component regression (PCR) transforms the correlated variables into a set of uncorrelated variables called principal components, thereby combating the problem of multicollinearity. The characteristic of orthogonality of the new set of variables reduces severity of the consequences of modelling non-orthogonal data. In addition, the advantage of the PCR is that it can be used as a data reduction technique. This means that it uses only a subset of the principal components by considering only the variables that have higher variability in the regression model. When compared to the OLS method, PCR is a better method as it tends to only include the variables that account for the most variability in the data (Alibuhtto and Peiris, 2015). This method is used to solve the problem of multicollinearity by eliminating the model's instability and reducing the variances of the regression coefficients (Ambra and Sarnacchiaro, 2010).

The weakness in using principal component regression is that if the eigenvalues are equal then it would mean that the principal components are not unique. Moreover, there might be loss of vital information when only the variables with the highest variability are modelled in the data.

Characteristics of Principal Component Analysis

a. Using the correlation matrix $\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}$, the eigenvalues are obtained through the following equation:

$$\left|\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*} - \lambda_j \boldsymbol{I}\right| = 0, j = 1,2,\dots,p$$

Otherwise in SAS, a command called '*COLLINOINT*' can be used to obtain the eigenvalues. To assess which of the principal components should be included in the model, one can use the Kaiser criterion, whereby only the principal components associated with eigenvalues that are

greater than 1 are included in the model (Thupeng *et al*., 2018). Alternatively, a scree plot can be used to determine the principal components to be included in the model.

b. Since the eigenvalues are a sum of the total variance in the model, then

$$\text{Total Variance} = \sum_{j=1}^{p} \lambda_j, \text{ and}$$

$$\text{Proportion of Variation} = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

The researcher would have to stipulate the amount of variance to be explained in the model, hence the variance percentage is subjective.

c. The eigenvectors represent how the principal component variables are related to the original variables (Ryan, 2009). The computation of eigenvectors is depicted below:

$$(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*} - \boldsymbol{\lambda_j} I)\boldsymbol{\alpha_j} = 0, where\ \boldsymbol{\alpha_j}\ \text{are the associated eigenvectors}$$

d. Let the standardized variables be defined by X^* and the eigenvectors be a_j, where

$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2p} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pp} \end{bmatrix}$, then it follows that the principal component $\boldsymbol{Z}$ is defined by:

$$\mathbf{Z} = \boldsymbol{X}^{*}\boldsymbol{\alpha}$$

$$= \boldsymbol{X}^{*}(\alpha_1, \ldots \alpha_p)$$

$= X_1^*\alpha_1 + X_2^*\alpha_2 + \cdots + X_p^*\alpha_p$. Showing that the principal component $\mathbf{Z}$ is a linear combination of the original variables $\mathbf{X}^*$.

e. The regression estimates from the principal components would now be $\widehat{\boldsymbol{\beta}}_{\boldsymbol{pc}} = (\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{Z}\boldsymbol{Y}^{*}$, where principal components are Z and Y^* is the standardized dependent variable. Then the estimated dependent variable is $\widehat{\boldsymbol{Y}}_{\boldsymbol{pc}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\boldsymbol{pc}}$.

3.2 Assessing the model competences

3.2.1 t-statistic and the confidence intervals of the parameter estimates

When modelling data in the presence of multicollinearity, the overall model is likely to be statistically significant, while the individual parameter estimates are not significant (Gujarati, 2003). The t-statistic is calculated as $t = \frac{\beta_j}{s_j\sqrt{C_{jj}}}$, where β_j is the parameter estimate of the associated independent variable and C_{jj} is the diagonal element of $(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*})^{-1}$. Then it follows that if we test for the significance of the estimates we would have:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

3.2.2 F-statistic for the overall model

The overall model test is done through the F-test, whereby the null hypothesis is determined as $H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$ and the alternative hypothesis becomes $H_1: At\ least\ one\ \beta_p \neq 0, where\ j = 1,2, \ldots, p$. Therefore, the test statistics to be evaluated is $F = \frac{MSR}{MSE} \sim F_{(p;n-p-1)}$.

The MSR and MSE denote the mean square for regression and mean square error, respectively. The F-test is then compared to the set level of p-value to see if the null hypothesis will be rejected or not.

3.2.3 Coefficient of Determination R^2

Another form of measurement that can be used to test for the validity of the model would be the coefficient of determination that basically assesses the total variability that is explained by the explanatory variables. The formula to determine this variability is depicted below:

$$R^2 = \left[\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}\right] = \left[\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}\right]$$

Where the Y_i represents the observed values, $\hat{Y}_i$ represents the predicted values and $\bar{Y}_i$ is the average of the observed values. In multivariate regression, the more robust form of R^2 is the one that considers the degrees of freedom. This robust coefficient of determination is called the $Adjusted\ R^2$ and is determined as per the below equation:

$$R^2{}_{adj} = 1 - \left[\frac{(1-R_2)(n-1)}{n-p-1}\right].$$

3.3 Out-of-time Testing

To analyse if the ridge regression and the principal component regression were successful in in the alleviation of multicollinearity, an out-of-time testing for 2016 to 2018 was used. Additionally, since the NPL ratios for the specified period are already known, the model that is closer to predicting the actual value of the NPL ratio for Kenya and Nigeria was used to test its validity and accuracy.

CHAPTER 4

DATA EXPLORATION AND MODEL FITTING

4 Data Exploration

4.1 Data Structure

The dependent variable under consideration is the non-performing loans (NPLs) ratio, for the respective portfolios acquired from the credit department at Bank X (name of bank kept anonymous due to data privacy compliance) from 2006 to 2018. The NPL ratio is based on the retail business, that is personal and business banking segments.

The modelling utilises data from different sources. The data was limited due to the establishment date of Bank X in the Nigeria and Kenya, hence data prior to 2010 is non-existent. Data on financial reporting from the two countries mentioned (Bank X monthly portfolio data) was used in the analysis. The two data sets were selected as case studies to enable us to develop predictable models that are free of multicollinearity problems. Data from Nigeria represents a broad banking sector cutting across small, medium to large economies of West Africa. Similarly, the Kenya data represents the economies of East Africa. In addition, due to their strict regulatory bodies, Nigeria was chosen due to its complex market, while Kenya was chosen as it recently had a regulatory stipulation regarding interest rates that affected all Banks.

The monthly and quarterly macroeconomic data from 2006 to 2018 for both Kenya and Nigeria was used in the analysis. The macroeconomic variables are considered as explanatory variables in this study. An outline and definition of these explanatory variables follows:

- *Interest Rate* – is the rate charged on the assets by the lender to the borrower. For this research, the lender would be Bank X, while the borrower would be the customers of Bank X.
- *Gross Domestic Product (Real Rate)* – represents the total value of the goods and services produced in a country for a period of one year.
- *Crude Oil Price* – is the spot price of one barrel of crude oil.
- *Treasury Bill Rate* – are mainly short-term money market instruments that are issued by the central bank on behalf of the government to curb the liquidity shortfalls in the economy.
- *Interbank Call Rate* – is a short-term money market rate which allows for banks, corporations and other financial institutions to borrow and lend money at the interbank call money market. The interbank rates are normally for overnight or at most weekly.
- *Lending Rate* – is the interest or rather the premium that a financial institution would charge for offering a loan.
- *Maximum Lending Rate (Max Lending Rate)* – is associated with the lending rate, as it is essentially the interest rate that the creditor or financial institution can charge the borrower.
- *Inflation Rate* – the rate at which prices increase over time, and ultimately impacting on the purchasing value of goods.
- *Monetary Policy Rate* – is one of the instruments that the central bank uses to control or curb inflation within the economy as well as control the money supply.
- *Deposit Rate* – is the interest that financial institutions would pay out for deposits.
- *M1 Money Supply* – is the type of money that consists of coins and notes that are currently in circulation in the economy.

- *M2 Money Supply* – in addition to consisting of M1 as explained above, M2 also constitutes short-term deposits and certain money market transfers.
- *Exchange Rate* – is the value of one currency for the purpose of conversion. For this research, the exchange rate used would be local currency to US dollar.
- *Foreign Exchange Reserves (USD bn)* – refers to the assets that are held by a central bank in various foreign currencies and are used to back liabilities on their own issued currency to influence monetary policy.
- *Central Bank Rate (%)* – refers to the interest rate that the central banks set for commercial banks.

Statistical techniques are applied to find the determinants of the NPL, first by assessing the presence of multicollinearity, and then offering solutions to the problem of multicollinearity.

4.2 Data Management

Trusted sources for data acquisition were used. These included readily available data from the central banks of Nigeria and Kenya, respectively. The data acquired from the respective central banks' websites was particularly based on monetary policy instruments such as inflation rate, monetary policy rate and interest rate to mention a few. Other variables like Treasury bill rate, exchange rate, monetary rate, maximum lending rate, interbank call rate and foreign exchange reserves were also mined from the websites. Data was also sourced from the economist intelligence unit website. The NPL ratios were acquired from Bank X in the respective jurisdictions, where data sourced from the central banks' data depositories was missing.

Some of the macroeconomic data can only be sourced quarterly and not monthly. For such data, a decision to use a proxy was taken, such that the quarterly value would be the same for the three months in a quarter. This is assuming that the values do not vary too much.

Data organisation using SAS

The software used for all data analysis was statistical analysis software (SAS). To appropriately import the respective datasets into the SAS environment, the CONTENTS procedure was used to identify the variables that will be used for the analysis. In addition, some of the procedures used included MEANS and UNIVARIATE programs. The MEANS procedure was used to determine the descriptive statistics of the variables that included average values, standard deviation, minimum and maximum values of each variable, while FREQ procedure assisted in identifying any missing information that might exist in the dataset as well as possible errors in the dataset. On the other hand, the UNIVARIATE procedure was employed to test for normality as well as help to identify outliers.

4.3 Issues and types of multicollinearity

One of the assumptions of a classical linear regression model is that the explanatory variables cannot have any correlation with each other (Gujarati, 2003). When this assumption is violated, then there exists a problem of multicollinearity that encompasses several non-desirable effects on the explanatory variables. The effect being that the explanatory variables become statistically insignificant, when they are supposedly significant (Daoud, 2017). Extensive literature exist that outlines the consequences of multicollinearity in difference disciplines other than statistics. According to Cheong, Kwak and Tang (2014), near-perfect multicollinearity appeared to have contributed to the fragility of the world trade organisations estimate,

particularly when dealing with structural variables that measure the general agreement on tariffs and trade membership status of any country pairs.

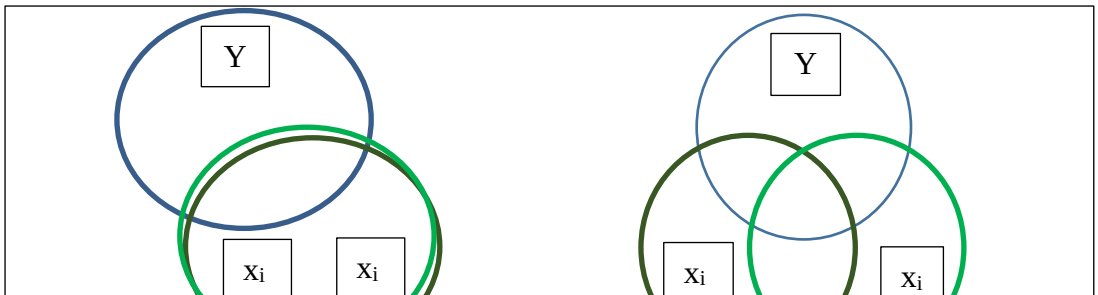
The nature of multicollinearity must be examined and ruled out in time series data (Gujarati, 2003) or where the researcher creates mathematically extra variables from existing variables. There is a distinguishing factor between perfect multicollinearity (exact) and less than perfect multicollinearity as explained by Gujarati (2003).

Perfect / exact multicollinearity

Suppose we have the linear equation as, $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_p X_p = 0$, where $\lambda_1, \lambda_2 \ldots \lambda_p$ are constants and non-zero. Consider X_1 as a function of the other variables, then we have $X_1 = \frac{\lambda_2 X_2}{\lambda_1} + \cdots + \frac{\lambda_p X_p}{\lambda_1}$, which implies that X_1 is exactly linearly related to the other explanatory variables (Gujarati, 2003).

Near perfect multicollinearity

Consider a linear equation $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_p X_p + \upsilon_i = 0$, where $\lambda_1, \lambda_2 \ldots \lambda_p$ are constants and non-zero and υ_i is a stochastic error term. Let X_1 be a function of the other variables, then $X_1 = \frac{\lambda_2 X_2}{\lambda_1} + \cdots + \frac{\lambda_p X_p}{\lambda_1} + \frac{\upsilon_i}{\lambda_1}$, which implies that X_1 is linearly related to the other explanatory variables as well as to the error term. This relationship is not exact, but rather near perfect (Guajarati, 2003). Figure 4.1 demonstrates graphically the difference between perfect and less than prefect multicollinearity.



4.3.1 Diagnosis for Multicollinearity

Multicollinearity can be tested through correlation matrix, variance inflation factor and condition number.

Partial correlation matrix

Commands for generating a partial correlation matrix are embedded in many statistical software. For instance, in the SAS software, the 'CORR procedure' call for matrix can be generated once a researcher can decide on the explanatory variables that have high correlation. The correlations can either be positive or negative in nature. The partial correlation matrix is usually depicted by:

$$R_{XX} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

Thupeng *et al*. (2018) suggest that a partial correlation that is more than 80% is indicative of the presence of multicollinearity. The magnitude of this percentage is relative and subject to the interpretation of the researcher on what they deem to be 'high correlation'. This subjectivity and lack of scientific validation has limitations, as outlined by Daoud (2017) and echoed by Thompson *et al.* (2017).

Variance Inflation Factor

The variance inflation factor (VIF) measures the correlation between explanatory variables. The value measure and quantify how much the variance is inflated by the correlation (O'Brien, 2007). The VIF is calculated as $VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination. A VIF value of 1 indicates no correlation among the explanatory variables, thus no multicollinearity, $1 < \text{VIF} \leq 5$ indicates moderate correlation, and VIF>5 confirms high correlation, a definite confirmation of multicollinearity problem.

Thompson *et al.* (2017), stipulates that it is easier to interpret $\sqrt{VIF_j}$ as the square root depicts the approximate measurement of how large the standard error of $\widehat{B_j}$ is compared to when it would have not been correlated to other variables. For example, suppose $VIF_j = 9$, then $\sqrt{VIF_j} = 3$, meaning that the standard error of $\widehat{B_j}$ is three times larger than if there was no collinearity with other explanatory variables.

Condition Number

Condition number (CN) tests for multicollinearity and incorporates the use of eigenvalues of the regression model. Consider **X'X**, a p x p matrix of sums of square and cross product of the explanatory variables. The **X'X** associated with the explanatory variables is used to compute the eigenvalues. Let λ denote the eigenvalues, where the condition number is calculated as:

$$CN = \sqrt{\frac{max_p(\lambda)}{min_p(\lambda)}},$$

where the $min_p(\lambda)$ is the minimum eigenvalue and $max_p(\lambda)$ is the maximum eigenvalue (Thompson *et al.*, 2017). The $CN < 10$ indicates no multicollinearity, $10 < CN < 30$ implies an acceptable multicollinearity and $CN > 30$ shows strong multicollinearity among the independent variables (Sinan and Alkan, 2015). In some cases, $CN \geq 15$ indicates presence of multicollinearity.

In all experiments in diagnosing multicollinearity, the researcher must be cognisant of any outliers that might exist in the data, especially since the dataset used in this paper involves macroeconomic data that is susceptible to outliers. Sinan and Alkan (2015) argue that outliers reduce the reliability of the diagnostics measure such as variance inflation factor (VIF) and condition number (CN). A robust method called minimum covariance determinant (MCD) that aims to improve and obtain robust versions of VIF and CN – that can be used to determine multicollinearity in the presence of outliers (Sinan and Alkan, 2015)

Three robust methods namely ordinary least squares, ridge regression and principal component regression analysis were compared in respect of addressing the problem of multicollinearity. Understanding the extent to which the multicollinearity problem is controlled, assist in the formulation of a predictable model for NPLs. The depicted trends in such a model are predicted using adjusted macroeconomic variables, which are quantitative in nature.

The initial phase in analysis involved descriptive statistical analysis to get a better understanding of the variables. The preliminary analysis provides summary statistics such as the central tendency, measure of dispersion, boxplots, etc. of each variable. With the central tendency, the mean, median, minimum, and maximum values of the fourteen variables were obtained. With the measure of dispersion, variances, standard deviation and standard errors which measure the spread of the data around the mean were obtained. The computation used the statistical software, SAS (SAS version 9.4).

4.4 Analysis of the Nigeria portfolio data

4.4.1 Data exploration – Nigeria portfolio data

The data exploration considered the quantitative variables from the non-performing loans ratio (dependent variable) to obtain summary statistics, which are presented in Table 1. Table 1 presents the means, minimum, maximum, and standard deviation of 14 variables, each with 84 observations. The dependent variable is non-performing loans ratio.

Table 4.1 Summary statistics of non-performing loans' data for Nigeria

Variable	Number of observations	Mean	Standard Deviation	Minimum	Maximum
Non-performing loans (NPL ratio) – Dependent variable	84	8.307	3.098	2.770	16.660
Gross domestic product (GDP)	84	4.529	3.327	-1.700	8.400
Crude oil (Crude)	84	88.044	28.807	30.660	128.280
Treasury bill (Tbill)	84	9.720	3.559	1.040	15.500
Interbank call rate (Icr)	84	11.400	7.075	0.770	36.642
Lending rate (Lrate)	84	16.769	0.698	15.730	19.905
Deposit rate (Drate)	84	7.792	1.742	4.100	12.240
M2 money supply (M2)	84	12.854	7.435	-2.300	27.760
Exchange rate (FX)	84	175.877	41.844	149.780	312.125
FX reserves (FXR)	84	35.440	6.262	24.000	48.800
M1 money supply (M1)	84	11.969	13.808	-10.400	49.950
Inflation (Infl)	84	11.151	2.961	7.770	18.855
Monetary policy rate (MPR)	84	10.954	2.519	6.000	14.400
Maximum lending rate (MLR)	84	24.731	1.883	21.750	28.855

The findings in Table 4.1 provide a snapshot of the structure of the Nigeria portfolio data, including the mean and standard deviation of the variables. The lending rate variable has the smallest standard deviation, indicative of high precision as the observations are in close proximity to the mean. Conversely, the exchange rate and Treasury bill variables exhibited relatively large standard deviations, thus less precision as values are farther away from the mean. In addition, the summary statistics of the variables are presented in a boxplot which

assists in assessing skewness and variability depicted by the data. Figure 4.2 presents the boxplots for each variable described in Table 4.1.

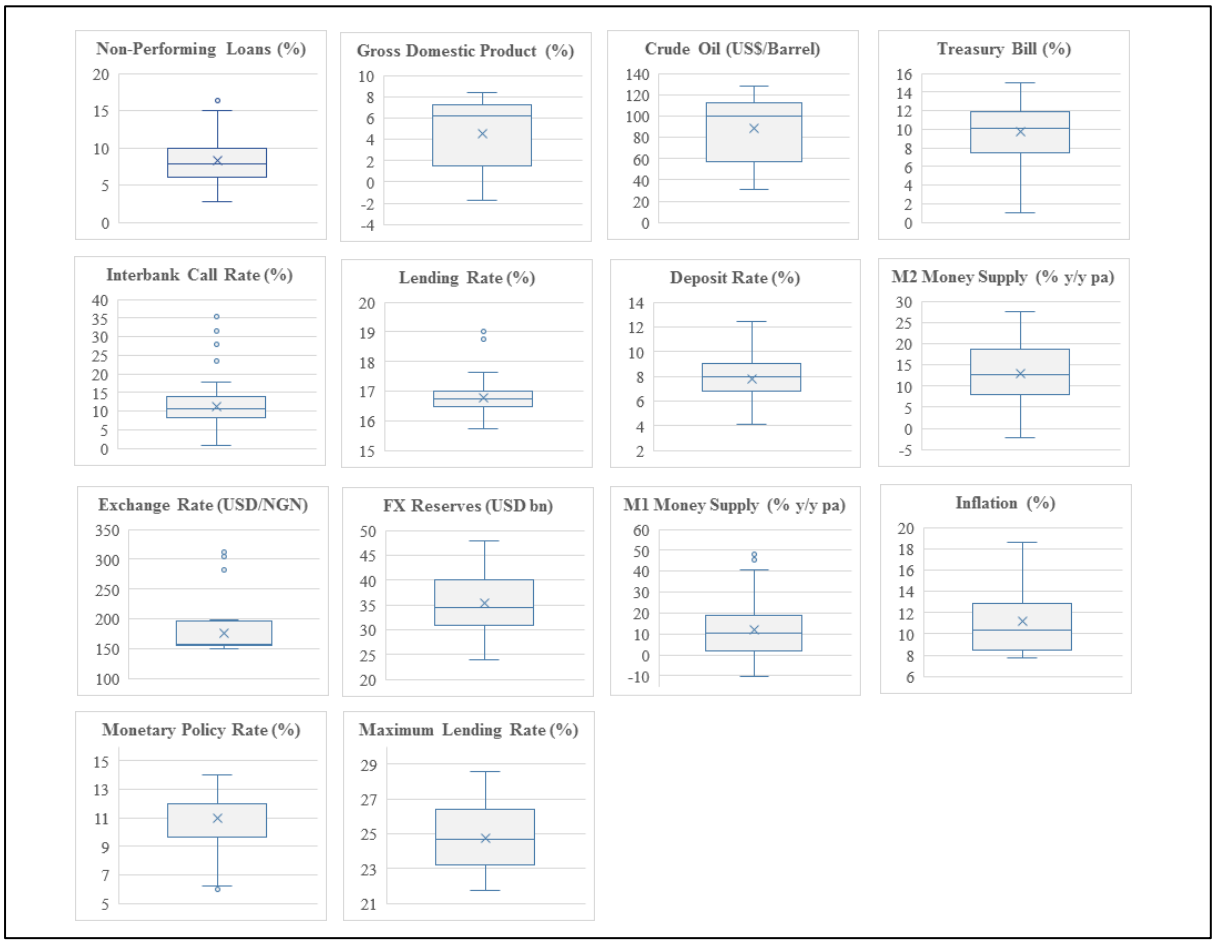


Figure 4.2 Boxplots of the variables in the Nigeria portfolio data (Dependent variable is Non-Performing Loans)

Using the information displayed by the boxplot, a summary in Table 4.2, in terms of variability, skewness and outliers is generated.

Table 4.2 Variability, skewness and presence of outliers as displayed in the boxplot for each variable

Variable	**Nature of variability**	**Nature of skewness**	**Presence of outliers**
Non-performing loans (NPL ratio) – Dependent variable	Small	Symmetric	Few
Gross domestic product (GDP)	Large	Negatively skewed	None
Crude oil (Crude)	Large	Negatively skewed	None
Treasury bill (Tbill)	Small	Symmetric	None
Interbank call rate (Icr)	Small	Symmetric	Many
Lending rate (Lrate)	Small	Symmetric	Few
Deposit rate (Drate)	Small	Symmetric	None
M2 money supply (M2)	Small	Symmetric	None
Exchange rate (FX)	Small	Symmetric	Few
FX reserves (FXR)	Small	Symmetric	None
M1 money supply (M1)	Small	Symmetric	Few
Inflation (Infl)	Small	Symmetric	None
Monetary policy rate (MPR)	Small	Symmetric	Few
Maximum lending rate (MLR)	Small	Symmetric	None

Figure 4.2 and Table 4.2 aided in identifying variables that had either small or large variation, were skewed or symmetric, and the presence of outliers. The variables showing presence of outliers are considered with care during the modelling as the results may be distorted. The variables that exhibited outliers are non-performing loans, interbank call rate, lending rate, exchange rate, M1 money supply and monetary policy rate.

Once, summary statistics were obtained and discussions around the results made, the next process involved the establishment of the adherence of the assumptions of the ordinary least squares, such as autocorrelation, stationarity, and normality. Some of the data was extrapolated due to the restricted availability of the data. For instance, suppose the quarterly inflation rate is

4%, then the inflation rate of the three months within that quarter is taken to be 4% as not much change is expected.

4.4.2 Checking the validity of the assumptions

Though there are various tests for the assumptions, we consider only the autocorrelation, stationarity, and normality. Table 4.3 (a-b) presents the testing results.

Table 4.3 The Durbin-Watson, stationarity and normality test using Nigeria portfolio data

(a) Durbin-Watson Statistic	Order 1 = 1.506		Order 2 = 2.139		Order 3 = 2.018		Order 4 = 2.030	
	Pr<DW = 0.000		Pr<DW = 0.435		Pr<DW = 0.358		Pr<DW = 0.460	
	Pr>DW = 1.000		Pr>DW = 0.565		Pr>DW = 0.642		Pr>DW = 0.540	
(b) Stationarity test		Lags	Rho	Pr<Rho	Tau	Pr<Tau	F	Pr>F
	Zero Mean	0	-1.162	0.444	-1.180	0.214		
		1	-1.133	0.448	-1.320	0.172		
	Single Mean	0	-6.441	0.302	-2.320	0.167	2.790	0.369
		1	-6.205	0.320	-2.480	0.125	3.210	0.262
	Trend Mean	0	-8.402	0.533	-3.140	0.104	8.690	0.002
		1	-8.421	0.531	-3.970	0.013	14.210	0.001
(c) Normality test	Variable		W Statistic	Pr<W	Variable		W Statistic	Pr<W
	Non-performing loans (NPL ratio) – Dependent variable		0.961	0.012	M2 money supply (M2)		0.978	0.163
	Gross domestic product (GDP)		0.857	<0.000	Exchange rate (FX)		0.585	<0.000
	Crude oil (Crude)		0.874	<0.000	FX reserves (FXR)		0.971	0.059
	Treasury bill (Tbill)		0.940	0.001	M1 money supply (M1)		0.941	0.001
	Interbank call rate (Icr)		0.872	<0.000	Inflation (Infl)		0.898	<0.000
	Lending rate (Lrate)		0.860	<0.000	Monetary policy rate (MPR)		0.761	<0.000
	Deposit rate (Drate)		0.966	0.026	Maximum lending rate (MLR)		0.944	0.001

The Durbin-Watson test for autocorrelation assumption for ordinary least squares is conducted to establish if there exists any relationship between the error terms. Table 4.3(a) presents the Durbin-Watson Statistic (DW) results, which tests the following:

H_0: Error terms are uncorrelated

H_1: There exist correlations amongst the error terms

Using the benchmark of 2, whereby if the DW value is closer or equal to 2, then the null hypothesis cannot be rejected. Then, from Table 4.3(a) above, the Durbin Watson value for the first order is 1.5, one might argue that this is indicative of serial correlations among the error terms. The null hypothesis cannot be rejected in orders above 1 as the values are closer to 2, hence the errors are uncorrelated.

Table 4.3(b) present testing results for stationarity. It tests if the data does not fluctuate with time or is not dependent on time. It tests data that is,

H_0: not stationary

H_1: stationary

Using results in Table 4.3(b), and at 5% significance level, we fail to reject H_0 for the single mean (p-value=0.262) and conclude non-stationarity. We reject H_0 for the trend at 0 lag (p-value=0.002) and first difference (p-value=0.001) and conclude that the data is stationary. For Rho and Tau, the p-values are all more than 0.05, which indicates that they are not significant.

Table 4.3(c) presents the Shapiro-Wilk test results for the normality assumption. Using the Proc Univariate function in SAS, we obtain the test statistics and p-values for the 14 variables. The test states,

H_0: Data is normally distributed

H_1: Data is not normally distributed

Using the $p - value$, at 5% significance level, we reject the null hypothesis for all variables except for M2 money supply (M2) and FX reserves (FXR). We conclude that most of the variables are not normally distributed. This means that the probability distributions of the variables are unknown.

In conclusion, the dataset does not have autocorrelation. This means that estimation of the parameters is likely to be precise as the confidence intervals are likely to be narrower. In addition, the dataset is proven to be stationary, as it does not change over time. In terms of the normality assumption, most of the variables were not normally distributed, hence the probability distributions of these would be unknown, making the hypothesis testing challenging.

4.4.3 Testing for Correlations – Pearson Correlation Coefficient

Table 4.4 presents Pearson correlation coefficients and associated p-values for testing for correlation among the 13 variables using the Nigeria portfolio data.

Table 4.4 Testing for Correlations – Nigeria portfolio data

Pearson Correlation Coefficients, N = 84													
	GDP	Crude	Tbill	Icr	Lrate	Drate	M2	FX	FXR	M1	Infl	MPR	MLR
GDP	1.000												
Crude	0.752	1.000											
Tbill	-0.195	0.305	1.000										
Icr	-0.328	-0.026	0.628	1.000									
Lrate	0.026	-0.257	-0.311	-0.171	1.000								
Drate	-0.313	-0.108	0.350	0.229	0.373	1.000							
M2	0.073	-0.012	-0.042	0.151	0.185	0.050	1.000						
FX	-0.809	-0.666	0.255	0.498	0.106	0.136	0.117	1.000					
FXR	0.645	0.657	-0.018	-0.204	0.134	0.184	0.107	-0.658	1.000				
M1	-0.376	-0.292	0.074	0.237	-0.021	-0.341	0.423	0.579	-0.502	1.000			
Infl	-0.181	-0.349	-0.080	0.128	0.338	-0.229	0.325	0.568	-0.443	0.740	1.000		
MPR	-0.663	-0.175	0.750	0.553	-0.134	0.591	-0.080	0.490	-0.139	0.034	-0.207	1.000	
MLR	-0.906	-0.640	0.256	0.307	0.113	0.569	-0.065	0.700	-0.361	0.126	-0.013	0.776	1.000

In Table 4.4, the correlations between all the variables are depicted, where several variables are highly correlated with one another. In observation, gross domestic product (GDP) and maximum lending rate (MLR) are negatively correlated at 90%, while with exchange rate (FX), the correlation stands at negative 80%. For the three variables mentioned above, the correlations are negative, meaning that when GDP increases, the MLR and FX move in the opposite direction, hence the observed decrease. Since these are independent variables, near perfect multicollinearity might be present. For some variables like maximum policy rate (MPR) and maximum lending rate (MLR), the correlation is rather positive at 78%. The MLR also shows some positive correlation with FX at 70%, while GDP and crude oil (Crude) also exhibit

a positive correlation of 75%. The Treasury bill (Tbill) is positively correlated with MPR at 75%, while M1 money supply (M1) and inflation (Infl) shows a correlation of 74%.

The other variables that have correlations in the 60% range include FXR and GDP at 65%, while GDP exhibits a negative correlation of 66% with MPR. Crude is the other variable that displays a negative correlation of 67% and 64% with FX and MLR respectively, but positively correlated with FXR at 66%. Treasury bill (Tbill) also displays a positive correlation with interbank call rate (Icr). It should be noted that correlation does not imply causality, particularly from observation without more information or data to support the notion.

Model Fitting

4.4.4 Ordinary Least Squares using Nigeria portfolio data

Using the 14 variables and non-performing loans as the dependent variable, an ordinary least square model was fitted. Table 4.5 presents the analysis results which include analysis of variance (ANOVA), R-square, estimates of regression model and variance inflation values.

Table 4.5 Ordinary least squares output – Nigeria portfolio data

Analysis of Variance (Nigeria portfolio data)					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	13	726.030	55.849	55.560	<.0001
Error	70	70.368	1.005		
Corrected Total	83	796.399			

Root MSE	1.003	R-Square	0.912
Dependent Mean	8.307	Adj R-Sq	0.895
Coeff Var	12.070		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-21.410	7.075	-3.030	0.004	0.000
Gross domestic product (GDP)	1	-0.264	0.193	-1.360	0.177	34.164
Crude oil (Crude)	1	-0.039	0.011	-3.720	0.000	7.639
Treasury bill (Tbill)	1	0.029	0.124	0.240	0.814	16.040
Interbank call rate (Icr)	1	0.015	0.025	0.600	0.553	2.603
Lending rate (Lrate)	1	0.814	0.288	2.830	0.006	3.338
Deposit rate (Drate)	1	0.246	0.153	1.610	0.113	5.893
M2 money supply (M2)	1	-0.048	0.022	-2.190	0.032	2.236
Exchange rate (FX)	1	0.027	0.010	2.740	0.008	13.707
FX reserves (FXR)	1	0.247	0.040	6.120	<.0001	5.265
M1 money supply (M1)	1	0.005	0.022	0.240	0.810	7.277
Inflation (Infl)	1	0.182	0.108	1.680	0.098	8.486
Monetary policy rate (MPR)	1	-1.109	0.234	-4.740	<.0001	28.662
Maximum lending rate (MLR)	1	0.630	0.310	2.030	0.046	28.128

From an economic point of view, the fitted model (full model), irrespective of the significance status of the parameters, has the form:

$$\boldsymbol{NPL} = -21.410 - 0.264\boldsymbol{GDP} - 0.039\boldsymbol{Crude} + 0.029\boldsymbol{TBill} + 0.015\boldsymbol{Icr} + 0.814\boldsymbol{Lrate} + 0.246\boldsymbol{Drate} - 0.048\boldsymbol{M2} + 0.027\boldsymbol{FX} + 0.247\boldsymbol{FXR} + 0.005\boldsymbol{M1} + 0.182\boldsymbol{Infl} - 1.109\boldsymbol{MPR} + 0.630\boldsymbol{MLR} \quad (1)$$

The results in Table 4.5 indicate that the model is significant at 5% significance level (p-value<0.0001). The Adjusted R^2 is at 0.895, denoting that at least 89% of the total variability in the dependent variable (NPLs) is explained by the independent variables. The $MSE \cong 1.005$, indicating the absolute value of the data points to the regression line. The parameter of estimates indicated that the intercept term, crude oil price, lending rate, M2 money supply, exchange rate, FX reserves, monetary policy rate and maximum lending rate were all significant. The rest of the variables were not significant. Taking the crude oil price for instance, for every unit change in price of crude oil, the NPL decreases by 0.04, holding other variables constant. For the lending rate, a unit increase corresponds with a 0.8 increase in the NPL value, holding other variables constant.

The variance inflation factor (VIF) helps in determining the variable that may be involved in multicollinearity. Several methods can be employed to determine the arbitrary value that can be used as a measure. The mathematical measurement is determined by $1/(1 - R^2)$, whereby this would give 11.364. The variables that have a VIF value that is more than 11.364 are gross domestic product (GDP), Treasury bill (Tbill), exchange rate (FX), FX reserves (FXR), monetary policy rate (MPR) and maximum lending rate (MLR), indicating the presence of multicollinearity. The effect of these high VIF values make most of the parameter estimates unstable and not reliable even though some are significant.

The instability of the parameter estimates of ordinary least squares, as revealed by the VIF of some variables, gives rise to explore other analysis that can combat multicollinearity, such as ridge regression and principal component analysis.

Table 4.6 depicts the reduced fitted model that takes into account the variables that are significant at 5% significance level. This comprises of a reduced number of variables from 13 independent variables down to 6.

Table 4.6 Ordinary least squares output – Nigeria portfolio data: fitted model

Analysis of Variance					
Source	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	687.777	114.629	81.260	<.0001
Error	77	108.622	1.411		
Corrected Total	83	796.399			

Root MSE	1.188	R-Square	0.864
Dependent Mean	8.307	Adj R-Sq	0.853
Coeff Var	14.298		

Parameter Estimates						
Variable	**Degrees of Freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-15.867	3.782	-4.200	<.0001	0.000
Crude Oil (Crude)	1	-0.067	0.007	-9.240	<.0001	2.531
Lending rate (Lrate)	1	1.212	0.223	5.440	<.0001	1.424
M2 money supply (M2)	1	-0.043	0.019	-2.260	0.027	1.152
Exchange rate (FX)	1	0.047	0.006	7.990	<.0001	3.491
FX reserves (FXR)	1	0.245	0.036	6.860	<.0001	2.933
Monetary policy rate (MPR)	1	-0.602	0.067	-9.010	<.0001	1.670

A reduced final fitted model becomes:

$$\boldsymbol{NPL} = -15.876 - 0.067\boldsymbol{Crude} + 1.212\boldsymbol{Lrate} - 0.043\boldsymbol{M2} + 0.040\boldsymbol{FX} + 0.245\boldsymbol{FXR} - 0.602\boldsymbol{MPR} \quad (2)$$

The results depicted in Table 4.6 illustrate that the reduced final model is significant at 5% significance level (p-value<0.0001). The Adjusted R^2 is at 0.853 (85.3%), while the $MSE \cong 1.411$, indicating the absolute value of the data points to the regression line. The significant variables are crude oil, lending rate, M2 money supply, exchange rate, FX reserves and monetary policy rate. To check if multicollinearity is a factor in the results attained, we use the mathematical criterion of $1/(1-R^2)$=7.331. All the variance inflation factors range between a minimum of 1.424 and a maximum of 2.933. All these values are below 7.331, indicative of no problem of multicollinearity.

To further understand if the OLS fitted model is robust in predicting the Nigeria NPLs, a plot of the residuals against the predicted values of the NPLs is analysed as seen below in Figure 4.3.

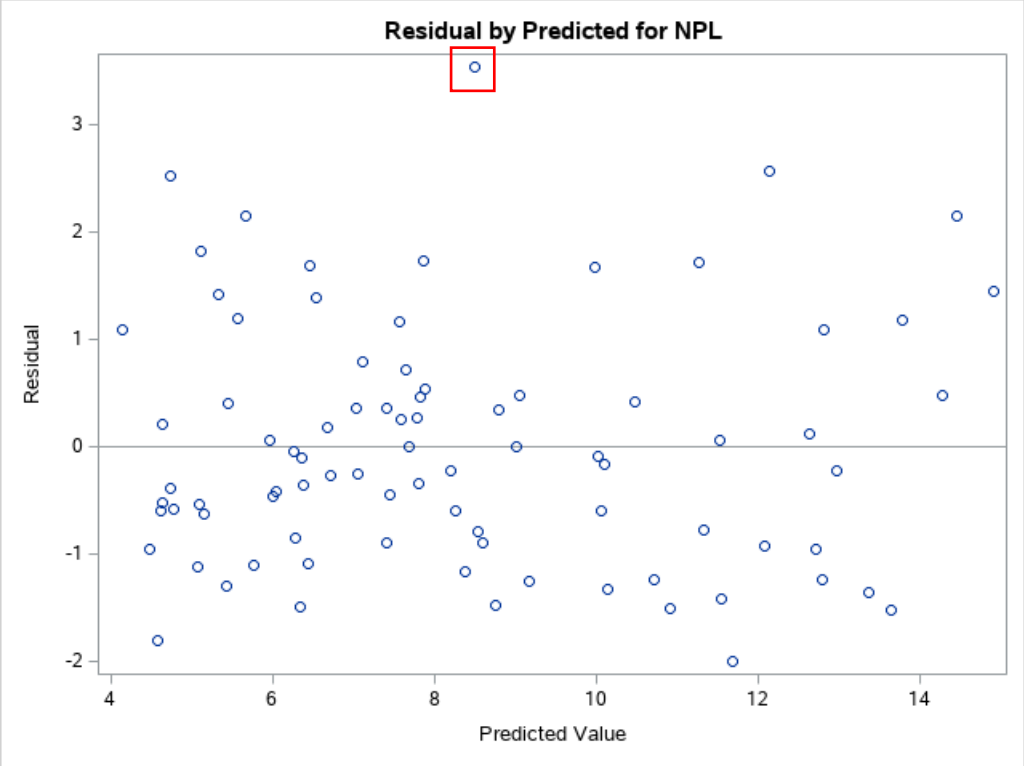


Figure 4.3 Residual plot of the fitted OLS model for the Nigeria portfolio

The x-axis on the scatterplot shows the predicted values of the NPLs, while the residuals are on the y-axis.

Figure 4.3 shows that the residuals are random, suggesting homoscedasticity, that is, constant variance of the residuals. One of the assumptions of classical ordinary squares is that the residuals should have constant variance, hence in this instance this is upheld (Gujarati, 2003). The red box outlines some notable outliers observed. Further analysis is shown on the plots of the residuals against the individual variables as shown in Figure 4.4.

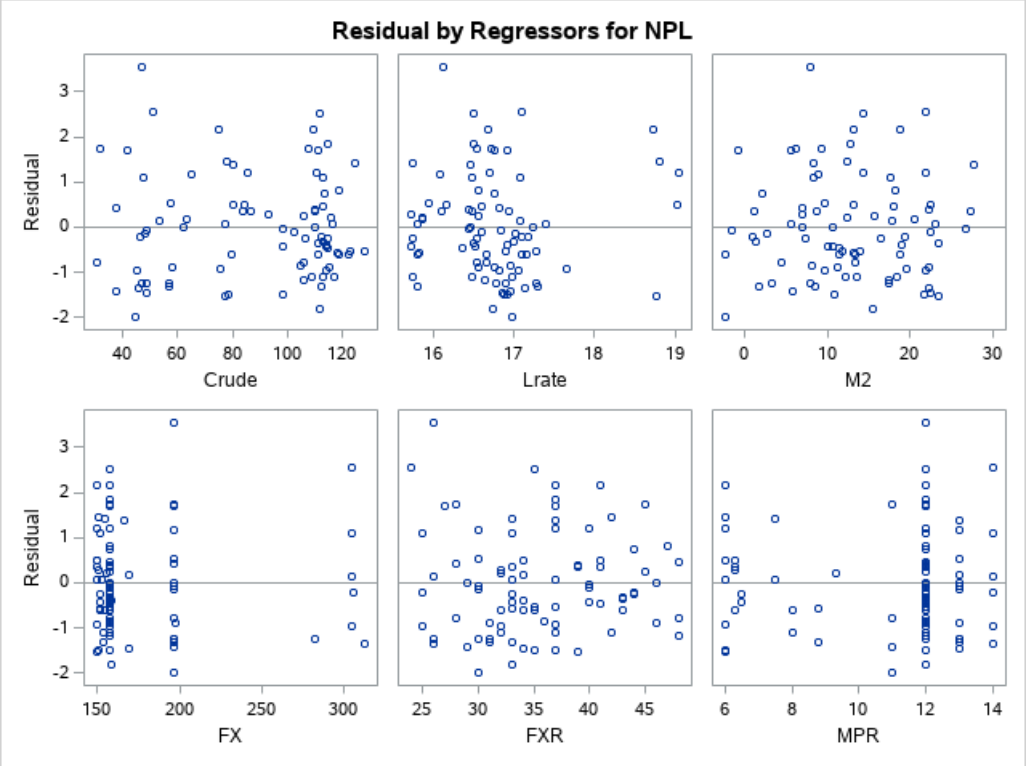


Figure 4.4 Residual by the individual independent variables of the fitted OLS model for the Nigeria portfolio

Figure 4.4 show that for crude oil, M2 money supply and FX reserves, there is a random pattern which depicts that there is no correlation between the residuals and the associated independent variable. For exchange rate (FX), the pattern is more concentrated between the 150 and 200 range and resembles a quadratic relation with the residuals. For the lending rate, the concentration is on the 16 to 17 range, with some outliers. The monetary policy rate is more concentrated on the 12 to 14 range but has fewer data points below 10, also showing a quadratic relationship with the residuals.

In conclusion, there are notable outliers on the scatterplots, but these do not necessarily affect the predictability of the model, where homoscedasticity is concerned.

4.4.5 Variable selection (post multicollinearity) – Nigeria portfolio data

Variable selection for ordinary least squares is about removal of variables that are causing multicollinearity within the data. There are many ways of conduction variable selection of variables. In this instance, variable selection entails removing the variable that is highly collinear with another variable and has the least correlation with the dependent variable. From the results, the two variables that are omitted are gross domestic product (GDP) and maximum policy rate (MPR). The resultant effect is depicted in Table 4.7.

Table 4.7 is an illustration of the results attained by removing some variables that exhibit higher variance inflation factors. This comprises of 8 variables, down from the original 13.

Table 4.7 Ordinary Least Squares post variable selection – Nigeria portfolio data

Analysis of Variance					
Source	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	632.831	79.104	36.270	<.0001
Error	75	163.567	2.181		
Corrected Total	83	796.399			

Root MSE	1.477	R-Square	0.795
Dependent Mean	8.307	Adj R-Sq	0.773
Coeff Var	17.778		

Parameter Estimates						
Variable	**Degrees of Freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-12.178	4.963	-2.450	0.017	0.000
Crude oil (Crude)	1	-0.084	0.009	9.660	<.0001	2.397
Interbank call rate (Icr)	1	-0.028	0.027	1.020	0.310	1.430
Lending rate (Lrate)	1	0.984	0.353	2.790	0.007	2.310
Deposit rate (Drate)	1	-0.018	0.128	-0.140	0.889	1.881
M2 money supply (M2)	1	-0.039	0.027	-1.420	0.159	1.582
FX reserves (FXR)	1	0.212	0.046	4.560	<.0001	3.222
M1 money supply (M1)	1	-0.004	0.022	-0.170	0.864	3.579
Inflation (Infl)	1	0.438	0.102	4.280	<.0001	3.494

Based on Table 4.7, it is evident that the model was significant, as indicated by the $p - value$ that is less than 0.001, while the $MSE \cong 2.181$. The $Adjusted\ R^2$ is at 77.3%, showing that circa 77.3% of the variability in the dependent variable is explained by the independent variables. From the parameter estimates, the independent variables that were significant at 5% significance level, were crude oil, lending rate, FX reserves and inflation.

Table 4.8 depicts the reduced fitted model of the OLS post variable selection, depicting only significant variables. The number of independent variables is further reduced from 8 to only 4.

Table 4.8 Ordinary Least Squares post variable selection – Nigeria portfolio data: fitted model

Analysis of Variance					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	621.332	155.333	70.100	<.0001
Error	79	175.067	2.216		
Corrected Total	83	796.399			

Root MSE	1.489	R-Square	0.780
Dependent Mean	8.307	Adj R-Sq	0.769
Coeff Var	17.921		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-13.355	4.453	-3.000	0.004	0.000
Crude oil (Crude)	1	-0.083	0.009	-9.710	<.0001	2.284
Lending rate (Lrate)	1	1.061	0.300	3.540	0.001	1.643
FX reserves (FXR)	1	0.199	0.043	4.600	<.0001	2.746
Inflation (Infl)	1	0.372	0.070	5.340	<.0001	1.593

Table 4.8 shows that the variable post selection of the model is significant (p-value=0.0001) at 5% significance level. The $MSE \cong 2.216$, while the $Adjusted\ R^2 = 0.769$.

On the other hand, using the table by Daoud (2017) as referenced in the Literature Review (Chapter 2), the variables that have a VIF value of more than 5 are highly collinear. From Table 4.8, none of the independent variables fit this criterion. This suggests that the remaining variables need to be further analysed to check if the results follow the expected logic, otherwise, multicollinearity might still be a problem.

The residual plots of the variable selection model are depicted in Figure 4.5, to identify any patterns that might hinder the robustness of the predicted variable.

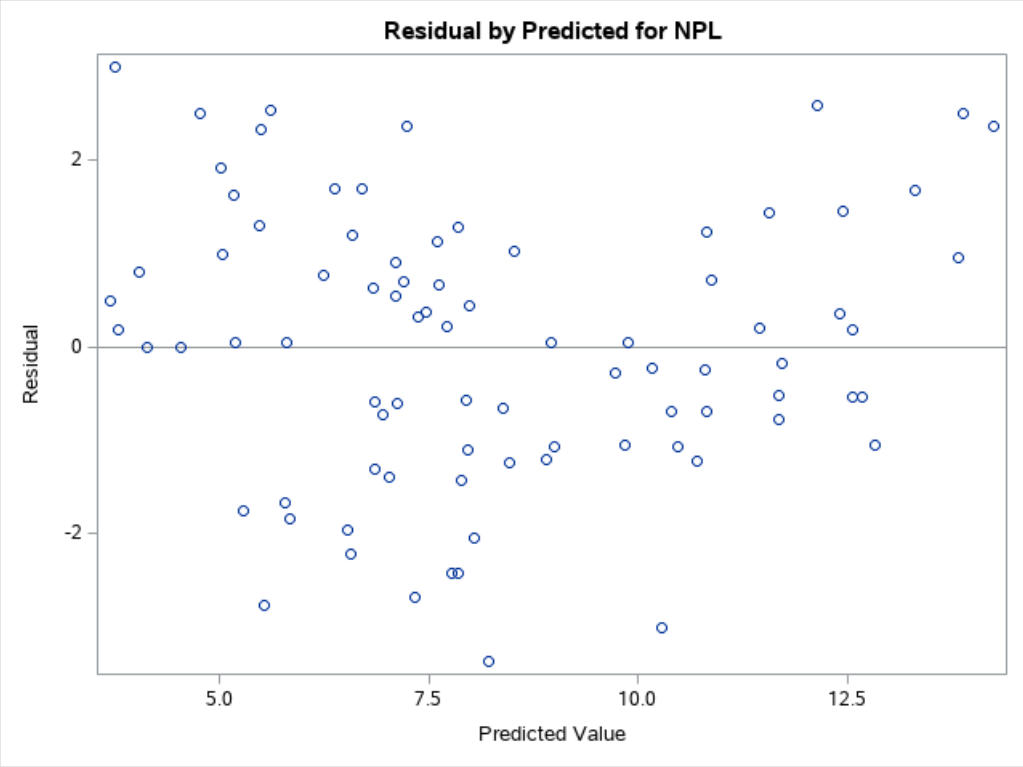


Figure 4.5 Residual plot of the OLS post variable selection of the Nigeria portfolio data

Figure 4.6 shows that the residuals are random, suggesting homoscedasticity, which is a constant variance of the residuals. The plots of the residuals against the individual variables in Figure 4.5 also shows that crude oil, FX reserves and inflation have random patterns, while the lending

rate random pattern is concentrated in the 16 to 17 range but with some outliers. In conclusion, there is no problem of heteroscedasticity.

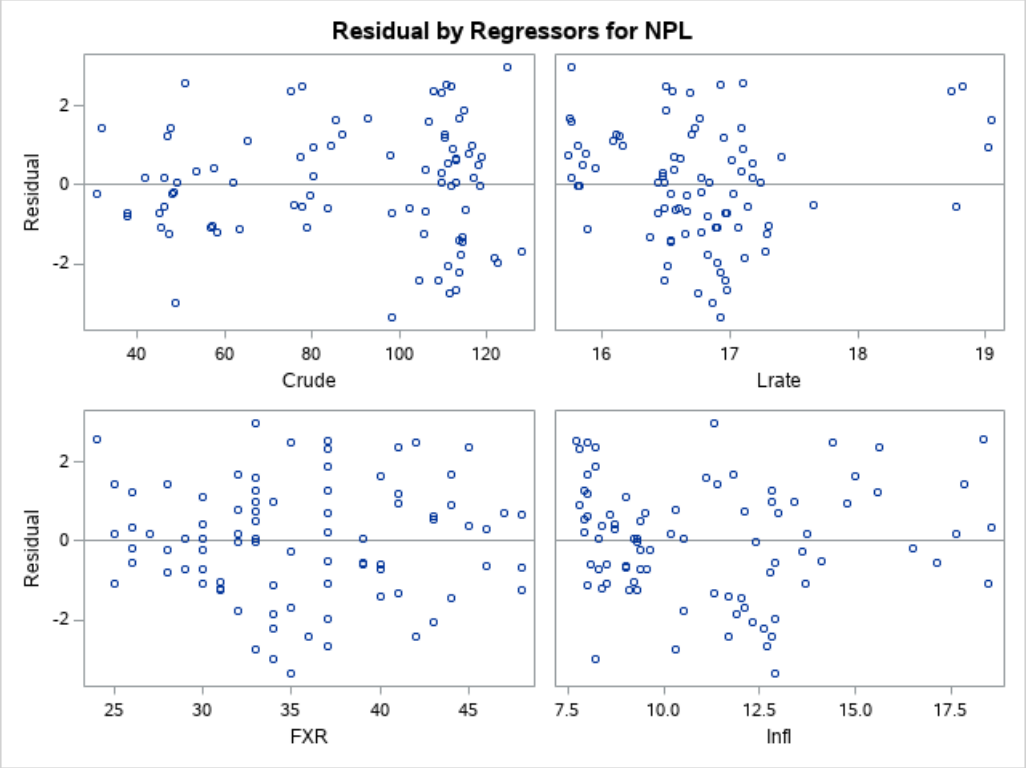


Figure 4.6 Residual plots by the individual independent variables for Nigeria portfolio

4.4.6 Ridge Regression – Nigeria portfolio data

The ridge regression attempts to find a value of K that can almost disintegrate the linear relationship between the variables by introducing some biasness to the model. The ideal way to find this value is through ridge trace, despite the subjectivity aspect of this method. Due to the high number of explanatory variables, SAS automatically divided the explanatory variables through two ridge trace plots as depicted in Figure 4.6a and 4.6b.

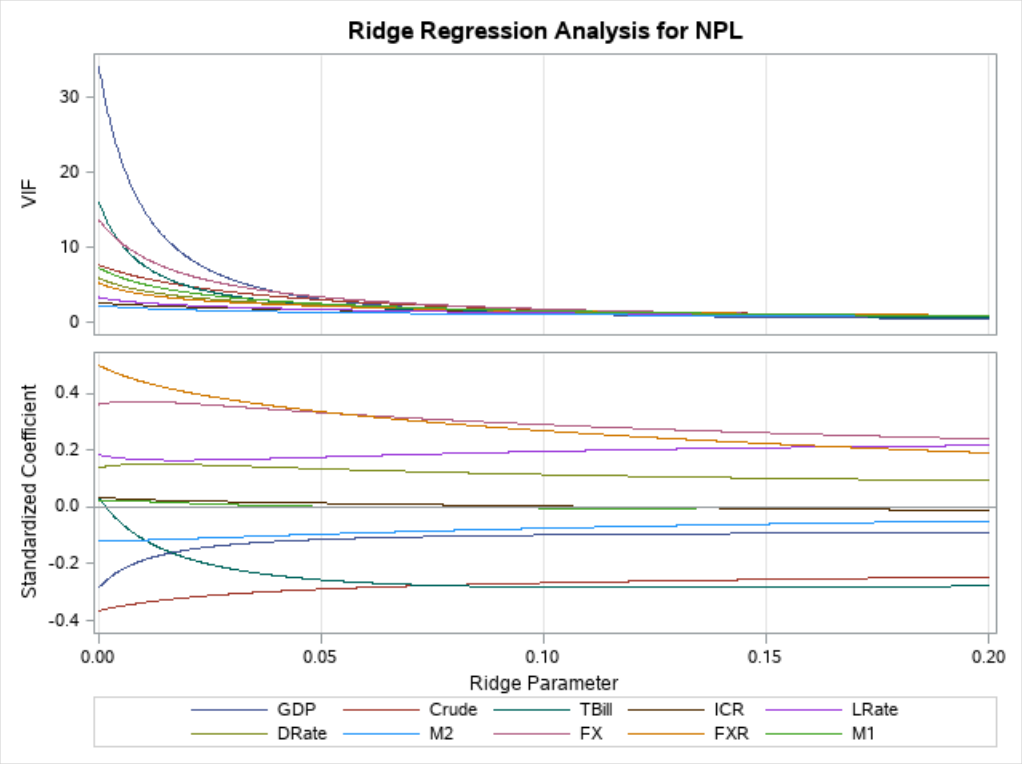


Figure 4.7 Ridge Trace a – for Nigeria portfolio

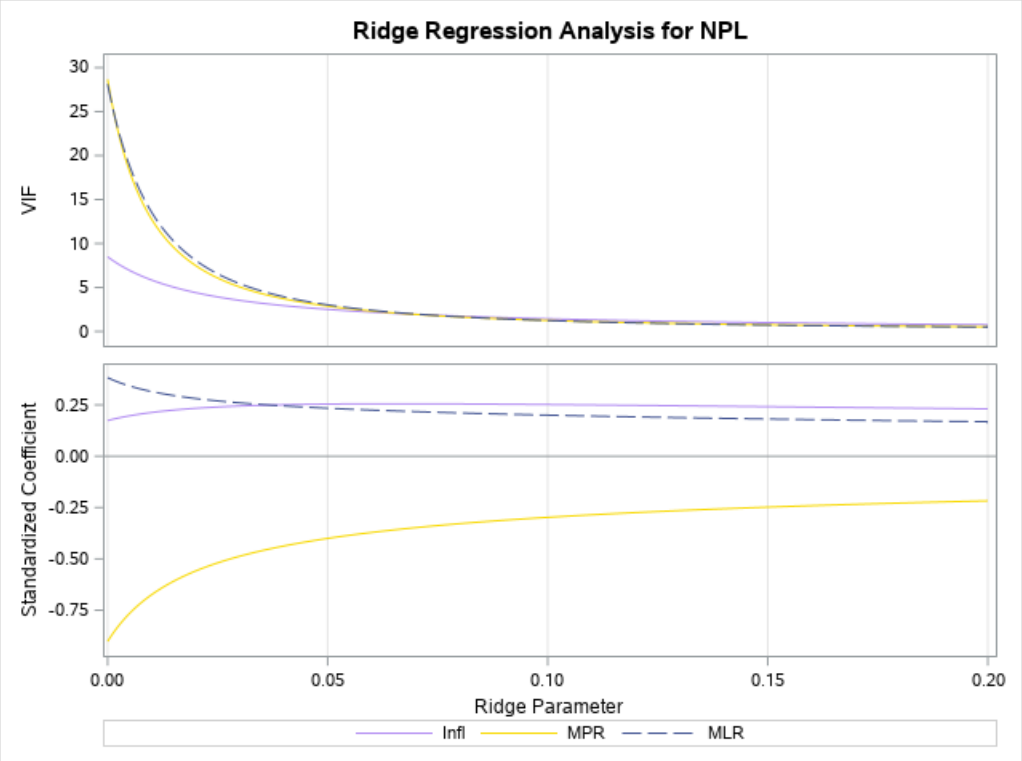


Figure 4.8 Ridge Trace b – for Nigeria portfolio

Figure 4.7 and Figure 4.8 show that the VIFs for most variables converged between 0 and 0.2. A closer look at the plots reveals that this value tends to sway more to being closer to 0.1, hence the subjective value of K chosen for this research paper would be 0.11.

Table 4.9 depicts the reduced model that only shows the significant variables, where the ridge regression parameter $K = 0.11$. The ridge parameter was chosen where the variance inflation factors are closer to 1. The table also gives a comparison when $K = 0$.

Table 4.9 Ridge Regression output – Nigeria portfolio data: fitted model

TYPE	RIDGE	RMSE	Intercept	Crude	Lrate	M2	FX	FXR	MPR	**RSQ**
PARMS	0.00	0.383	0.000	-0.619	0.273	-0.102	0.628	0.494	-0.490	0.864
SEB	0.00	0.384	0.042	0.067	0.050	0.045	0.079	0.072	0.054	
RIDGEVIF	0.11			1.329	0.929	0.856	1.459	1.320	0.985	
RIDGE	0.11	0.418	0.000	-0.511	0.326	-0.046	0.423	0.265	-0.352	
RIDGESEB	0.11	0.419	0.046	0.053	0.044	0.043	0.055	0.053	0.046	

Table 4.9 shows that at 5% significance level, the variables that are significant are crude oil, lending rate, M2 money supply, exchange rate, FX reserves and the monetary policy rate. The $R^2 \cong 0.864$, shows that most of the variability is explained by the variables, while $MSE \cong 0.864$. Following on the concept of correlation transformation, the standardized ridge regression coefficients have to be transformed back in order to attain the original variables. The regression coefficients attained would be transformed using the following formulae:

a) For the intercept: $\bar{y} - \sum_{j=1}^{n} \hat{\beta}_j \frac{S_y}{S_x} \bar{x}_j$

b) For the other coefficients: $\hat{\beta}_j \frac{S_y}{S_x}$, where the $\hat{\beta}_j$ represents the regression coefficients and $\frac{S_y}{S_x}$, $\bar{x}_j$ and $\bar{y}$ can be obtained from Table 1, that shows the descriptive data of the variables.

For the reduced fitted model, the ridge regression coefficients as depicted in Table 4.9 becomes:

$$\boldsymbol{NPL}^{*} = -0.511\boldsymbol{Crude}^{*} + 0.326\boldsymbol{Lrate}^{*} - 0.046\boldsymbol{M2}^{*} + 0.423\boldsymbol{FX}^{*} + 0.265\boldsymbol{FXR}^{*} - 0.352\boldsymbol{MPR}^{*} \quad (3)$$

Transforming the ridge regression coefficients to the original variables equation (1) becomes

$$\boldsymbol{NPL} = 5.585 - 0.055\boldsymbol{Crude} + 0.143\boldsymbol{LRate} - 0.019\boldsymbol{M2} + 0.031\boldsymbol{FX} + 0.131\boldsymbol{FXR} - 0.433\boldsymbol{MPR} \quad (4)$$

The model defined in equation (4) is the one used in the prediction of NPLs for ridge regression.

The residual plots also assist in gaining better understanding of the data as shown in Figure 4.9 and Figure 4.10.

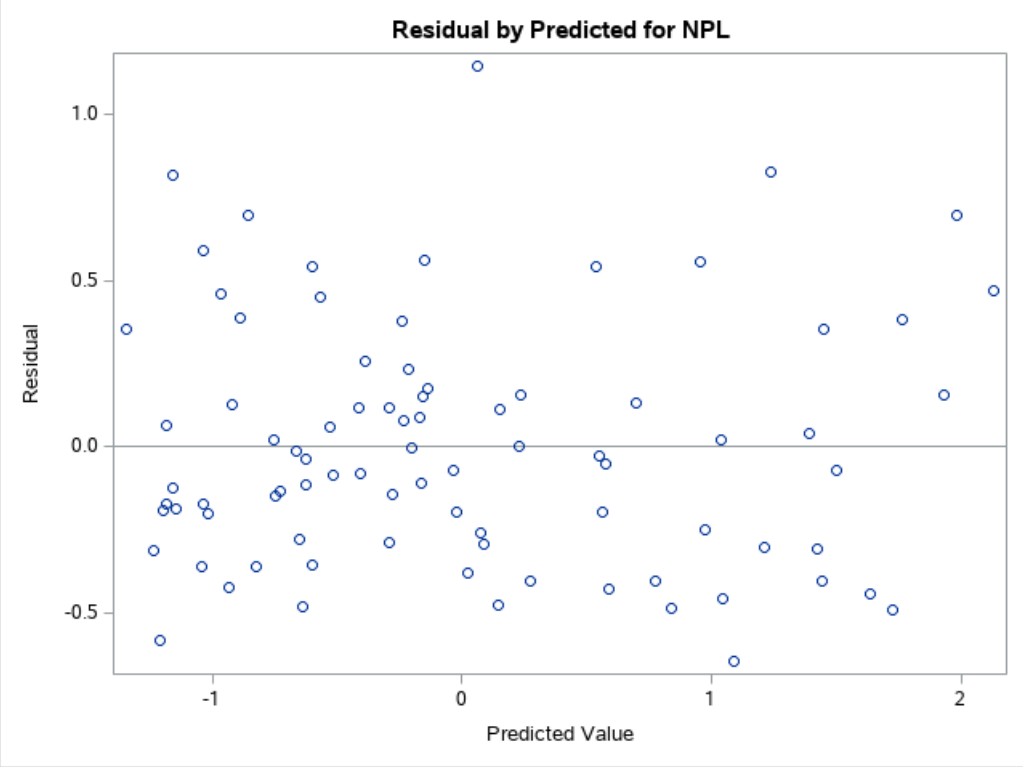


Figure 4.9 Residual plot of fitted ridge regression for Nigeria portfolio

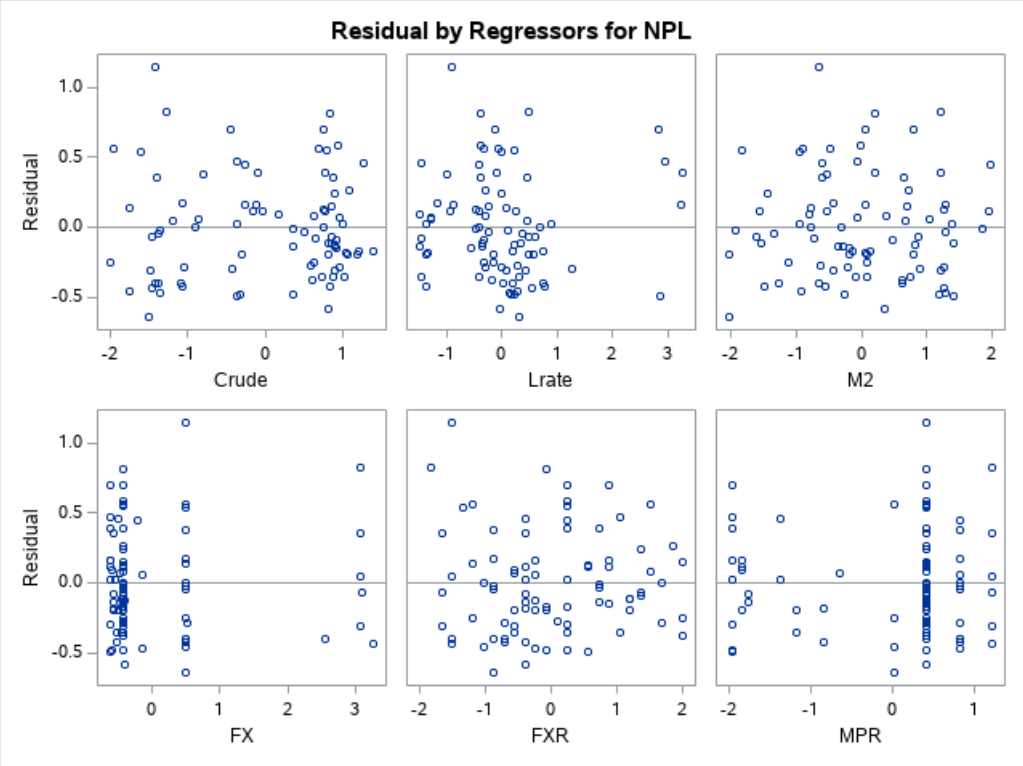


Figure 4.10 Residual plots of individual independent variables for fitted ridge regression for Nigeria portfolio

The residual plots of the fitted ridge regression are similar to those of the fitted ordinary least squares. Figure 4.9 show a random pattern between the residuals and the predicted values, implying no problem of heteroscedasticity, while the individual plots in Figure 4.10 also show that there are outliers within some variables.

4.4.7 Principal Component Analysis – Nigeria portfolio data

Principal component analysis on the Nigeria portfolio data led to the results presented in Table 4.10. Table 4.10 contains the eigenvalues for the 13 variables and proportion of variation accounted for by each of them. Principal component analysis was operated on a correlation matrix.

Table 4.10 Eigenvalues of the Correlation Matrix – Nigeria portfolio data

Eigenvalues of the Correlation Matrix (Nigeria portfolio data)				
PC	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
1	4.802	1.964	0.369	0.369
2	2.839	0.990	0.218	0.588
3	1.849	0.302	0.142	0.730
4	1.547	0.843	0.119	0.849
5	0.705	0.251	0.054	0.903
6	0.453	0.149	0.035	0.938
7	0.304	0.129	0.023	0.962
8	0.175	0.022	0.013	0.975
9	0.153	0.055	0.012	0.987
10	0.098	0.060	0.008	0.994
11	0.038	0.016	0.003	0.997
12	0.022	0.005	0.002	0.999
13	0.017		0.001	1.000

Using the Kaiser Criterion as previously stated in the Literature Review, it can be deduced from Table 4.10 that the eigenvalues that have values that are greater or equal to one account for approximately 85% of the variability in the data. The difference column also adds insights as it shows how much variance there is from one eigenvalue to the next. Between eigenvalue one and two, there is about 1.09 difference, however, as we go down the table, the difference between eigenvalue 4 and 5 is about 0.84.

The data in Table 4.10 is further verified by the scree plot (Figure 4.11) as the trend depicted starts to flatten between four and five, indicating the number of components to be retained as they are deemed independent of multicollinearity.

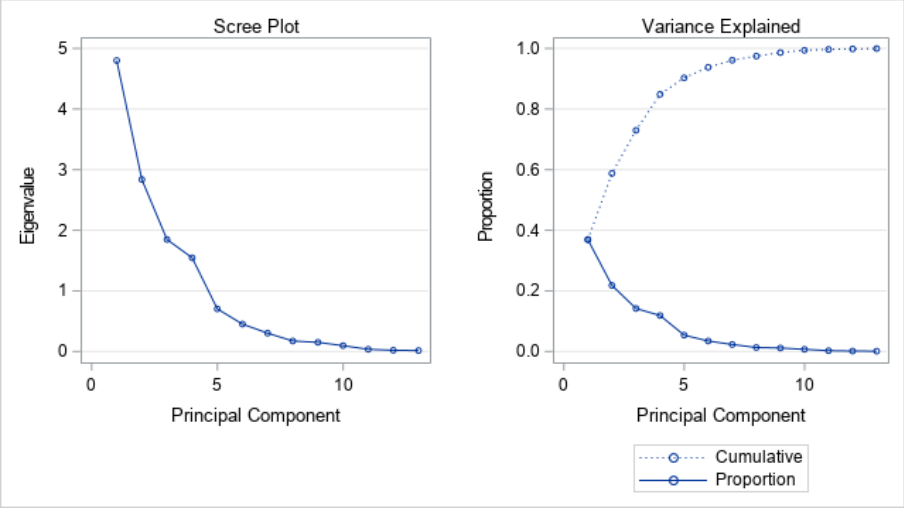


Figure 4.11 Scree Plot for Nigeria portfolio

Table 4.11 presents the eigenvectors for the four selected PCs accounting for 85% of the total variation. The dominating variables in each of the PC are bolded, based on a coefficient with a magnitude greater 0.3.

Table 4.11 Eigenvectors of the Nigeria portfolio data

Variable	**Prin1**	**Prin2**	**Prin3**	**Prin4**
Gross domestic product (GDP)	**-0.425**	-0.024	0.146	0.168
Crude oil (Crude)	**-0.318**	0.222	**0.364**	0.128
Treasury bill (Tbill)	0.168	**0.371**	**0.427**	0.081
Interbank call rate (Icr)	0.245	0.182	**0.395**	0.143
Lending rate (Lrate)	0.027	-0.137	**-0.419**	**0.531**
Deposit rate (Drate)	0.154	**0.372**	-0.291	**0.378**
M2 money supply (M2)	0.036	-0.178	0.190	**0.557**
Exchange rate (FX)	**0.423**	-0.126	0.050	0.022
FX reserves (FXR)	**-0.304**	0.248	-0.037	**0.350**
M1 money supply (M1)	0.229	**-0.364**	**0.335**	0.097
Inflation (Infl)	0.175	**-0.437**	0.154	0.245
Monetary policy rate (MPR)	**0.317**	**0.397**	0.069	0.029
Maximum lending rate (MLR)	**0.387**	0.179	-0.257	-0.016

The purpose of the eigenvectors in a nutshell is to depict the strength of the relationship between the principal component and the original independent variable. Worth noting is that

Table 4.11 depicts the eigenvectors of the data and does not include the dependent variable as the main reason for this exercise is to reduce the number of collinear independent variables in the data. From Table 4.11 it can be observed that the first principal component prin1, is an average measure that has large positive associations with exchange rate (42%), maximum lending rate (39%) and monetary policy rate (32%), while attaining large negative associations with gross domestic product (43%), crude oil (32%) and FX reserves (30%). PC2 measures the contrast of variables (Tbill (37%), Drate (37%), MPR (40%) against (M1 (36%), Infl (44%)). PC3 measures the average of variables (Crude (36%), Tbill (43%), Icr (40%), M1 (34%) against Lrate (42%)). PC4 measures the average of variables (Lrate (53%), Drate (38%), M2 (56%), FXR (35%)).

PC1 explains the most variability and the positive associations that are interpreted as the variables that move in the same direction with the dependent variable. In this instance, exchange rate, maximum lending rate and monetary policy rate would increase the NPLs, while gross domestic product, crude oil and FX reserves effects would decrease the NPLs. This applies to PC2 and to some extent PC3 and PC4. The positive associations increase the NPLs while negative associations tend to decrease the NPLs.

Table 4.12 demonstrate the analysis of variance (ANOVA) for the 4 principal components that account for 85% of the variability of the data. This data comprises of 13 variables and 84 observations.

Table 4.12 Principal Component Regression results – Nigeria portfolio data

Analysis of Variance					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	64.185	16.046	67.370	<.0001
Error	79	18.815	0.238		
Corrected Total	83	83.000			

Root MSE	0.488	R-Square	0.773
Dependent Mean	-1,56E-16	Adj R-Sq	0.762
Coeff Var	-3,13E+17		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	0.000	0.053	0.000	1.000	0.000
Prin1	1	0.171	0.024	7.010	<.0001	1.000
Prin2	1	-0.303	0.032	-9.540	<.0001	1.000
Prin3	1	-0.411	0.039	-10.440	<.0001	1.000
Prin4	1	0.193	0.043	4.490	<.0001	1.000

Table 4.12 depicts the regression analysis over the principal components attained from the data. The Analysis of Variance shows that model is significant, while the $MSE \cong 0.238$ and the associated $Adjusted\ R^2 = 0.762$ (76.2%). The table also illustrates that the principal components are all significant, with variance inflation factors all equal to 1.

$$\boldsymbol{NPL} = 0.171\boldsymbol{PC1} - 0.303\boldsymbol{PC2} - 0.411\boldsymbol{PC3} + 0.193\boldsymbol{PC4} \quad (5)$$

The residual analysis for the principal components also shows that the variance of the residuals is constant and not correlated with the predicted values, as shown in Figure 4.12 and Figure 4.13.

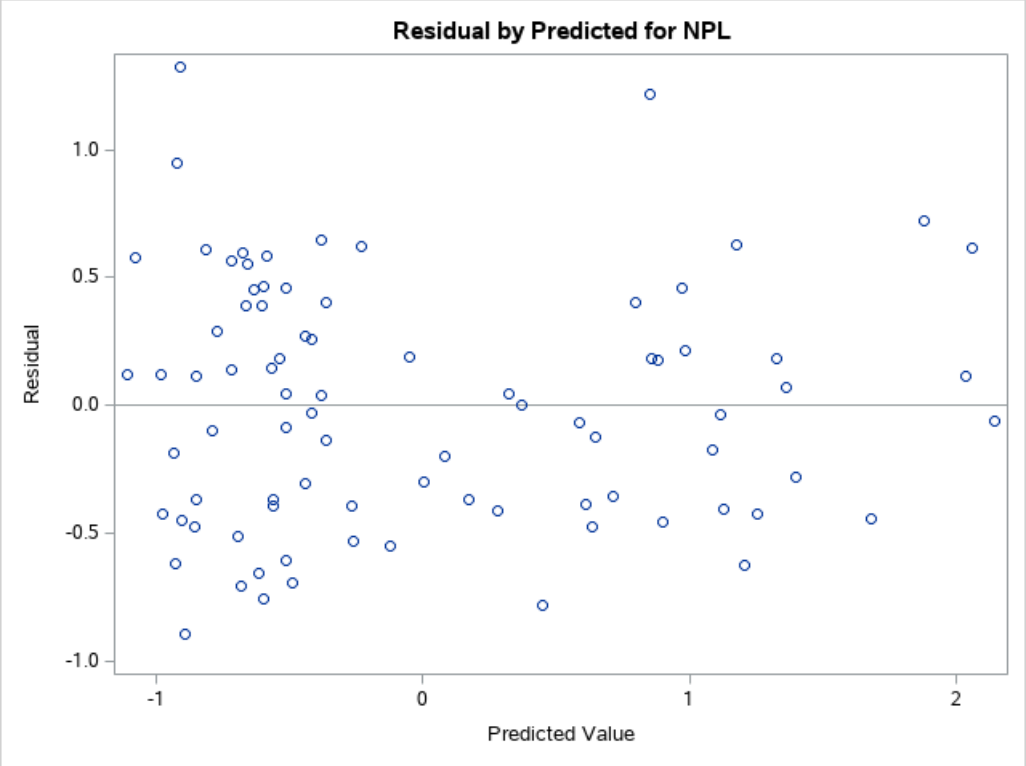


Figure 4.12 Residuals of the predicted values for the 4PC for Nigeria portfolio

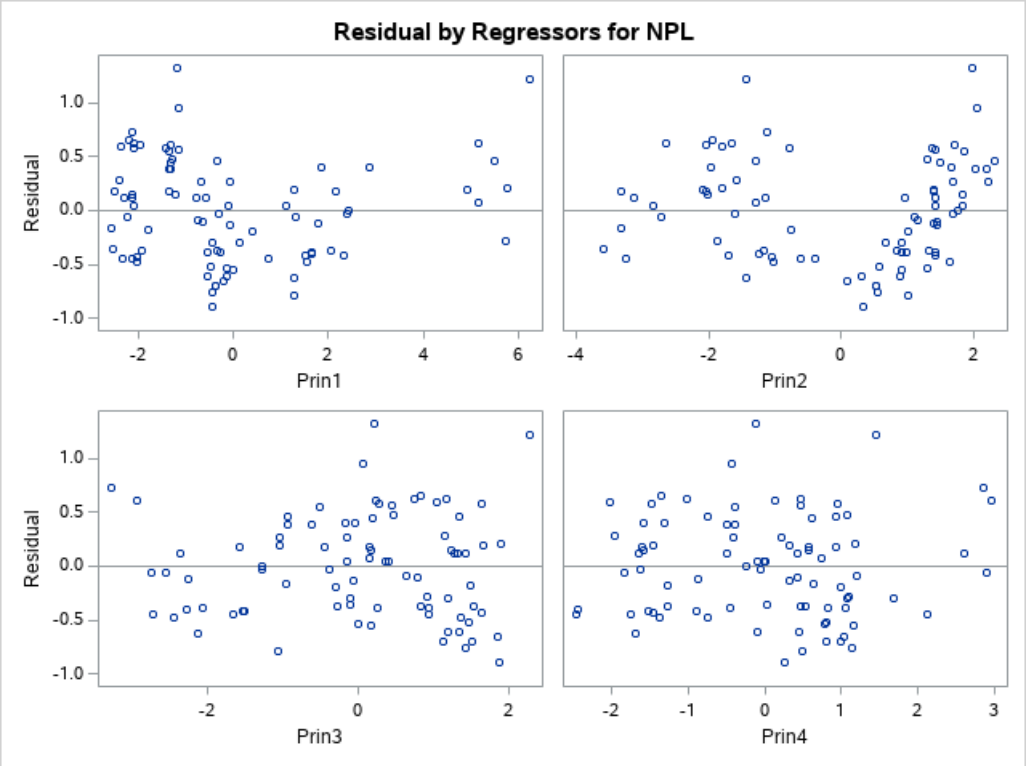


Figure 4.13 Residuals of the predicted values for the individual principal components for Nigeria portfolio

Figure 4.12 shows that the residuals have a random pattern, while in Figure 4.13, PC1, PC2 and PC4 have random patterns with some outliers. On the other hand, PC2's scatterplot has some quadratic relationship with the residuals (Gujarati, 2003). In conclusion, the overall model for the 4 principal components does not have a problem of heteroscedasticity, although there are outliers that are evident.

4.5 Analysis of Kenya portfolio data

4.5.1 Descriptive statistics – Kenya portfolio data

The data exploration considered the quantitative variables on the non-performing loans, to obtain summary statistics presented in Table 4.13. Table 4.13 presents the means, minimum, maximum, and standard deviation of 12 variables each with 108 observations.

Table 4.13 Summary statistics on Kenya non-performing loans data

Summary (Kenya portfolio data)					
Variable	**Number of observations**	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Non-performing loans (NPL ratio) – Dependent variable	108	5.947	3.345	2.640	19.080
Gross domestic product (GDP)	108	5.364	2.217	0.500	11.600
Treasury bill (Tbill)	108	8.417	3.689	1.600	21.650
Interbank call rate (Icr)	108	7.874	5.067	0.980	28.900
Lending rate (Lrate)	108	15.697	2.026	12.900	20.300
Deposit rate (Drate)	108	7.162	2.474	3.600	13.700
M2 money supply (M2)	108	17.540	4.763	10.600	35.000
Exchange rate (FX)	108	82.197	10.045	62.000	105.300
FX reserves (FXR)	108	4.827	1.706	2.500	8.600
M1 money supply (M1)	108	16.211	8.406	-3.000	33.600
Inflation (Infl)	108	8.516	4.994	1.850	19.720
Central Bank Rate (Crate)	108	9.331	3.049	5.750	18.000

The findings in Table 4.13 provide a snapshot of the structure of the Kenya data. The standard deviation that measures the proximity of data points to mean is shown in the table to provide more information on the structure of the data. From Table 4.13, it is evident that FX reserves (FXR) has the lowest standard deviation, while exchange rate (FX) has the highest. Gross domestic rate, lending rate and deposit rate have lower standard deviation, indicative of close proximity to the mean. To show the dispersion and variability of the data, the box plot below presents this data more efficiently.

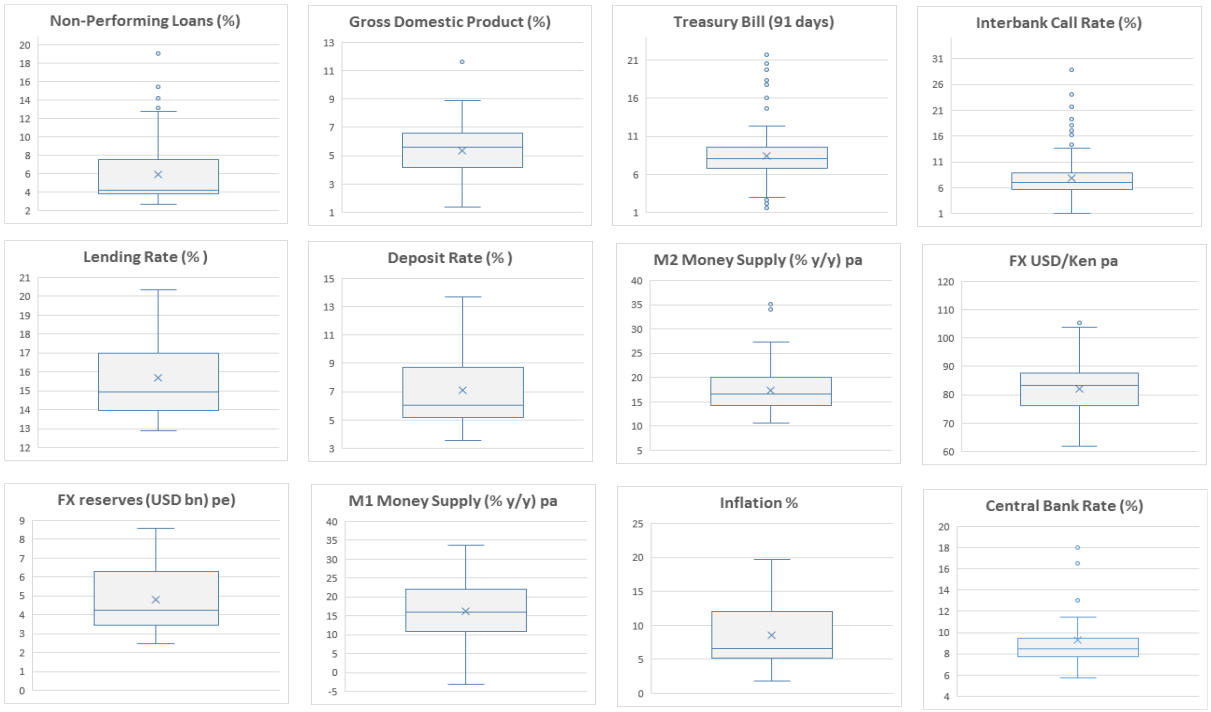


Figure 4.14 Box plot of the variables in the Kenya portfolio data (Dependent variable is Non-Performing Loans)

Figure 4.14 demonstrates the variability of the data. This includes the skewness of the data as well as displaying any outliers that might be present within the data. Table 4.14 provides a summary of the variability of the Kenya portfolio data.

Table 4.14 Summary table for skewness and outliers – Kenya portfolio data

Variable	**Nature of variability**	**Nature of skewness**	**Presence of outliers**
Non-performing loans (NPL ratio) – Dependent variable	Large	Positively skewed	Many
Gross domestic product (GDP)	Small	Negatively skewed	Few
Treasury bill (Tbill)	Small	Symmetric	Many
Interbank call rate (Icr)	Small	Symmetric	Many
Lending rate (Lrate)	Large	Positively skewed	None
Deposit rate (Drate)	Large	Positively skewed	None
M2 money supply (M2)	Small	Negatively skewed	Few
Exchange rate (FX)	Small	Symmetric	Few
FX reserves (FXR)	Large	Positively skewed	None
M1 money supply (M1)	Small	Positively skewed	None
Inflation (Infl)	Small	Symmetric	None
Central bank rate (Crate)	Small	Symmetric	Few

Figure 4.14 and Table 4.14 both summarise the features of the variables included in the dataset. The non-performing loans ratio (dependent variable), Treasury bill and the interbank call rate exhibit many outliers within the data.

4.5.2 Testing for Assumptions – Kenya portfolio data

There are various tests that can be conducted for assumptions as cited by Gujarati (2003), for this research paper, only autocorrelation, stationarity and normality are analysed as presented in Table 4.15 (a-c).

Table 4.15 Auto-correlation, Stationarity, and normality test

(a) Auto-correlation (Durbin-Watson Statistic)	Order 1 = 0.905		Order 2 = 1.372		Order 3 = 1.635		Order 4 = 1.849	
	Pr<DW = <0.0001		Pr<DW = <0.0001		Pr<DW = 0.012		Pr<DW = 0.166	
	Pr>DW = 1.000		Pr>DW = 1.000		Pr>DW = 0.988		Pr>DW = 0.834	
(b) Stationarity test (Augmented Dickey Fuller Unit Root Test)		Lags	Rho	Pr<Rho	Tau	Pr<Tau	F	Pr>F
	Zero Mean	0	-3.732	0.182	-1.360	0.160		
		1	-2.362	0.290	-1.050	0.262		
	Single Mean	0	-15.875	0.027	-2.910	0.048	4.240	0.074
		1	-11.588	0.084	-2.400	0.144	2.880	0.336
	Trend Mean	0	-26.840	0.011	-4.030	0.010	8.200	0.006
		1	-23.239	0.027	-3.650	0.030	6.810	0.037
(c) Normality test (Shapiro -Wilk)			W Statistic	Pr<W			W Statistic	Pr<W
	Non-performing loans (NPL ratio)		0.781	<0.0001	M2 money supply (M2)		0.897	<0.0001
	Gross domestic product (GDP)		0.964	0.005	Exchange rate (FX)		0.973	0.025
	Treasury bill (Tbill)		0.880	<0.0001	FX reserves (FXR)		0.919	<0.0001
	Interbank call rate (Icr)		0.858	<0.0001	M1 money supply (M1)		0.988	0.447
	Lending rate (Lrate)		0.905	<0.0001	Inflation (Infl)		0.860	<0.0001
	Deposit rate (Drate)		0.924	<0.0001	Central Bank Rate (Crate)		0.770	<0.0001

The Durbin-Watson test is used to test for autocorrelation within the data as it assesses the existence of relationships between the error terms. The analysis is derived from the following hypothesis:

H_0: Error terms are uncorrelated

H_1: There exist correlations amongst the error terms

Using the benchmark of 2, whereby if the DW value is closer or equal to 2, then the null hypothesis cannot be rejected. Then, from Table 4.15 (a) in the first order, the Durbin Watson is at 0.905, that is indicative of serial correlations among the error terms. For high orders, the values are closer to 2, hence the null hypothesis cannot be rejected, concluding that the error terms are uncorrelated.

Table 4.15 (b) presents testing results for stationarity. It tests if the data does not fluctuate with time or is not dependent on time. It tests if data is,

H_0: Data is not stationary

H_1: The data is stationary

From Table 4.15(b), using the 5% significance level, for the single mean, the null hypothesis can be not be rejected, when the lag is 0 and 1. For the trend, we reject the H_0 at lag 0, (p-value=0.006) and lag 1 (p-value=0.037) and conclude stationarity, For Rho and Tau, the p-values are mostly more than 0.05, with the exception of single mean at lag 0 and trend mean at lag 0 and 1. For the p-values, where the values are less than 5%, this is indicative of significance.

Table 4.15(c) presents the Shapiro-Wilk test results for the normality assumption. Using the Proc Univariate function in SAS, we obtain the test statistics and p-values for the 12 variables. The test states,

H_0: Data is normally distributed

H_1: Data is not normally distributed

Using the 5% significance level, we reject the null hypothesis for all variables except for M1 money supply. We conclude that most of the variables are not normally distributed, with the exception of M1 money supply where the p-value = 0.447. This means that the probability distributions of the variables are unknown.

In conclusion, the dataset does not have autocorrelation for high orders. This means that estimation of the parameters is likely to be precise as the confidence intervals are likely to be narrower. In addition, the dataset is proven to be stationary, as it does not change over time. In terms of the normality assumption, most of the variables were not normally distributed, hence the probability distributions of these would be unknown, making the hypothesis testing challenging.

4.5.3 Testing for Correlations – Pearson Correlation Coefficient: Kenya portfolio data

Table 4.16 presents Pearson correlation coefficients and associated p-values for testing for correlation among the 11 variables using the Kenya portfolio data.

Table 4.16 Testing for correlations – Kenya portfolio data

Pearson Correlation Coefficients, N = 108											
	GDP	Tbill	Icr	Lrate	Drate	M2	FX	FXR	M1	Infl	Crate
GDP	1.000										
Tbill	-0.330	1.000									
Icr	-0.206	0.846	1.000								
Lrate	-0.152	0.656	0.607	1.000							
Drate	-0.200	0.761	0.690	0.928	1.000						
M2	0.284	-0.302	-0.395	-0.254	-0.239	1.000					
FX	0.122	0.409	0.381	0.496	0.547	0.051	1.000				
FXR	0.101	0.283	0.244	0.510	0.591	0.157	0.753	1.000			
M1	0.469	-0.518	-0.415	-0.600	-0.606	0.415	-0.348	-0.145	1.000		
Infl	-0.511	0.424	0.422	0.048	0.050	-0.310	-0.024	-0.279	-0.308	1.000	0.279
Crate	-0.234	0.762	0.802	0.767	0.800	-0.327	0.174	0.107	-0.551	0.279	1.000

The Table 4.16 depicts the correlations between the variables. In observation, strong positive correlations exist between the deposit rate (Drate) and the lending rate (Lrate) at 92.8%. On the other hand, the deposit rate is also highly correlated with the central bank rate (Crate) and Treasury bill (Tbill) at 80.0% and 76.1% respectively.

The Treasury bill exhibits positive correlations with the central bank rate at 76.2%, while with the interbank rate, the correlation is higher at 84.6%. The interbank call rate is also positively correlated with the central bank rate at 80.2% and with the lending rate at 76.7%. The exchange rate (FX) also exhibits positive correlation with FX reserves (FXR) at 75.3%.

Model Fitting

4.5.4 Ordinary Least Squares – Kenya portfolio data

Using the 11 independent variables and non-performing loans as dependent variable, an ordinary least square model was fitted. Table 4.17 presents the analysis results which include analysis of variance (ANOVA), R-square, estimates of regression model and variance inflation values.

Table 4.17 Ordinary Least Squares output – Kenya portfolio data

Analysis of Variance					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	11	745.125	67.739	14.370	<.0001
Error	96	452.388	4.712		
Corrected Total	107	1197.514			

Root MSE	2.171	R-Square	0.622
Dependent Mean	5.947	Adj R-Sq	0.579
Coeff Var	36.501		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	37.123	4.324	8.590	<.0001	0.000
Gross domestic product (GDP)	1	0.182	0.139	1.310	0.192	2.144
Treasury bill (Tbill)	1	-0.137	0.138	-0.990	0.325	5.900
Interbank call rate (Icr)	1	0.000	0.117	0.000	0.998	7.963
Lending rate (Lrate)	1	-1.148	0.304	-3.780	0.000	8.604
Deposit rate (Drate)	1	1.177	0.419	2.810	0.006	24.442
M2 money supply (M2)	1	0.102	0.059	1.750	0.084	1.773
Exchange rate (FX)	1	-0.160	0.043	-3.690	0.000	4.301
FX reserves (FXR)	1	-0.644	0.310	-2.080	0.041	6.363
M1 money supply (M1)	1	-0.224	0.044	-5.060	<.0001	3.135
Inflation (Infl)	1	0.178	0.066	2.710	0.008	2.435
Central Bank Rate (Crate)	1	-0.518	0.253	-2.050	0.043	13.493

$$\begin{aligned} \boldsymbol{NPL} = 37.123 &+ 0.182\boldsymbol{GDP} - 0.137\boldsymbol{Tbill} + 0.0003\boldsymbol{ICR} - \mathbf{1.148}\boldsymbol{Lrate} \\ &+ 1.177\boldsymbol{Drate} + 0.102\boldsymbol{M2} - 0.160\boldsymbol{FX} - 0.644\boldsymbol{FXR} - 0.224\boldsymbol{M1} \\ &+ 0.178\boldsymbol{Infl} - 0.518\boldsymbol{Crate} \end{aligned} \quad (6)$$

Table 4.17 (analysis of variance) shows that the model is significant at 0.001 when compared to 5% significance level. The $MSE \cong 4.712$ indicates the standard deviation of the residuals or rather how far the predicted values are from the actual values. The $Adjusted\ R^2 =$

0.579 (57.9%), stating the proportion of the variability that is explained by the independent variables. The only variables that are significant are the lending rate, deposit rate, exchange rate, FX reserves, M1 money supply, inflation, and the central bank rate.

The variance inflation factor (VIF) assists in determining the variables that might be causing multicollinearity. Various methods can be used to define the arbitrary value that can be used as a measure. In using the mathematical method of $1/(1-R^2)$, then using the $R^2 = 62.2\%$, then the equation gives 2.646. The variables that have VIFs more than 2.646 are Treasury bill (Tbill), interbank rate (Icr), lending rate (Lrate), deposit rate (Drate), exchange rate (FX), FX reserves (FXR), M1 money supply (M1) and central bank rate (Crate). These variables are indicative of the presence of multicollinearity. The effect of multicollinearity in these variables is that the parameter estimates might not be stable nor reliable, even though some are significant.

The presence of multicollinearity in these variables for ordinary least squares gives a reason to explore other regression methods other than ordinary least squares, such as ridge regression and principal component analysis.

Table 4.18 depicts the reduced fitted model that takes into account the variables that are significant at 5% significance level. This comprises of a reduced number of variables from 11 independent variables down to 6.

Table 4.18 Ordinary least squares output – Kenya portfolio data: fitted model

Analysis of Variance					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	669.519	222.840	43.810	<.0001
Error	104	528.995	5.086		
Corrected Total	107	1197.514			

Root MSE	2.255	R-Square	0.558
Dependent Mean	5.947	Adj R-Sq	0.546
Coeff Var	37.922		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	37.361	2.758	13.550	<.0001	0.000
Lending rate (Lrate)	1	-0.988	0.146	-6.790	<.0001	1.831
Exchange rate (FX)	1	-0.147	0.025	-5.850	<.0001	1.334
M1 money supply (M1)	1	-0.238	0.033	-7.310	<.0001	1.570

A reduced fitted model that only accounts for parameters that are significant becomes:

$$\boldsymbol{NPL} = 37.361 - 0.988\boldsymbol{Lrate} - 0.147\boldsymbol{FX} - 0.238\boldsymbol{M1} \quad (7)$$

Table 4.18 results illustrate that the reduced final model is significant at 5% significance level (p-value<0.0001). The Adjusted $R^2 = 0.546$ (54.6%), while the $MSE \cong 5.086$, indicating the absolute value of the data points to the regression line. The significant variables are the lending rate, exchange rate and M1 money supply. To check if multicollinearity is a factor in the results attained, we use the mathematical criterion of $1/(1 - R^2)$ =2.264. All the variance inflation factors are below the criterion, indicative of no high collinearity problem.

To further understand if the OLS fitted model is robust in predicting the NPLs, a plot of the residuals against the predicted values of the NPLs is analysed as seen in Figure 4.15.

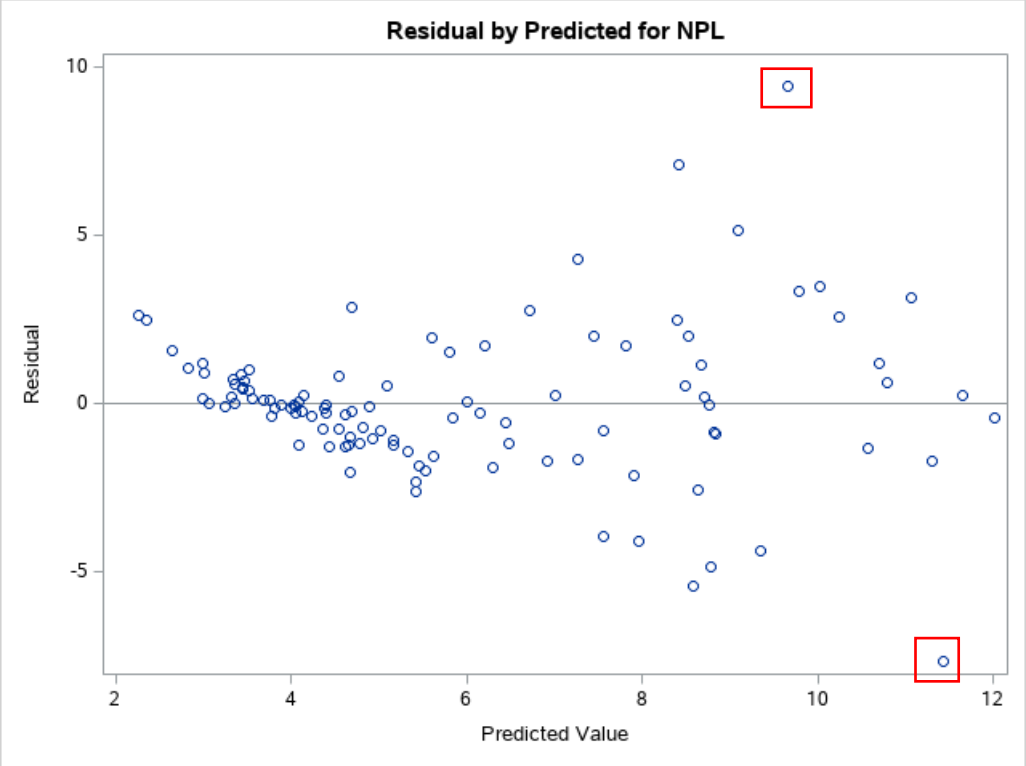


Figure 4.15 Residual plot of the fitted OLS model for the Kenya portfolio

Figure 4.15 shows that the residuals do not resemble a horizontal band that suggest homoscedasticity. The residuals variance is dense when the predicted value are between 2 and 6 and display a random pattern in the latter value of the predicted values of NPLs. This leads to the possibility of heteroscedasticity that is caused by either outliers in the data or that the data used involves a wider range of values. In the Figure 4.15, the outliers are shown in the red boxes.

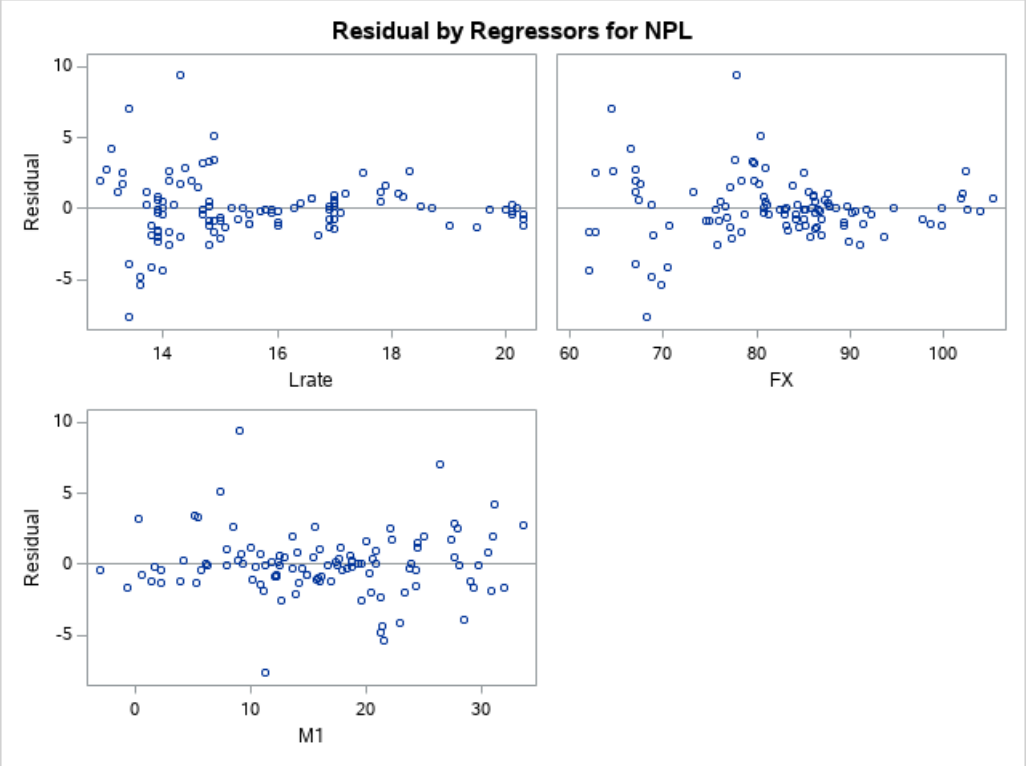


Figure 4.16 Residual plot of the fitted OLS model for the individual independent variables for Kenya portfolio

Figure 4.16 shows that for lending rate, exchange rate and M1 money supply, there is some pattern that shows correlation between the variables and the associated residuals.

In conclusion, the predicted values that are attained from the data that does not have residual values with constant variance are prone to have large variances.

4.5.5 Variable Selection (Post Multicollinearity) – Kenya portfolio data

Variable selection for ordinary least squares is about removal of variables that are causing multicollinearity within the data. This involves variables that have a variance inflation factor (VIF) value of more than 5. The variables omitted are lending rate (VIF=8.604), deposit rate (VIF=24.442), central bank rate (VIF=13.493), interbank call rate (7.963) and FX reserves (VIF=6.363). The resultant effect is depicted below in Table 4.19:

Table 4.19 Ordinary Least Squares output post variable selection – Kenya portfolio data

Analysis of Variance					
Source	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	494.821	98.964	14.370	<.0001
Error	102	702.693	6.889		
Corrected Total	107	1197.514			

Root MSE	2.625	R-Square	0.413
Dependent Mean	5.947	Adj R-Sq	0.384
Coeff Var	44.133		

Parameter Estimates						
Variable	**Degrees of Freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	22.469	2.683	8.370	<.0001	0.000
Gross domestic product (GDP)	1	0.109	0.152	0.720	0.476	1.760
M2 money supply (M2)	1	0.131	0.061	2.140	0.034	1.323
Exchange rate (FX)	1	-0.223	0.029	-7.590	<.0001	1.350
M1 money supply (M1)	1	-0.144	0.042	-3.450	0.0001	1.913
Inflation (Infl)	1	0.145	0.060	2.400	0.018	1.413

Table 4.19 shows that the model is significant at 5% significance level (p-value<0.0001), while the $MSE \cong 6.889$. The $Adjusted\ R^2$ is at 0.384, showing that circa 38.4% of the variability in the dependent variable is explained by the independent variables. From the parameter estimates, the independent variables that were significant at 5% significance level, were M2 money supply, exchange rate, M1 money supply and inflation. The variance inflation factor is lower when using the criterion of $1/(1 - R - square)$ and as a result, the only variables that fall outside the range are gross domestic product and M1 money supply.

Table 4.20 depicts the reduced fitted model of the OLS post variable selection, depicting only significant variables. The number of independent variables is further reduced from 5 to only 4.

Table 4.20 Ordinary Least Squares post variable selection – Kenya portfolio data: Fitted model

Analysis of Variance					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	491.293	122.823	17.910	<.0001
Error	103	706.221	6.857		
Corrected Total	107	1197.514			

Root MSE	2.618	R-Square	0.410
Dependent Mean	5.947	Adj R-Sq	0.387
Coeff Var	44.028		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	22.475	2.677	8.400	<.0001	0.000
M2 money supply (M2)	1	0.129	0.061	2.120	0.037	1.320
Exchange rate (FX)	1	-0.216	0.028	-7.780	<.0001	1.214
M1 money supply (M1)	1	-0.131	0.037	-3.510	0.001	1.529
Inflation (Infl)	1	0.127	0.055	2.320	0.022	1.169

Table 4.20 shows that the model is significant (p-value<0.001) at 5% significance level. The $MSE \cong 6.586$, while the $Adjusted\ R^2 = 0.387$. On the other hand, using the table by Daoud (2017) as reference in the Literature Review (Chapter 2), the variables that have a VIF value of more than 5 are highly collinear. Table 4.20, it can be deduced that none of the independent variables fit this criterion. This suggests that the remaining variables need to be further analysed to check if the results follow the expected logic, otherwise, multicollinearity might still be a problem.

To further check if there are any patterns to be concerned about within the data, a plot of the residuals against the predicted values of the dependent variable is analysed as in Figure 4.17 and Figure 4.18

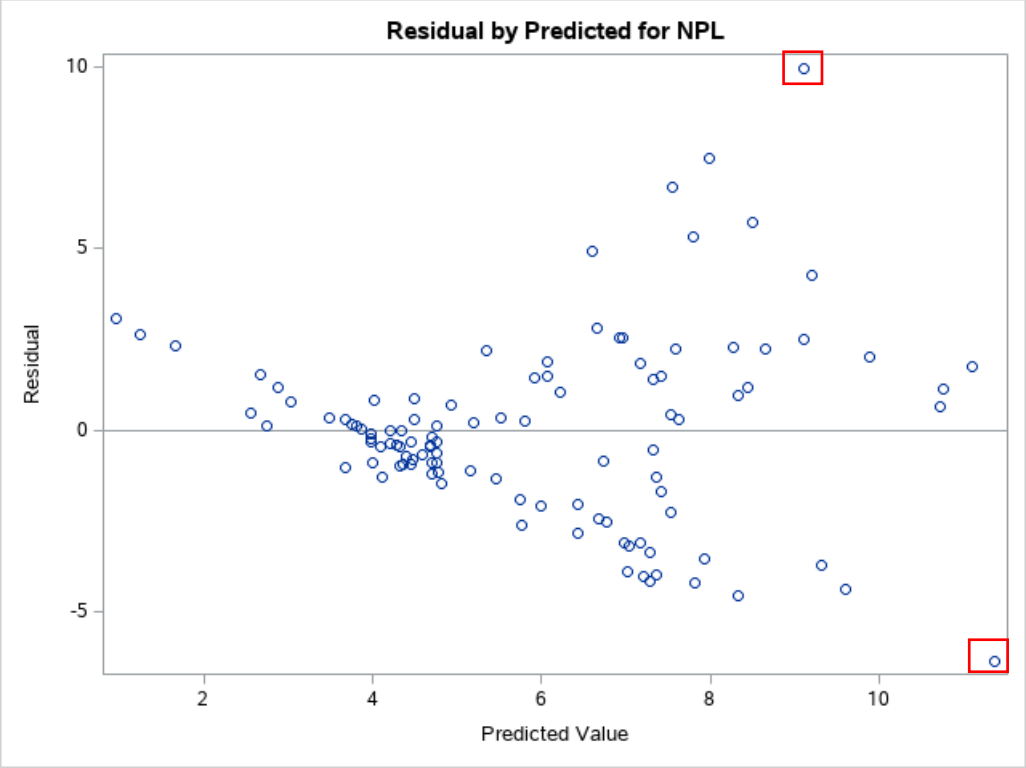


Figure 4.17 Residual plot for the OLS post variable selection for Kenya portfolio

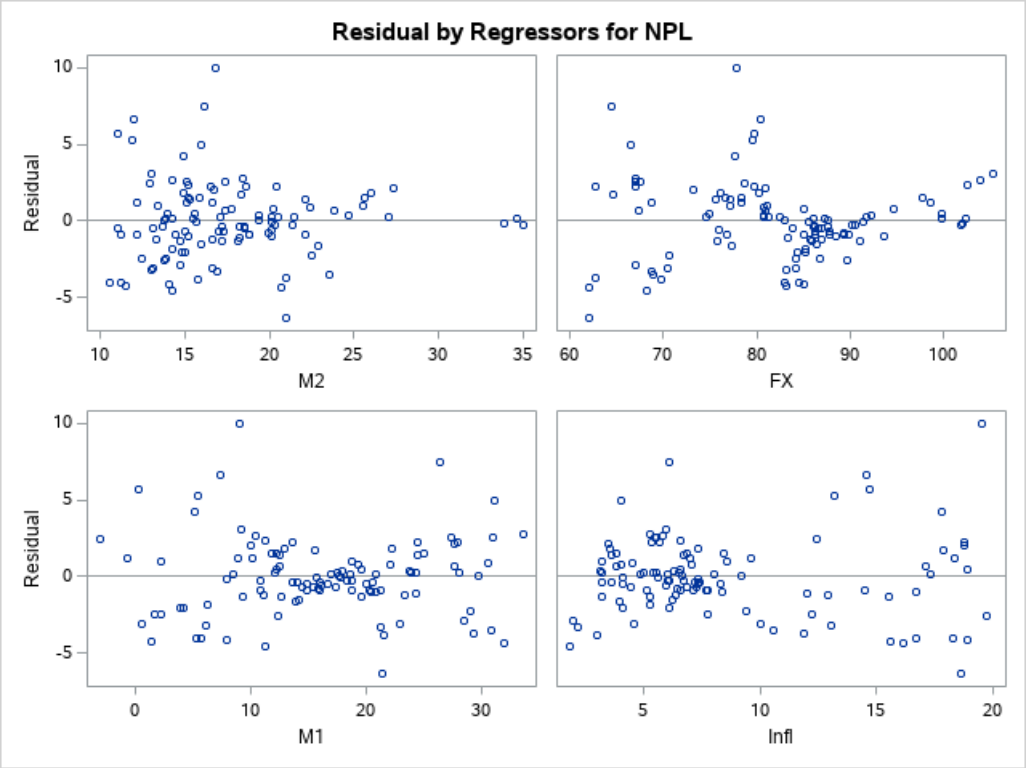


Figure 4.18 Residual plot for the OLS post variable selection for individual regressors for Kenya portfolio

Figure 4.17 shows that there the variance of the residual is not constant. The residual variance is dense when the predicted values are between 4 and 6 and display a random pattern for the higher values of the predicted NPLs. Also notable is the presence of outliers in the data, as shown by the red boxes. Figure 4.18 also shows that there are patterns for the residuals against the regressors. M2 money supply resembles some random pattern but has outliers. In conclusion, the data used has heteroscedasticity and this affects the predictability of the model as the variance are inflated.

4.5.6 Ridge Regression – Kenya portfolio data

Ridge regression is a method that is used to combat the problem of multicollinearity by finding a value of the ridge regression parameter K that can almost disintegrate the linear relationship between the variables by introduction of some biasness to the model. One way of finding the value of the ridge regression parameter is through ridge trace, despite the subjectivity aspect of this method. Due to the high number of explanatory variables, SAS automatically divided the explanatory variables through two ridge trace plots as depicted in .

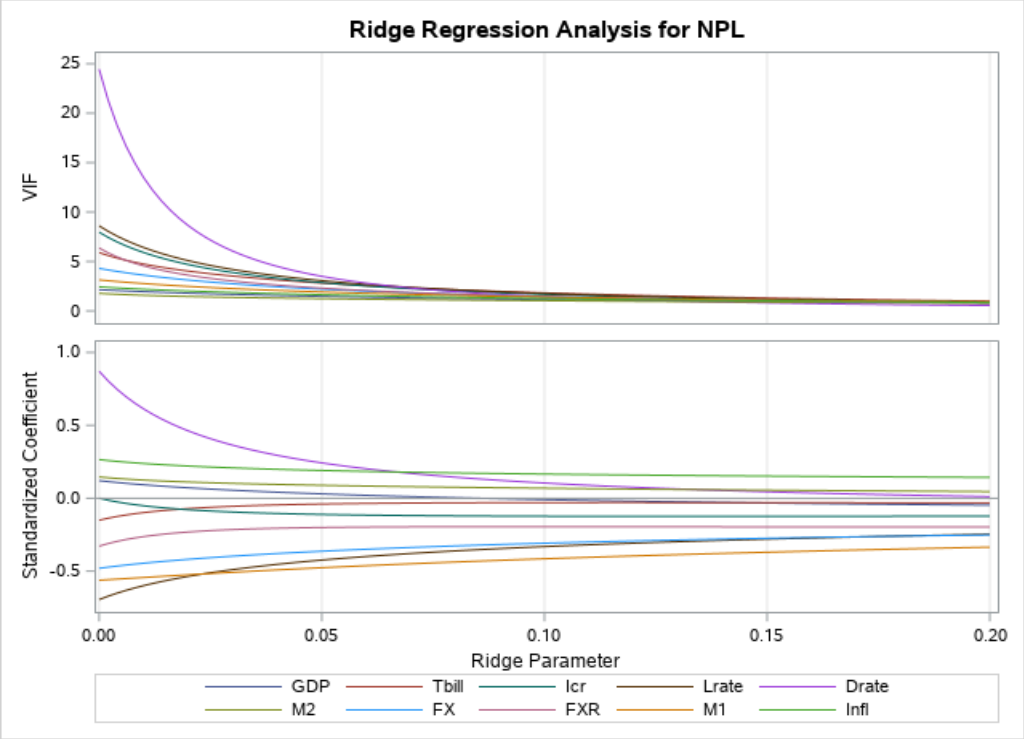


Figure 4.19 Ridge Trace a – for Kenya portfolio

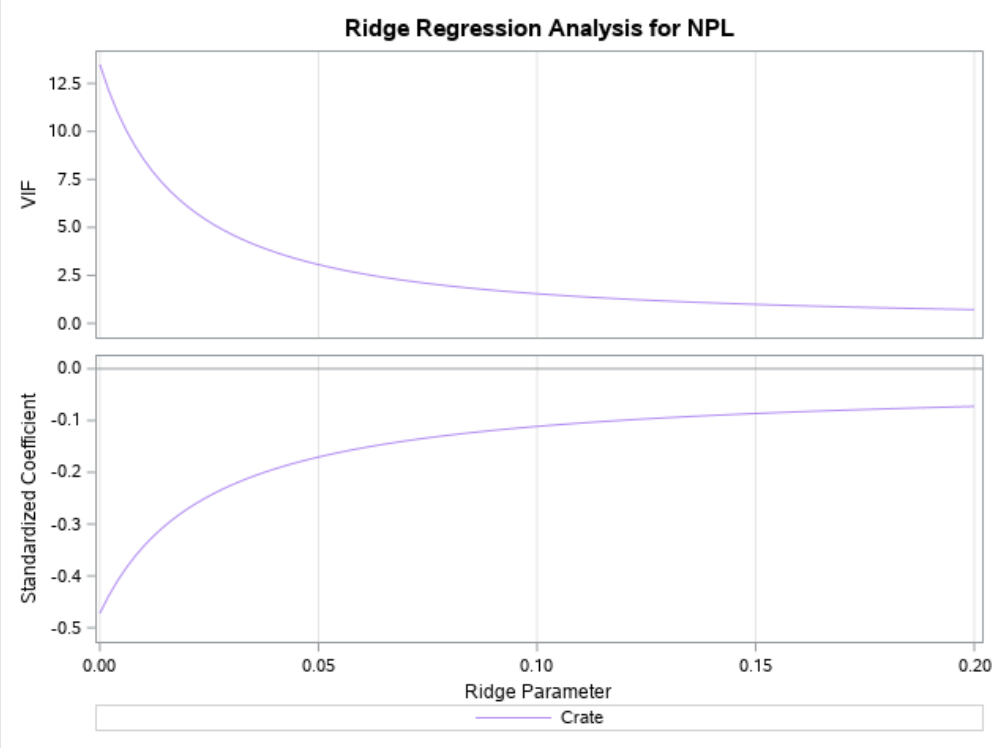


Figure 4.20 Ridge Trace b – for Kenya portfolio

Figure 4.19 and Figure 4.20 showed that the VIFs for most variables converged between 0.0 and 0.2. A closer look at the plots, reveals that this value tends to sway more to being closer to 0.1. The subjective value of K chosen for this research paper would be 0.12.

Table 4.21 displays the output for the ridge regression when the value of the ridge regression parameter $K = 0$.

Table 4.21 Ridge regression output – Kenya portfolio data

Analysis of Variance					
Source	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	11	745.125	67.739	14.370	<.0001
Error	96	452.388	4.712		
Corrected Total	107	1197.514			

Root MSE	2.171	R-Square	0.622
Dependent Mean	5.947	Adj R-Sq	0.579
Coeff Var	36.501		

Parameter Estimates						
Variable	**Degrees of Freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	5.947	0.209	28.470	<.0001	0.000
Gross domestic product (GDP)	1	0.404	0.307	1.310	0.192	2.144
Treasury bill (Tbill)	1	-0.504	0.510	-0.990	0.325	5.900
Interbank call rate (Icr)	1	0.001	0.592	0.000	0.998	7.963
Lending rate (Lrate)	1	-2.325	0.616	-3.780	0.000	8.604
Deposit rate (Drate)	1	2.914	1.038	2.810	0.006	24.442
M2 money supply (M2)	1	0.488	0.279	1.750	0.084	1.773
Exchange rate (FX)	1	-1.608	0.435	-3.690	0.000	4.301
FX reserves (FXR)	1	-1.099	0.529	-2.080	0.041	6.363
M1 money supply	1	-1.881	0.372	-5.060	<.0001	3.135
Inflation (Infl)	1	0.887	0.327	2.710	0.008	2.435
Central bank rate (Crate)	1	-1.580	0.771	-2.050	0.043	13.493

The analysis from Table 4.21 show that at 5% significance level, the model is significant (p-value<0.0001), with the associated $Adjusted\ R^2 = 0.579$, while the $MSE \cong 4.712$. The independent variables that are significant are lending rate, deposit rate, exchange rate, FX reserves, M1 money supply and inflation. Some of the results are in line with the initial expectations. For a unit change in inflation, the NPLs will increase by 0.887. This is what is expected, as inflation would increase the instalment rates, and this might be over the affordability range of the customer. On the other hand, for a unit change in exchange rate, the NPLs will decrease by 1.608. This means that the exchange rate would be favourable to the disposable income of the customer, hence the proclivity to repay debts.

Table 4.22 depicts the reduced model that only shows the significant variables, where the VIFs are closer to 1 and the ridge regression parameter is chosen as $K = 0.12$. The table also gives a comparison for the coefficients when ridge parameter $K = 0$.

Table 4.22 Ridge Regression output – Kenya portfolio data - fitted model

TYPE	**RIDGE**	**RMSE**	Intercept	Lrate	FX	M1	**RSQ**
PARMS	0	0.674	0.000	-0.599	-0.440	-0.597	0.558
SEB	0	0.674	0.065	0.088	0.075	0.082	.
RIDGEVIF	0.12	.	.	1.109	0.946	1.024	.
RIDGE	0.12	0.687	0.000	-0.475	-0.455	-0.455	.
RIDGESEB	0.12	0.687	0.066	0.070	0.067	0.067	.

From Table 4.22 we see that at 5% significance level, the variables that are significant are lending rate, exchange rate and M1 money supply. The $R^2 \cong 0.558$ shows that the variability is explained by the variables. Following on the concept of correlation transformation, the standardized ridge regression coefficients have to be transformed back in order to attain the original variables.

For the reduced fitted model, the ridge regression coefficients as depicted in Table 4.22 becomes:

$$\boldsymbol{NPL}^{*} = -0.475\boldsymbol{Lrate}^{*} - 0.455\boldsymbol{FX}^{*} - 0.455\boldsymbol{M2}^{*} \quad (8)$$

Transforming the ridge regression coefficients to the original variables equation (8) becomes

$$\boldsymbol{NPL} = 33.650 - 0.784\boldsymbol{Lrate} - 0.152\boldsymbol{FX} - 0.181\boldsymbol{M1} \quad (9)$$

The model defined in equation (9) is the one used in the prediction of NPLs for ridge regression.

To further understand the data, the residuals for the ridge regression are shown below in Figure 4.21 and Figure 4.22

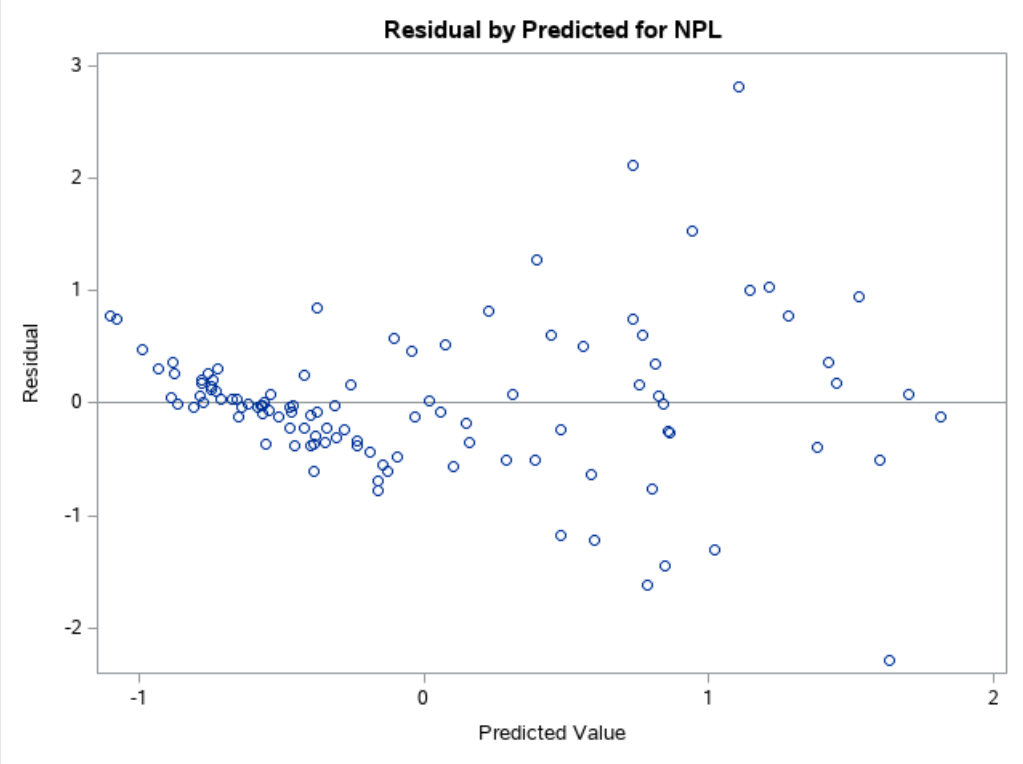


Figure 4.21 Residual plot for the fitted ridge regression predicted values for Kenya portfolio

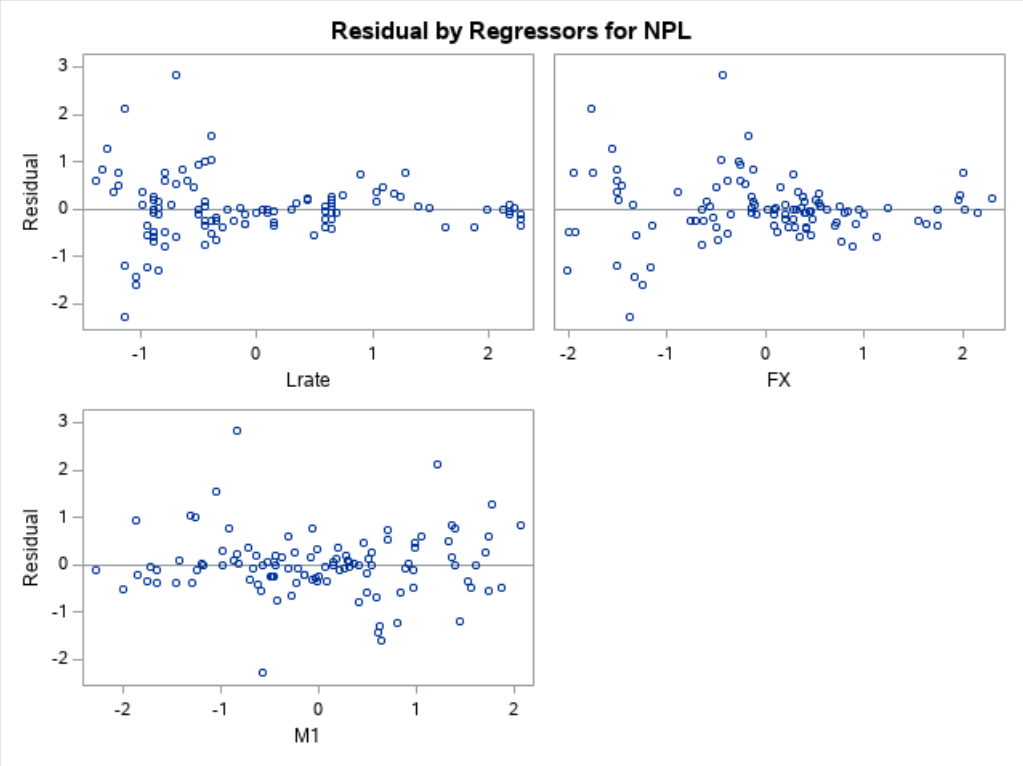


Figure 4.22 Residual plot for the fitted ridge regression for the individual regressors for Kenya portfolio

Figure 4.21 and Figure 4.22 resemble the residual plots of the fitted OLS plots. Based on these illustrations, it is evident that there is heteroscedasticity involved, which affects the robustness of the ability of the ridge regression to predict the NPLs.

4.5.7 Principal Component Analysis – Kenya portfolio data

The principal component analysis on the Kenya portfolio data led to the results presented in Table 4.23. Table 4.23 contains the eigenvalues for the 11 variables and proportion of variation accounted for by each of them. Principal component analysis was operated on a correlation matrix.

Table 4.23 Eigenvalues of the correlation matrix – Kenya portfolio data

Eigenvalues of the Covariance Matrix				
PC	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
1	5.239	2.982	0.476	0.476
2	2.258	1.315	0.205	0.682
3	0.942	0.104	0.086	0.767
4	0.838	0.147	0.076	0.843
5	0.691	0.248	0.063	0.906
6	0.442	0.200	0.040	0.946
7	0.243	0.099	0.022	0.968
8	0.144	0.025	0.013	0.982
9	0.119	0.060	0.011	0.992
10	0.058	0.031	0.005	0.998
11	0.027		0.002	1.000

Table 4.23 shows that about 68.2% of the total variation is explained by 2 components. This means that from the data, only 2 variables can be used instead of the 11 variables. The principal component analysis is known as a data reduction technique. This is further supported by the scree plot depiction in Figure 4.23. The line starts to flatten from around 2, hence we can subjectively have only 2 variables in the model that can explain the maximum variability.

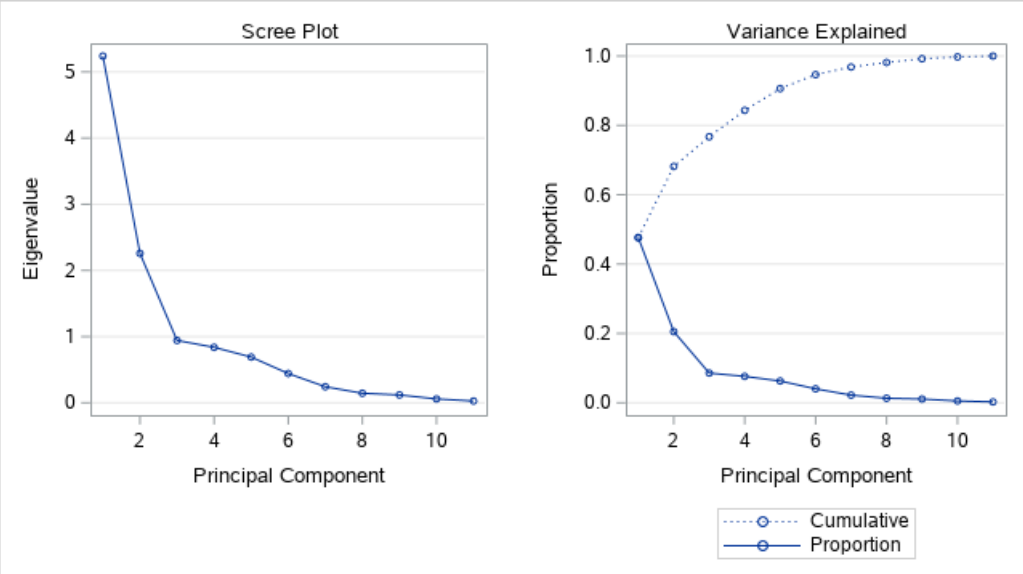


Figure 4.23 Scree plot for Kenya portfolio

Table 4.24 presents the eigenvectors for the two selected PCs accounting for 68.2% of the total variation. The dominating variables in each of the PC is bolded based on coefficients with a magnitude greater 0.3.

Table 4.24 Eigenvectors of the correlation matrix on the Kenya portfolio data

Variables	**Prin1**	**Prin2**
Gross domestic product (GDP)	-0.155	**0.395**
Treasury bill (Tbill)	**0.385**	-0.071
Interbank call rate (Icr)	**0.369**	-0.082
Lending rate (Lrate)	**0.381**	0.163
Deposit rate (Drate)	**0.404**	0.176
M2 money supply (M2)	-0.178	**0.330**
Exchange rate (FX)	0.235	**0.408**
FX reserves (FXR)	0.191	**0.512**
M1 money supply (M1)	**-0.314**	0.135
Inflation (Infl)	0.150	**-0.457**
Central bank rate (Crate)	**0.372**	-0.103

The purpose of the eigenvectors is to depict the strength of the relationship between the principal component and the original independent variable. From Table 4.24, it can be observed that the first principal component, PC1, is an average measure and has large positive associations with Treasury bill (39%), interbank call rate (37%), lending rate (38%), deposit rate (40%), central bank rate (37%), while attaining negative association with M1 money supply (31%). PC2 measures the contract of variables gross domestic product (40%), M2 money supply (33%), exchange rate (41%), FX reserves (51%) against inflation (46%).

From the grouped variables explained, the positive associations increase the NPLs while negative associations tend to decrease the NPLs. Since PC1 explains the most variability, M1 money supply would decrease the NPLs, while Treasury bill, interbank call rate, lending rate, deposit rate and central bank rate would increase the NPLs. This would also apply to PC2 that

implies that inflation would decrease the NPLs, although this goes against the logic as inflation typically inflates values.

Table 4.25 below shows the regression output of the two principal components that have eigenvalues that are more than 1 and cumulatively explains 68.2% of the total variation.

Table 4.25 Principal component regression output – Kenya portfolio data

Analysis of Variance					
Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	45.071	22.535	38.210	<.0001
Error	105	61.929	0.590		
Corrected Total	107	107.000			

Root MSE	0.768	R-Square	0.421
Dependent Mean	-2,37E-16	Adj R-Sq	0.410
Coeff Var	-3,23E+17		

Parameter Estimates						
Variable	**Degrees of freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	0.000	0.074	0.000	1.000	0.000
Prin1	1	-0.150	0.032	-4.640	<.0001	1.000
Prin2	1	-0.366	0.049	-7.410	<.0001	1.000

$$\boldsymbol{NPL} = -0.150\boldsymbol{PC1} - 0.366\boldsymbol{PC2} \quad (10)$$

Table 4.25 shows the overall model is significant at a p-value that is less than 0.001. The $MSE \cong 0.590$, indicating the absolute fit of the model. The variability that is explained by the independent variables is at 0.410 (41%) when looking at the $Adjusted\ R^2$, while the R^2 that does not account for degrees of freedom lies at 0.421. The regression analysis on the principal components indicates that there is no multicollinearity as the VIFs are all one.

To gain better understanding of the predictability of the model using principal components, the residual plots are shown in Figure 4.24 and Figure 4.25.

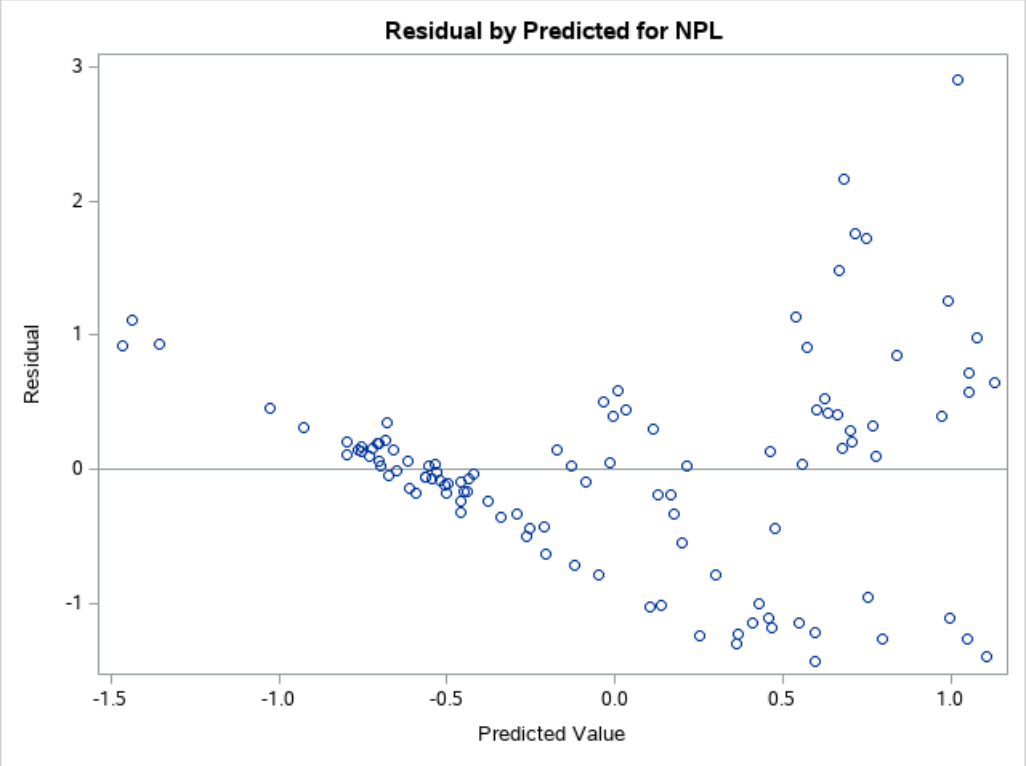


Figure 4.24 Residual plots of the principal components regression residuals against the predicted values for Kenya portfolio

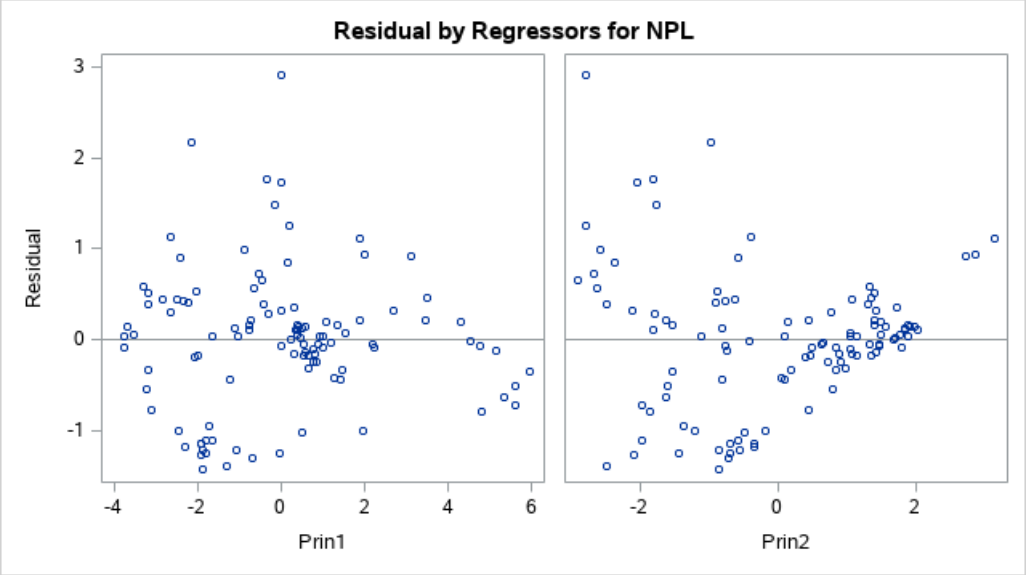


Figure 4.25 Residual plots of the principal components regression residuals against PCs for Kenya portfolio

Figure 4.24 show that there is a pattern, and that the variance of the residuals is not constant. For the individual PC, in Figure 4.25, PC1 shows some random pattern, while PC2 shows a relationship with the residuals. The conclusion is that there is a possibility of heteroscedasticity.

CHAPTER 5

DISCUSSIONS

5 Discussion

5.1 Case Study – Nigeria portfolio data

5.1.1 Descriptive Analysis Discussion – Nigeria portfolio data

The central limit theorem states that for samples sizes larger than 30, the sample mean would approach normal distributions, hence in this scenario, though the normality assumption is violated, the estimates can still be used (Islam, 2018). For the Nigeria portfolio dataset, some variables failed the normality test. The data has time series structure hence the need to test for stationarity. The results show that the dataset is stationary, meaning that the variables are not dependent on time and can be used for forecasting purposes. This is important as one of the objectives is to investigate the predictive power of the ordinary least squares, ridge regression and principal component analysis.

The presence of multicollinearity in the data was detected using the Pearson Correlation Coefficient metric. High correlation between gross domestic product (GDP) and maximum lending rate (MLR) (r=0.91) exist. This implies that when the lending rates are high, the consumers are likely to not afford the loans. Similarly, the exchange rate (FX) appreciates and strengthens as the GDP increases and adds to economic growth. The maximum policy rate (MPR) informs the MLR, thus when one increases the other is expected to increase too. Another robust method that could have been explored is Spearman Correlation method. According to Hauke and Kossowski (2011), Spearman is a better method than Pearson.

In the Nigeria context, sovereign revenue is mainly driven by crude oil (Crude). This notion is partially supported by the analysis from Pearson Correlation Coefficient (r=0.75) between crude oil and GDP. This is also supported by the high negative correlation between crude oil and non-performing loans (NPLs). For countries whose sovereign revenue is highly dependent on crude oil, higher crude oil prices would stimulate growth (Khamdelwal, Miyajima and Santos, 2016), thus fewer delinquent loans are expected. GDP and foreign exchange reserves (FXR) are positively correlated (r=0.65). Economic stimulus results in GDP growth, which leads to an increase in FX reserves. The reserves are used in the future as a catalyst for monetary policy, to stabilise the currency, as well as to improve the country's credit worthiness (Akamobi and Ugwanna, 2017).

5.1.2 Comparative Assessment of Models – Nigeria portfolio data

The Ordinary Least Squares (OLS) model is significant, with $Adjusted\ R^2 = 0.89$ and $MSE \cong 1.01$, but not all the parameters are significant. The reduced final model has $MSE \cong 1.411$, while the $Adjusted\ R^2 = 0.853$. At 5% significance level, the variables that were significant were crude oil (Crude), lending rate (Lrate), M2 money supply (M2), exchange rate (FX), FX reserves (FXR) and monetary policy rate (MPR). All the variables exhibited lower variance inflation factors when compared to the criterion based on the mathematical method of $1/(1 - R^2)$.

Significant crude oil rate (p-value=0.0001) implies that for every unit change in crude oil, the NPL decreases by 0.067, holding other variables constant. Crude oil is the determinant of sovereign revenue in Nigeria; hence it is expected that with more crude oil, there would be more economic activity that would ideally translate to more revenues being generated to alleviate the level of NPLs. In addition, for a unit change in the lending rate, the NPLs increase

by 1.212. This is expected as the central banks would lend money to the commercial banks; hence the banks will have a premium on top of the MPR in order to be able to make profit, making the lending rate higher. When the lending rate is higher, customers can find it difficult to repay their debt obligations (Khamdelwal, Miyajima and Santos, 2016).

Similarly, for a unit change in FX reserves, the NPLs increase by 0.245, holding other variables constant. For exchange rate, for every unit change, the NPLs increase by 0.047, holding others constant. Typically, the central bank would use the exchange rate to control the economic activity within a country. Depending on the level of interest rates, if they are higher, then there can be currency appreciation, while lower interest rates can cause depreciation of the currency.

In addition, detection of multicollinearity that necessitated the variable selection does not result in efficient nor adequate outcome. Instead, the variable selection criteria that excluded the independent variables with high correlations led to the $Adjusted\ R^2 = 0.77$, a decrease which is dependent on the number of variables included in the analysis. The remaining variables that were significant were crude oil, lending rate, FX reserves and inflation. For every unit change in crude oil, the NPL decreases by 0.083, keeping all else constant. This is the expected result. In contrast, the FX reserves show that for every unit change, the NPL increases by 0.199. Although FX reserves can be used to improve the country's creditworthiness (Akamobi and Ugwanna, 2017) , this might not filter to individuals. The MSE increased to 2.216, hence, the variable selection method is not ideal when using the comparative metric MSE.

From Figure 5.1 presents the forecasted NPLs for OLS for the period January 2017 – December 2018, using the final fitted model (only significant variables used) as depicted below:

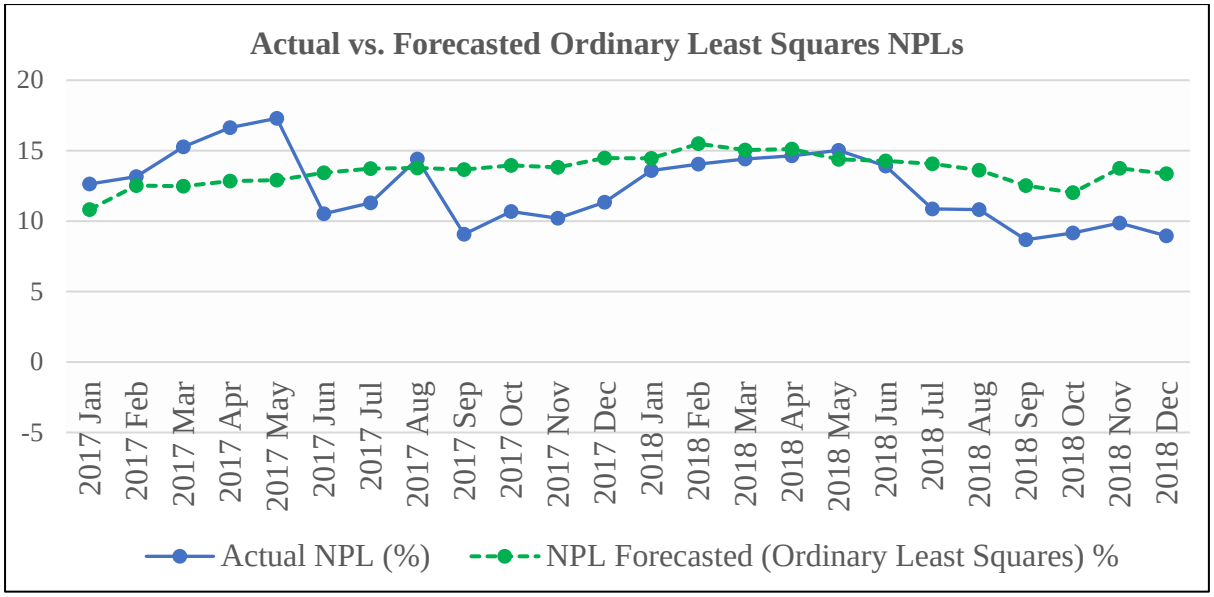


Figure 5.1 Forecasted NPLs using Ordinary Least Squares (Fitted reduced model) on Nigeria portfolio data

From Figure 5.1 the forecasted values of NPLs are trending slightly above the actual NPLs. For the actual non-performing loans ratios, there are notable outliers from January 2017 to May 2017 as this is due to the roll forward of accounts into NPL as the sovereign revenue shrunk due to the significant decrease in oil prices during that period in review. The less NPLs during the latter part of the months is explained by the written off portfolio as the Central Bank of Nigeria urged the banks to significantly lower their NPL rates to desirable rates. For Bank X, this was a period of accelerated write-off of delinquent accounts in the portfolio to normalise the NPLs to acceptable levels rates. The forecasted trend does not show the volatility of NPLs during January and May 2017 but shows a smooth trend.

The ridge regression parameter, on the other hand, is chosen through the ridge trace instead of mathematical programming. At $K = 0.11$, the variance inflation factors are closer to 1, while the $Adjusted\ R^2 = 0.853$.

The reduced fitted model above is used to graphically represent the trend of the actual NPLs against the ridge regression predictions as shown in Figure 5.2.

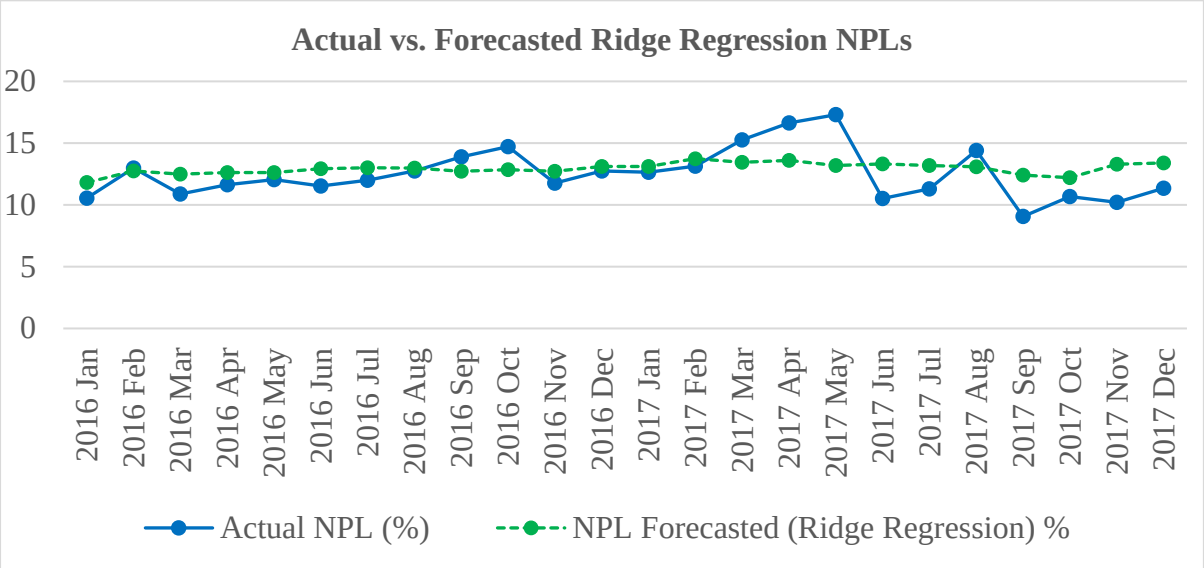


Figure 5.2 Forecasted NPLs using Ridge regression (Fitted reduced model) on Nigeria portfolio data

When comparing Figure 5.1 and Figure 5.2, the ridge regression forecasted the NPLs better than the fitted ordinary least squares, as observed in Figure 5.2. The figures present the fitted models using the data from January 2017 – December 2018 NPLs.

To further understand the complexity of forecasting the NPLs in the Nigeria context, Figure 5.3 shows the trend of the GDP rates.

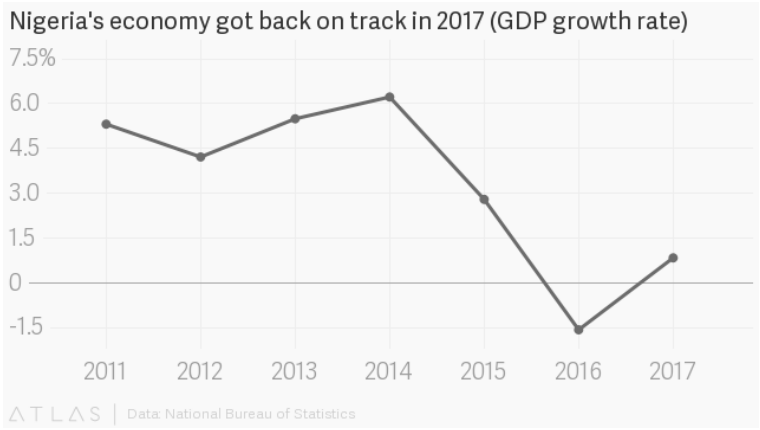


Figure 5.3 GDP Rates in Nigeria

From Figure 5.3, the 2016 value is a possible outlier because GDP contracted up to -1.6%, following the decline in oil exportation caused by political instability with Niger Delta militants. The graph supports the uptick of the economy in 2017 as economic policies and oil prices amid other economic drivers start to improve. According to Zahari *et al.* (2014), such shocks in the economy render the results of ridge regression less reliable. Polat and Turkan (2016) ascertain that the predictive power of a model is negatively affected by the presence of outliers. The presence of outliers prevents the ridge regression model from capturing the impact of the macroeconomic variables on the NPLs.

An improvement on the shortcoming noted in the OLS and RR models is the PC Regression model. The PC Regression based on the correlation matrix, produces at least 4 eigenvalues, accounting for up to 85% of the total variation. The interest in this study is not on the reduction of dimensionality, but to address the multicollinearity problem. In addressing the problem, we fit a PC regression model which considers the PCs as independent variables.

The interpretation of the significant PCs considers the following relationship among the dominant original variables per PC:

PC1: Measures the average of (FX, MLR, MPR) against the average of (GDP, Crude, FXR).

PC2: Measures the average of (Tbill, Drate, MPR) against the average of (M1, Infl).

PC3: Measures the average of (Crude, Tbill, Icr, M1) against (Lrate).

PC4 : Measures the average of (Lrate, Drate, M2, FXR).

For comparison, the metric that was used was MSE. The MSE for principal component analysis ($MSE \cong 0.238$), was better than that of OLS ($MSE \cong 1.411$), but higher than that of ridge regression $MSE \cong 0.147$. Principal component analysis manages to combat the problem of multicollinearity within the data. As literature remarks through Nduka and Ijomah (2012), principal component brings about challenges as it explains the independent variables, rather than the dependent variable, hence even through the selected components, relevance of the chosen components for inclusion in the model is not guaranteed. For this dataset, Ridge regression produced biased estimates and managed to predict the NPL with some degree of precision. The precision of ridge regression can be enhanced when other challenges such as outliers are corrected (Shariff and Ferdaos, 2017, Polat and Turkan, 2016).

5.2 Case Study – Kenya portfolio data

5.2.1 Descriptive Analysis Discussion – Kenya portfolio data

Within the Kenya portfolio data, the assumption of autocorrelation is tested and proved to be adhered to, whereby from order 2 and upwards, the null hypothesis could not be rejected, which adheres to the notion that there exists no relationship between the error terms. This is crucial as if there was any form of correlation between the error terms, then the results from OLS would be deemed not precise nor accurate. For OLS results to have higher precision levels, the error terms should follow the random pattern. In terms of the first order, the Durbin-Watson is at 0.905, which is further from the benchmark of 2, hence the error terms at the p-value of 0.05,

the null hypothesis of uncorrelated error terms is rejected. This would infer that at order 1, the error terms might be correlated.

The results show that the Kenya portfolio data exhibits stationarity characteristics for the trend, while zero and single mean, the data does not adhere to the stationarity assumption. This can be expected as time series data fluctuates with time.

The normality assumption, when tested, showed that the data does not adhere to this assumption. The central limit theorem states that for samples sizes larger than 30, the sample mean would approach normal distributions, hence in this scenario, though the normality assumption is violated, the estimates can still be used (Islam, 2018). At 5% significance level, the only variable that exhibited normality characteristics was M1 money supply.

To be certain that no relationship exists between the independent variables, a multicollinearity test should be conducted. For the Kenya portfolio data, there are some correlations that are worth mentioning: between Treasury bill and interbank call rate at r=0.85. This implies that interbank call rates inform the demand trends of Treasury bills.

The interbank call rate shows high correlation with the central bank rate at r=0.80 as interbank rates are determined by the central bank rates. On the other hand, the lending rate and deposit rate were also highly correlated at r=0.93. This stems from the notion that deposits are used to offer loans to customers. The fee for the lending rate is pre-determined by the deposit rate as banks have to make some profit from the loans that they provide, while simultaneously being able to pay off the fees of the deposit. Additionally, the lending rate is pre-determined by the central bank rate, hence the high correlation of r=0.77 between these two variables. In Kenya's

context, the lending rates was capped at 4% above the central bank rate in an effort to curb the inflow of non-performing loans (Muriuki, Mathuva and Egondi, 2017). The same phenomenon also applies to deposit rates and the central bank rate that displayed a high correlation of r=0.80.

5.2.2 Comparative Assessment of Models – Kenya portfolio data

The ordinary least squares (OLS) reflected an Adjusted $R^2 = 0.579$, while the MSE $\cong 4.71$. For the fitted model, the model is significant at 5% significance level (p-value<0.001), while the $Adjusted\ R^2 = 0.546$ and the $MSE \cong 5.086$. At 5% significance level, the variables that were significant were lending rate (Lrate), exchange rate (FX) and M1 money supply (M1). All the variables exhibited lower variance inflation factors when compared to the criterion based on the mathematical method $1/(1 - R^2)$. The plot of the residuals against the predicted values showed some pattern where the variance of the residuals was not constant. This can lessen the predictability of the model to be insufficient.

A significant p-value (p=<0.0001) implies that for a unit change in M1 money supply, the NPLs decrease by 0.238, while holding others constant. This is expected as more money circulating can improve the proclivity of the debtors to honour their repayment obligations (Tyona *et al*., 2017). For exchange rate, a unit change would cause the NPLs to decrease by 0.147, holding others constant. A favourable exchange rate would assist in decreasing the level of NPLs as it is supposed to boost economic activity (Ahmed *et al*, 2021)

On the other hand, an unexpected result is of the significant lending rates (p<0.0001). A unit change in lending rates results in the NPLs decreasing by 0.988, holding others constant. This is not in line with expectations as when lending rates increase, they increase the instalment amount that customers are supposed to pay, hence the inability of the customers to afford the

new levels of instalments and they subsequently default. On the other hand, due to the cap on lending rates, if they were higher, the cap would force the lending rates to be reduced, hence the proclivity of the customers to pay, and ultimately reduce the NPLs. For the purpose of this paper, the reason for this ambiguity outcome is multicollinearity.

In addition, to detect multicollinearity, the variable selection does not result in an adequate nor efficient outcome. The variable selection criteria excluded the variables with high correlations that led to $Adjusted\ R^2 = 0.387$. The MSE increased to 6.587, up from 5.086 (from equation 12), hence the variable selection method is not ideal when using the comparative metric MSE.

The OLS model fails to produce a robust forecasted model for predicting NPLs using macroeconomic variables, as shown in Figure 5.4. The fitted model only takes into account the significant variables from equation 12.

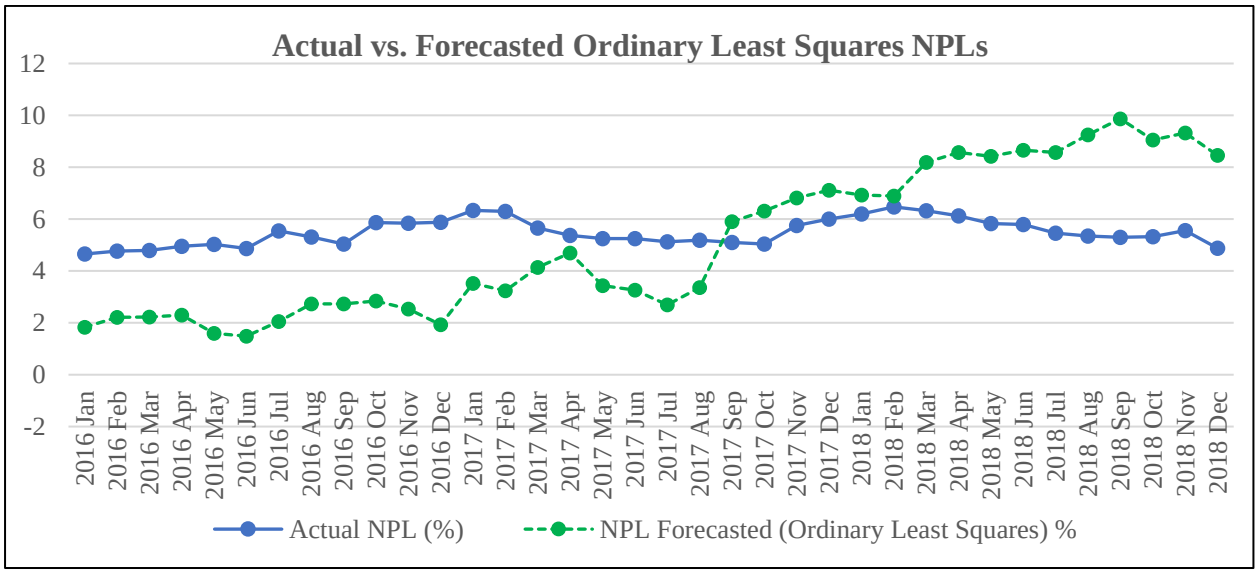


Figure 5.4 Forecasted NPLs using Ordinary Least Squares (Fitted model) on Kenya portfolio data

The Figure 5.4 shows that the fitted model does not accurately predict NPLs for the Nigeria data. The forecasted NPL ratios are trending below the actual NPL line but are higher from

October 2018 to December 2018. This can be explained by the heteroscedasticity that exists, as the residuals have a variance that is not constant. The other reason is the lower variability ($Adjusted\ R^2 = 0.546$).

The actual NPLs do not exhibit notable spikes, except between November and March. For Bank X, these are seasonal spikes which are expected as it encompasses the festive season, where most employers pay their employees early and customers tend to divert the funds for festivities. Similarly, between January and March, it is expected that funds will be diverted to other activities such as school fees, but the trend then stabilises post March. The other factor that is contributing to constant NPLs is that the Central Bank of Kenya had implemented a cap on lending rates that hindered credit growth, a factor that would have forced commercial banks to either perfect the art of acquiring good quality customers or be forced to get any customers without thorough assessment on their credit worthiness.

The ridge regression parameter on the one hand is chosen through the ridge trace instead of mathematical programming. At $K = 0.12$, the $MSE \cong 0.471$, while the $Adjusted\ R^2 = 0.579$. The fitted regression model at $K = 0.12$, the $MSE \cong 0.472$, while the $R^2 = 0.558$. Figure 5.5 shows the actual and the predicted values of the Kenya portfolio using the fitted ridge regression.

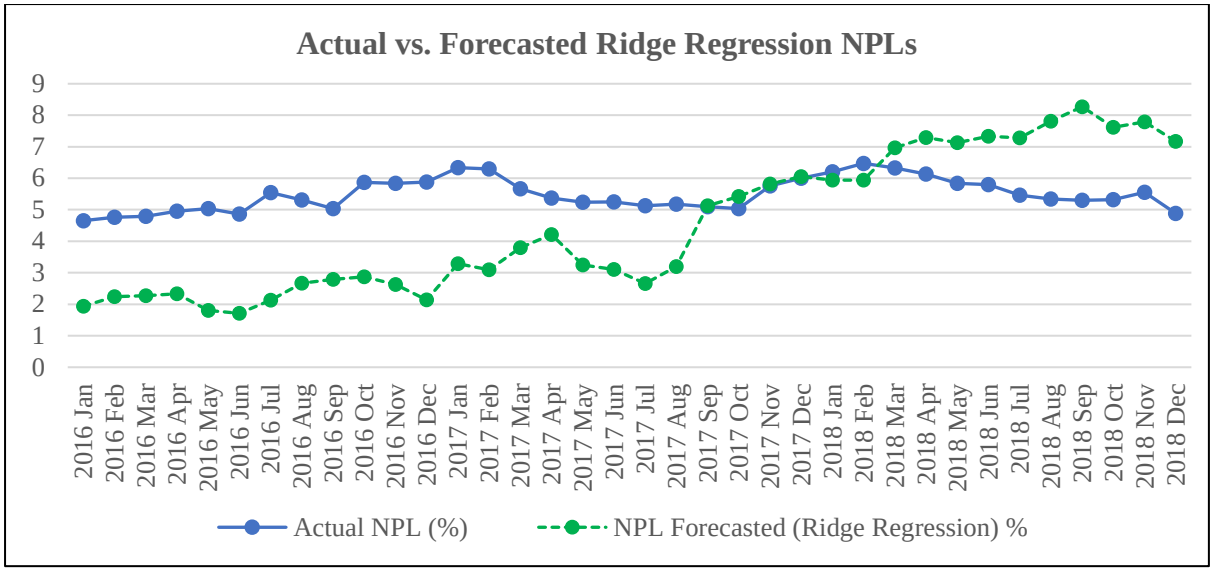


Figure 5.5 Forecasted NPLs using Ridge Regression (Fitted Model) on Kenya portfolio data

Figure 5.5 shows that the ridge regression does not predict the actual NPLs. The reason for the trend of the predicted values can be attributed to model specification and the model specification.

An improvement on the shortcomings noted in the ridge regression and ordinary least squares, is the principal component analysis. This regression model is based on the correlation matrix that produced at least 2 eigenvalues, accounting for 68.2% of the total variation. The purpose of this study is not on data reduction, but to address the problem of multicollinearity. The significant PCs are used in fitting a reduced model that provides forecast for the NPLs.

The interpretation of the significant PCs considers the following relationship among the dominant original variables per PC:

PC1: Measures the average of (Tbill, Icr, Lrate, Drate, Crate) against (M1).

PC2: Measures the average of (GDP, M2, FX, FXR) against (Infl).

The MSE for principal component analysis ($MSE \cong 0.590$) was better than that of OLS ($MSE \cong 5.086$), but higher than that of ridge regression ($MSE \cong 0.472$). Principal component brings about challenges as it explains the independent variables, rather than the dependent variable, hence even through the selected components, relevance of the chosen components for inclusion in the model is not guaranteed (Nduka and Ijomah (2012).

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6 Conclusions

The objective of this research study was to assess an optimal model among the three outlined models that can eliminate the problem of multicollinearity within the Africa portfolio data for Bank X, while simultaneously finding the best model that can forecast non-performing loans rates using macroeconomic variables.

The ordinary least squares (OLS) model fails to correct for multicollinearity in the Kenya and Nigeria portfolios. Forecasting the NPLs for Nigeria showed that the predicted NPLs were closer to the actual variances, though in the presence of multicollinearity, hence unreliable. For Kenya, OLS fails to correct for multicollinearity and is inefficient in predicting NPLs.

Ridge regression managed to solve for multicollinearity and is close to predicting the NPLs in in the Nigeria context but fails to accurately forecast NPLs for the Kenya portfolio. The trajectory of the trends in the Nigeria context mirrors that of the actual NPLs and the variances between the forecasted and actual values were small. These variances can be due to the biasness that is introduced when solving for multicollinearity using ridge regression. For the Kenya dataset, ridge regression fails to predict NPLs with precision and the predicted trends do not mirror the actual trends.

The principal component regression (PCR) model managed to reduce the effect of multicollinearity by producing an effective model that forecasts NPLs, although interpretation

becomes ambiguous. The interpretation of the significant determinant using the PCR model is done through the relation between the dominant original variables in each PC.

The findings significantly contribute to the banking sector's approach of data reduction and adjusting for outliers, correlation, and the multicollinearity effect. On the other hand, the findings also show that ridge regression in the Nigeria context can be used to forecast NPLs using the macroeconomic variables, while in the context of Kenya, more interrogation of the data is needed in order to be able to forecast. In addition, the two countries chosen are extreme representation of what can potentially happen; hence all the other countries would fall within this range.

Further Work

Future work can be expanded for the Partial Least Squares and other robust models such as GARCH (Generalised Autoregressive Conditional Heteroskedasticity), as the nature of the data is that of time series. Some of the variables did not conform to the normality assumption, which resulted in ordinary least squares failing in this regard. Hence other regression or forecasting models that do not conform to the normality assumption could be used in future studies.

Limitations

Although multicollinearity can be present in variables used for binary logistic regression modelling, for the purposes of this work, only solutions applicable to linear regression models in the near-perfect multicollinearity were applied. Heteroscedasticity was not explored methodically as the primary objective of this paper was to solving for multicollinearity.

BIBLIOGRAPHY

1. Adeboye, N.O., Fagoyinbo, I.S. and Olatayo, T.O. 2014. 'Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients', *Journal of Mathematics,* Volume 10, pp. 16-20.
2. Adegoke, A.S., Adewuyi, E., Ayinde, K. and Lukman, A.F. 2016. 'A comparative study of some Robust Ridge and Liu Estimators', *Science World Journal*, Vol 11, No.4, pp. 17-20.
3. Adeola, O. and Ikpesu, F. 2016. 'Macroeconomic Determinants of Non-Performing Loans in Nigeria: An Empirical Analysis', *Proceedings of the International Conference for Bankers and Academics*.
4. Ahmed, S, M., Majeed, E., Thalassinos, E., and Thalassinos, Y. 2021. 'The Impact of Bank Specific and Macro-Economic Factors on Non-Performing Loans in the Banking Sector: Evidence from an Emerging Economy', *Journal of Risk and Financial Management* 14: 217.
5. Al-Hassan, Y.N. 2010. Performance of a new Ridge Regression estimator', *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 9 (1), pp. 23-26.
6. Alibuhtto, M.C. and Peiris, T.S.G. 2015. 'Principal Component Regression for Solving Multicollinearity Problem', *5th International Symposium*, pp. 231-238.
7. Ambra, A.D. and Sarnacchiaro, P. 2010. Some Data Reduction Methods to Analyse the Dependency with Highly Collinear Variable: A Simulation Study', *Asian Journal of Mathematics and Statistics, 3 (2),* pp. 69-81.
8. Akamobi, O.G. and Ugwanna, O.T. 2017. 'Determinants of foreign reserve in Nigeria', *Journal of Economics and Sustainable Development*, Vol 8, No:20, pp. 58-67.Ayinde, K., Lukman, A.F. and Arowolo, O.T. 2015. 'Combined Parameters Estimation Methods of

Linear Regression Model with Multicollinearity and Autocorrelation', *Journal of Asian Scientific Research*, 5(5), pp. 243-250.

9. Bager, A., Roman, M., Algedih, M., and Mohammed, B. 2017. 'Addressing Multicollinearity In Regression Models: A Ridge Regression application', *Journal of Social and Economic Statistics*, Bucharest University of Economic Studies, Vol. 6(1), pp. 30-45.
10. Bagheri, A. and Midi, H. 2009. 'Robust Estimations as a Remedy for Multicollinearity Caused by Multiple High Leverage Points', *Journal of Mathematics and Statistics*, 5(4), pp. 311-321.
11. Bagya, L.H., Gallo, M., and Srinivasan, M.R. 2018. 'Comparison of regression models under multicollinearity', *Electronic Journal of Applied Statistical Analysis*, Vol. 11, Issue 01, pp 340-368. Available from: https://www.researchgate.net/publication/324820109_Comparison_of_regression_models_under_multicollinearity/link/5ae4469b0f7e9b9793c476c8/download [26 July 2020].
12. Belsley, A., Kuh, E., and Welsch, R.E. 2013, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, 2nd Edition.
13. Chai ,T. and Draxler, R.R. 2014. 'Root mean square error (RMSE) or Mean Absolute Error (MAE) – Arguments against avoiding RMSE in the literature', *Geoscientific Model Development*, 7, pp. 1247-1250.
14. Chandrasekhar, C.K, Bagyalakshmi, H., Srinivasan, M.R. and Gallo, M. 2016. 'Partial Ridge Regression under multicollinearity', *Journal of Applied Statistics*, Vol. 43, No.13, pp. 2462-2473.
15. Chatterjee, S. and Hadi, A.S. 2006. *Regression Analysis by Example*. 4th Edition, John Wiley and Sons Inc.

16. Chen, J. and Tindall, M.L. 2014. 'Hedge Fund Replication Using Shrinkage Methodologies', *The Journal of Alternative Investments*, 17, pp. 26-48.
17. Cheong, J., Kwak, D.W. and Tang, K.K. 2014. 'The WTO puzzle, multilateral resistance terms and multicollinearity' *Applied Economics Letters*, Vol.21, No.13, pp. 928-933.
18. Ciampi, F. and Gordini, N. 2008. 'Using Economic-Financial Ratios for Small Enterprise Default Prediction Modelling: An Empirical Analysis', *Oxford Business and Economics Conference Program*.
19. Daoud, J.L. 2017. 'Multicollinearity and Regression Analysis', *Journal of Physics: Conference Series*, 949.
20. Dorugade, A.D. and Kashid, D.N. 2010. 'Alternative Method for Choosing Ridge Parameter for Regression', *Applied Mathematical Sciences*, Volume 9, pp. 447-456.
21. Dorugade, A.D. 2014. 'New Ridge Parameters for Ridge Regression', *Journal of the Association of Arab Universities for Basic and Applied Sciences*, Volume 15, pp. 94-99.
22. Dorugade, A.D. 2016. 'Adjusted ridge estimator and comparison with Kibria's method in linear regression', *Journal of the Association of Arab Universities for Basic and Applied Sciences*, Volume 21, pp. 96-102.
23. Duzan, H. and Shariff, N.S.B.M. 2016. 'Solutions to the Multicollinearity Problem by Adding some Constant to the Diagonal', Journal of Modern Applied Statistical Methods, 15(1), Article 37.
24. El-Dereny, M. and Rashwan, N.I. 2011. 'Solving Multicollinearity Problem Using Ridge Regression Methods', *International Journal of Contemporary Mathematical Sciences*, 6 (12), pp. 585-600.
25. Erickson, M. 2018. 'The Effects of Capping Interest Rate on Profitability of Kenya Commercial Bank', *IOSR Journal of Economics and Finance*, 9(2), pp. 34-37.

26. Farhan, M., Sattar, A., Chaudhry, A.H. and Khalil, F. 2012. 'Economic Determinants of Non-Performing Loans: Perception of Pakistani Bankers', *European Journal of Business and Management*, Vol 4, No.19.
27. Garcia, J., Salmeron, R., Garcia, C., and Martin, M.M.L. 2016. '*Standardization of Variables and Collinearity Diagnostics in Ridge Regression*', International Statistical Review, 84, 2, pp. 245-266.
28. Gidigbi, M.O. 2017. 'An Assessment of the Impact of Banking Reforms on Economic Growth and Bank Performance in Nigeria', *CBN Journal of Applied Statistics*, Vol 8, No.2, 143-162.
29. Gordinsky, A. 2016. 'New Facts in Regression Estimation under Conditions of Multicollinearity', *Open Journal of Statistics*, 6, pp. 842-861.
30. Gorgees, H.M. 2017. 'The Comparison Between Different Approached to Overcome the Multicollinearity Problem in Linear Regression Models', *Ibn Al-Haitham Journal for Pure and Appl. Sci.*, Vol 31.
31. Gujarati D.N., 2003, Basic Econometrics, 4th Edition.
32. http://uweconsoc.com/ols-blue-and-the-gauss-markov-theorem/ [Accessed on 2018-09-15, 7:20PM].
33. Hauke, J. and Kossowski, T. 2011. 'Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the same sets of Data', *Quaestiones Geographicae*, 30(2), pp 87-93.
34. Herawati, N., Nisa, K., Setiawan, E., Nusyirwan, and Tiryono. 2018. 'Regularised Multiple Regression Methods to Deal with Severe Multicollinearity', International Journal of Statistics and Applications. 8(4), pp.167-172.
35. Islam, M.R., 2018. 'Sample Size and Its Role in Central Limit Theorem (CLT)', *Computational and Applied Mathematics Journal.* Vol. 4, No. 1, pp. 1-7.

36. Junttila, V. and Laine, M. 2017. '*Bayesian principal component regression model with spatial effects for forest inventory variables under small field sample size, Remote Sensing of Environment'*, Remote Sensing of Environment, 192, pp. 45-57.
37. Khamdelwal, P., Miyajima, K. and Santos, A. 2016. '*The Impact of Oil Prices on the Banking System in the GCC',* IMF Working Paper, Middle East and Central Asia Department, 16/161.
38. Kumar, A. and Goyal, P. 2011. 'Forecasting of air quality in Delhi using principal component regression technique', *Atmospheric Pollution Research*, Vol 2, 431-444.
39. Kumari, S.S.S. 2008. 'Multicollinearity: Estimation and Elimination', *Journal of Contemporary Research in Management*, pp. 87-95.
40. Lipovetsky, S. 2010. 'Enhanced Ridge Regressions' *Mathematical and Computer Modelling*, 51, pp. 338-348.
41. Lokesh, G., Maurya, S.B., Koutu, G.K., Singh, S.K., Shukla, S.S. and Mishra, D.K. 2017. 'Characterization of rice (Oryza sativa L.) genotypes using Principal Component Analysis including Scree plot & Rotated Component Matrix', *International Journal of Chemical Studies*, 5(4), pp. 975-983.
42. Ma, S. and Dai, Y. 2011. 'Principal Component analysis-based methods in bioinformatics studies', *Briefings in Bioinformatics*, 12(6), pp. 714-722.
43. Maestre, F.T. and Escudero A. 2009. 'Is the patch size distribution of vegetation a suitable indicator of desertification processes?' *Ecological Society of America*, 90 (7), pp. 1729-1735.
44. Mansoon, K. and Shukur, G. 2011. 'A Poisson Ridge Regression estimator', *Economic Modelling*, Volume 28, pp. 1475-1481.

45. Mardikyan, S. and Cetin, E. 2008. 'Efficient Choice of Biasing Constant for Ridge Regression', *International Journal of Contemporary Mathematical Sciences*, 3 (11), pp. 526-536.
46. Marinoiu, C. 2017. 'Classic and Modern in Regression Modelling', *Economic Insights – Trends and Challenges*, Vol 6, pp. 41-50.
47. Messai, A.S. and Jouini, F. 2013. 'Micro and Macro Determinants of Non-Performing Loans', *International Journal of Economics and Financial Issues*, Vol 3, No. 4, pp.852-860.
48. Muniz, G., Kibria, B.M. and Shukur, G. 2012. '*On Developing Ridge Regression Parameters: A Graphical investigation*', Department of Mathematics and Statistics, Florida International University, Volume 10.
49. Muriuki, F., Mathuva, E. and Egondi, P. 2017. 'Influence of Interest Rate Capping on Financial Performance of Commercial Banks in Mombasa County, Kenya', *Imperial Journal of Interdisciplinary Research*, 3 (9).
50. Nduka, E.C. and Ijomah, M.A. 2012. 'The Effects of Perturbing Eigenvalues in the Presence of Multicollinearity', *Electronic Journal of Applied Statistical Analysis*, Vol 5, Issue 2, pp. 304-311.
51. O'Brien, R.M. 2017. 'Dropping Highly Collinear Variables from a model: Why it Typically is Not a Good Idea', *Social Science Quarterly*, Vol.98, No.1.
52. O'Brien, R.M. 2007. 'A Caution Regarding Rule of Thumb for Variance Inflation Factors', *Quality and Quantity*, 41, pp. 673-690.
53. Ogah, D.M. 2011. 'Assessing size and conformation of the body of Nigerian indigenous turkey', *Slovak J. Animal Sciences*, 44 (1), pp. 21-27.

54. Ogunjobi, E.O., Agunbiade, D.A. and Ayansola, O.A. 2017. 'Comparative Analysis of the Efficiencies on Methods of Handling Multicollinearity in Regression Analysis', *Annals. Computer Sciences Series*, Vol 15, No.2.

55. Ongore, V.O. and Kusa, G.B. 2013, 'Determinants of Financial Performance of Commercial Banks in Kenya', *International Journal of Economics and Financial Issues*, 3(1), pp. 237-252.

56. Panik ,M.. 2009. '*Regression Modelling, Methods, Theory and computation with SAS',* Stats Papers, 53, pp. 803-804.

57. Pepler, P.T. 2014. '*The Identification of Common Principal Components'*, Faculty of Economic and Management Sciences at Stellenbosch University.

58. PetroPedia Oil Price. Available from: https://www.petropedia.com/definition/7902/oil-price [Accessed on 3 October 2019].

59. Pindyck, R. S., and Rubinfeld, D. L., 1997, Econometric Models and Economic Forecasts, 4th Edition.

60. Polat, E. and Gunay, S. 2015. 'The comparison of Partial Least Squares Regression, Principal Component Regression and Ridge Regression with Multiple Regression for Predicting PM10 Concentration Level Based on Meteorological', *Journal of Data Science*, 13, pp. 663-602.

61. Polat, E. and Turkan, S. 2016. 'The Comparison of Classical and Robust Biased Regression Methods for Determining Unemployment Rate in Turkey: Period of 1985-2012', *Journal of Data Science*, 14, pp. 739-768.

62. Rulyasri, N., Achsani, N.A. and Mulyati, H. 2017. 'Effects of Macroeconomic Conditions on Non-Performing Loan in Retail Segments: An Evidence from the Indonesian Banking', *International Journal of Scientific and Research Publications*, 7(10), pp 208-217.

63. Shariff, N.S.M. and Ferdaos, N.A. 2017. ‘An application of robust Ridge Regression model in the presence of outliers to real data problem’, *Journal of Physics: Conference Series*, pp. 890.

64. Sheefeni J.P.S. 2015. ‘The Impact of Macroeconomic Determinants on Non-Performing Loans in Namibia’, *International Review of Research in Emerging Markets and the Global Economy,* 1(4).

65. Shlens, J. 2014. ‘*A Tutorial on Principal Component Analysis*’, Google Research.

66. Sinan, A. and Alkan, B.B. 2015. ‘A Useful approach to identify the multicollinearity in the presence of outliers’, *Journal of Applied Statistics*, Vol. 42, No.5, pp. 986-993.

67. Skenderi, N., Islami, X. and Mulolli, E. 2016. ‘The Influence of Macroeconomic Factors in the Failure of Returning Bank Credit in Kosovo’, *Mediterranean Journal of Social Science*, Vol 7, pp. 2039-2117.

68. Standard Bank Group, *Risk and Capital Management Report 2016*. Available from: http://www.ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments [Accessed on 19 January 2018].

69. Standard Bank Group, *Risk and capital management report and annual financial statements 2016*. Available from: http://reporting.standardbank.com/downloads/SBG_FY16_Risk%20and%20capital%20management%20report%20and%20AFS.pdf [Accessed on 19 January 2018].

70. Thompson, C.G., Kim, R.S., Aloe, A.M. and Becker, B.J. 2017. ‘Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from typical Regression Results’, *Basic and Applied Social Psychology*, Vol.39, No.2, 81-90

71. Thupeng, W.M., Mothupi, T., Mokgweetsi, B., Mashabe, B. and Sediadie, T. 2018., ‘A Principal Component Regression Model, for forecasting daily peak ambient ground level ozone concentrations, in the presence of Multicollinearity amongst precursor air pollutants

and local meteorological conditions: A case study of Maun', *International Journal of Applied Mathematics and Statistical Sciences*, Vol. 7, Issue 1.

72. Touny, M.A. and Shebab, M.A. 2015. 'Macroeconomic Determinants of Non-Performing Loans: Empirical Study of Some Arab Countries', *American Journal of Economics and Business Administration*, 11-22.
73. Tyona, T., Tyohemba, S. and Eya, C.I. 2017. 'Macroeconomic Determinants of Non-Performing Loans in Nigeria', *Imperial Journal of Interdisciplinary Research*, 3(7), pp. 662-665.
74. Ugoani, J.N.N. 2016. 'Non-Performing Loans Portfolio and Its Effect on Bank Profitability in Nigeria', *Independent Journal of Management and Production* (IJMandP), Vol 7, No.2.
75. Vlachos, I. and Kugiumtzis, D. 2013. 'Backward-in-Time Selection of the Order of Dynamic Regression Prediction Model', *Journal of Forecasting*, 32, pp. 685–701.
76. Waweru, N.M. 2009. 'Commercial Banking Crises in Kenya: Causes and Remedies', *African Journal of Accounting, Economics, Finance and Banking Research*, 4 (4).
77. Zahari, S.M., Ramli, N.M. and Mokhtar B. 2014. 'Bootstrapped Parameter Estimation in Ridge Regression with Multicollinearity and Multiple Outliers', *Journal of Applied Environmental and Biological Sciences*, 4(7S), pp. 150-156.
78. Zakari, Y., Yau, S.A. and Usman, U. 2018. 'Handling Multicollinearity; A Comparative Study of the Prediction Performance of some Methods Based on Some Probability Distributions', *Annals. Computer Science Series*, Vol 16, No. 1.
79. Zhang. G., Zhang, X. and Fen, H. 2016. 'Forecasting financial time series using a methodology on autoregressive integrated moving average and Taylor expansion', *Expert Systems*, Vol. 33, No.5, 501-516.

ANNEXURE

Annexure A – SAS Program: Nigeria

```
data Nigeria;
input NPL GDP Crude Tbill Icr Lrate Drate M2 FX FXR M1 Infl MPR MLR;
datalines;
16.36 8.4 77.60 3.72 2.61 18.82 12.4 12.4 150.32 42.0 -1.8 14.4 6.0 22.76
16.60 8.4 75.10 2.33 2.27 18.74 10.9 18.8 150.1 41.0 3.0 15.6 6.0 23.33
14.76 8.4 80.30 1.04 1.5 19.03 8.6 22.5 149.78 41.0 6.4 14.8 6.0 23.62
14.96 8.4 85.30 1.2 1.27 19.05 7.3 21.9 150.1 40.0 10.4 15 6.0 23.47
12.13 8.4 77.50 1.63 4.94 18.77 6.2 23.4 150.27 39.0 16.1 12.9 6.0 22.56
11.14 8.4 75.80 2.29 2.73 17.65 5 19.5 149.99 37.0 9.7 14.1 6.0 22.03
11.59 8.4 77.20 2.94 3.59 17.4 4.5 23.1 150.09 37.0 15.2 13 6.0 22.27
9.40 8.4 78.70 2.63 1.26 16.89 4.1 21.6 150.78 37.0 20.1 13.7 6.0 22.31
9.45 8.4 79.50 6.6 2.71 16.66 5.2 18.7 151.35 35.0 21.3 13.6 6.3 22.2
9.53 8.4 84.40 6.75 8.5 16.16 4.3 13.2 149.99 34.0 21.5 13.4 6.3 21.85
9.13 8.4 86.70 7.58 8.79 16.11 5.1 8.8 150.24 33.0 11.7 12.8 6.3 21.84
8.06 8.4 92.80 7.47 8.03 15.74 4.6 6.9 150.66 32.0 11.0 11.8 6.3 21.86
7.00 7.2 98.00 7.49 6.13 15.73 5 10.7 151.85 33.0 20.0 12.1 6.5 21.75
6.80 7.2 106.60 7.09 8.38 15.75 4.6 7.4 152.07 33.0 12.3 11.1 6.5 21.88
6.03 7.2 116.60 8.27 9.33 15.81 5.4 5.7 153.04 33.0 9.2 12.8 7.5 22.02
6.75 7.2 124.50 9.52 10.8 15.75 5.5 8.4 154.45 33.0 11.4 11.3 7.5 22.19
4.54 7.2 118.40 8.63 9.75 15.81 5.4 11.4 155.13 32.0 10.6 12.4 8.0 22.11
3.96 7.2 117.00 8.2 11.15 15.76 5.1 12.2 153.31 32.0 14.6 10.2 8.0 22.02
4.19 7.2 117.90 7.08 8.85 15.84 5.1 13.2 151.96 33.0 18.4 9.4 8.8 22.42
4.12 7.2 112.00 7.41 8.59 15.82 5.4 8.6 153.92 33.0 8.3 9.3 8.8 22.27
4.84 7.2 115.70 8.92 9.37 15.87 5.5 12.4 156.15 32.0 14.2 10.3 9.3 22.09
5.24 7.2 113.10 15 13.07 16.49 7.1 8.4 151.75 33.0 8.7 10.5 12.0 23.32
3.52 7.2 113.90 14.53 15.58 16.82 7.4 9.6 157.87 32.0 9.5 10.5 12.0 23.35
2.77 7.2 111.50 14.27 15.85 16.75 6.8 15.4 158.27 33.0 21.5 10.3 12.0 23.21
4.35 6.3 113.81 14.85 14.19 16.92 7.6 19 158.62 34.0 22.6 12.6 12.0 23.08
4.01 6.3 121.87 14.76 14.35 17.11 8 13.4 157.46 34.0 19.1 11.9 12.0 23.13
4.11 6.3 128.00 14.49 14.13 17.28 8.3 13.9 157.57 35.0 20.2 12.1 12.0 23.21
4.56 6.3 122.62 13.92 14.23 16.9 8.2 11.8 157.26 37.0 18.7 12.9 12.0 23.31
4.65 6.3 113.08 13.34 13.8 16.98 8.6 13.5 157.31 37.0 17.7 12.7 12.0 23.44
4.84 6.3 98.06 14.08 14.92 16.93 7.8 10.8 157.5 35.0 17.1 12.9 12.0 23.44
5.43 6.3 104.62 13.86 15.19 16.96 8.6 8.1 157.4 36.0 9.1 12.8 12.0 23.45
5.63 6.3 113.76 14.26 17.81 16.53 8.3 10.1 157.36 40.0 6.4 11.7 12.0 23.76
5.54 6.3 114.36 12.75 13.5 16.37 8.8 11.5 157.34 41.0 6.5 11.3 12.0 24.67
5.35 6.3 108.92 12.94 11.42 16.48 8.7 18.3 157.27 42.0 12.8 11.7 12.0 24.65
6.01 6.3 111.05 12.6 11.86 16.51 8.9 23.4 157.32 43.0 19.2 12.3 12.0 24.7
6.45 6.3 114.49 11.77 11.88 16.54 9.2 16.4 157.33 44.0 9.6 12 12.0 24.61
6.52 6.2 115.24 11.17 11.67 16.57 8.9 11.3 157.3 46.0 3.7 9 12.0 24.54
7.91 6.2 118.81 9.9 11.98 16.56 8.3 18.2 157.31 47.0 7.7 9.5 12.0 24.6
8.29 6.2 112.79 10.17 10.39 16.61 8 18.1 157.31 48.0 6.4 8.6 12.0 24.49
7.21 6.2 105.55 10.41 11.24 16.65 7.9 17.5 157.31 48.0 1.6 9.1 12.0 24.53
7.74 6.2 106.00 10.64 12.23 16.66 7.7 13.4 157.3 48.0 3.2 9 12.0 24.57
```

```
7.83  6.2   106.06 11.6  11.59 16.56 7.5   15.7  157.31 45.0 5.2  8.4     12.0  24.58
7.68  6.2   109.78 11.56 10.63 16.47 7.8   10.6  157.32 46.0 1.9  8.7     12.0  24.62
9.59  6.2   107.84 11.3  15.24 16.55 8     6.2   157.32 45.0 0.4  8.2     12.0  24.46
8.38  6.2   113.59 10.91 16.88 16.76 7.4   2.1   157.31 44.0 -1.5 8       12.0  25.11
7.99  6.2   112.29 10.8  11.08 17.1  7.5   0.9   157.36 44.0 -1.3 7.8     12.0  24.9
7.65  6.2   111.14 10.8  11.23 17.17 8.3   -2.2  157.28 43.0 -7.3 7.9     12.0  25
7.46  6.2   112.75 10.97 10.75 17.01 8     1.3   157.26 43.0 -5.2 8       12.0  24.9
7.78  3.8   110.19 10.81 10    16.95 9.3   1.2   157.31 41.0 -4.2 8       12.0  25.52
8.14  3.8   110.83 11.82 10.5  16.93 9.5   -0.8  157.31 37.0 -2.5 7.7     12.0  25.83
7.82  3.8   109.47 11.92 10.5  16.69 9.5   13.2  157.3  37.0 9.8  7.8     12.0  25.8
6.76  3.8   110.41 11.26 10.5  16.7  9.4   14.4  157.29 37.0 13.3 7.9     12.0  25.63
7.26  3.8   111.90 10.13 10.63 16.5  9.4   14.3  157.29 35.0 9.0  8       12.0  25.76
6.93  3.8   114.60 9.98  10.5  16.5  9.3   12.7  157.29 37.0 2.3  8.2     12.0  26.07
5.85  3.8   109.63 9.88  10.5  16.44 9.4   22.2  157.29 39.0 10.7 8.3     12.0  26.07
6.26  3.8   102.33 9.95  11.91 16.6  8.5   22.4  157.29 40.0 12.6 8.5     12.0  25.07
6.22  3.8   98.27  9.75  10.73 16.44 9.3   26.7  157.31 40.0 16.8 8.3     12.0  25.77
7.38  3.8   83.50  9.83  10.98 16.48 9.3   27.3  157.32 39.0 12.4 8.1     12.0  25.75
7.94  3.8   80.42  9.82  8.98  16.47 9.7   27.6  166.65 37.0 11.1 7.9     13.0  25.74
6.86  3.8   63.28  10.8  24.3  15.88 9.5   20.6  169.68 34.0 -1.8 8       13.0  25.91
7.28  1.5   48.81  11.2  10.21 16.86 9.6   22.4  169.7  34.0 4.8  8.2     13.0  25.97
7.69  1.5   58.09  10.88 23.5  16.77 9.5   22.3  197.5  31.0 -0.2 8.4     13.0  26.33
7.91  1.5   56.69  10.77 12.59 16.9  9     7.9   196.5  30.0 -8.3 8.5     13.0  26.61
8.43  1.5   57.45  10.23 24.24 15.95 8.7   9.6   196.5  30.0 -6.4 8.7     13.0  26.41
8.73  1.5   65.08  10.03 10.43 16.08 8.9   8.9   196.5  30.0 -9.2 9       13.0  26.43
9.01  1.5   62.06  9.95  10.85 17.24 10.3  7     196.5  29.0 -7.8 9.2     13.0  26.84
8.80  1.5   57.01  10    9.69  17.3  10.3  1.8   196.5  31.0 -10.4 9.2     13.0  27.03
9.47  1.5   47.09  10    27.92 17.29 10.3  3.3   196.5  31.0 -1.3 9.3     13.0  27.01
9.94  1.5   48.08  10.36 8.12  17.02 10.6  2.8   196.5  30.0 -2.8 9.4     13.0  26.99
9.92  1.5   48.90  9.11  3.22  16.84 9.1   -1.6  196.5  30.0 -7.9 9.3     13.0  27.01
9.69  1.5   44.82  5.62  0.84  16.98 6.5   -2.3  196.5  30.0 -1.5 9.37    11.0  27.02
10.12 1.5   37.80  4.57  0.77  16.96 6.9   5.9   196.5  29.0 24.1 9.55    11.0  26.84
10.55 -1.7  30.66  4.12  2.04  16.54 6.7   4.4   196.5  28.0 16.1 9.62    11.0  26.77
12.98 -1.7  31.70  4.91  2.67  16.72 6.8   9.3   196.5  28.0 34.8 11.38   11.0  26.73
10.89 -1.7  37.76  5.53  4.32  16.82 6.9   7     196.5  28.0 29.5 12.77   12.0  26.93
11.65 -1.7  41.60  7.27  3.75  16.77 6.8   5.7   196.5  27.0 27.2 13.72   12.0  26.88
12.04 -1.7  47.01  8.04  7.67  16.13 7     7.9   196.5  26.0 40.8 15.58   12.0  26.73
11.54 -1.7  48.46  8.32  35.26 16.78 6.9   17.4  282.5  26.0 45.5 16.48   12.0  26.93
12.01 -1.7  45.92  12.34 31.51 17.14 7.4   22.3  312.5  26.0 48.1 17.13   14.0  27.06
12.74 -1.7  46.15  14.93 24.25 17.18 8     19.3  305.5  25.0 36.8 17.61   14.0  27.21
13.90 -1.7  47.43  14    14.5  17.09 7.7   17.6  304.8  25.0 37.5 17.85   14.0  27.49
14.71 -1.7  51.00  13.96 36.42 17.1  8.3   21.8  304.5  24.0 48.4 18.33   14.0  27.69
11.77 -1.7  45.25  13.99 15.21 17.06 8.6   21.9  304.5  25.0 49.5 18.48   14.0  28.53
12.75 -1.7  53.48  13.96 10.39 17.09 8.8   17.8  304.5  26.0 31.5 18.55   14.0  28.55
;
run;
proc means data=Nigeria;
run;
```

```
**********Ordinary Least Regression with VIFs**********;
ods graphics on;
proc reg data=Nigeria plots=residuals plots=residualbypredicted;
model NPL = GDP Crude Tbill Icr Lrate Drate M2 FX FXR M1 Infl MPR MLR/vif tol collin;
run;

/*reduced model with only significant variables*/;
proc reg data=Nigeria plots=residuals plots=residualbypredicted;
model NPL = Crude Lrate M2 FX FXR MPR/vif;
run;
ods graphics off;

**********Ordinary Least Regression post Variable Selection with VIFs**********;
ods graphics on;
proc reg data=Nigeria plots=residuals plots=residualbypredicted;
model NPL = Crude Icr Lrate Drate M2 FXR M1 Infl/vif collin;
run;

/*reduced model with only significant variables*/;
proc reg data=Nigeria plots=residuals plots=residualbypredicted;
model NPL = Crude Lrate FXR Infl/vif collin;
run;
ods graphics off;

**********Standardising the Variables**********;
proc standard data = Nigeria mean=0 std=1 out=Nigeria1;
var NPL GDP Crude Tbill Icr Lrate Drate M2 FX FXR M1 Infl MPR MLR;
run;

proc means data=Nigeria1;
run;

**********Ridge Regression with standardized
variables**********
**********;
ods graphics on;
proc reg data=Nigeria1 outvif plots=residuals plots=residualbypredicted
      outest=b1 ridge=0 to 0.2 by 0.002;
  model NPL = GDP Crude Tbill Icr Lrate Drate M2 FX FXR M1 Infl MPR MLR;
run;
proc print data=b1;
run;

proc reg data= Nigeria1 outvif rsquare plots=residuals plots=residualbypredicted
     outest=b2 ridge=0.11 outseb;
 model NPL = GDP Crude Tbill Icr Lrate Drate M2 FX FXR M1 Infl MPR MLR/vif;
```

```
run;
proc print data=b2;
run;

/*reduced model with only significant variables*/;
proc reg data= Nigeria1 outvif rsquare plots=residuals plots=residualbypredicted
      outest=b3 ridge=0.11 outseb;
model NPL = Crude Lrate M2 FX FXR MPR/vif;
run;

proc print data=b3;
run;
ods graphics off;

**********************************************************Principal Component
Analysis with standardized
variables*************************************************************************
**********;
ods graphics on;

proc princomp data=Nigeria1 cov out=prin;
var GDP Crude TBill ICR LRate DRate M2 FX FXR M1 Infl MPR MLR;
run;

proc reg;
model NPL=prin1 - prin13/vif;
run;

proc reg plots=residuals plots=residualbypredicted;
model NPL=prin1 prin2 prin3 prin4 prin5 prin7 prin8 prin11 prin13/vif;
run;

proc reg plots=residuals plots=residualbypredicted;
model NPL=prin1 prin2 prin3 prin4/vif;
run;
ods graphics off;


**************************************************************Correlations
**********************************************************;


ods graphics on;

title 'Descriptive Statistics1';

proc corr data=Nigeria;

var NPL GDP Crude TBill ICR LRate DRate M2 FX FXR M1 Infl MPR MLR;

run;
```

```
ods graphics off;

ods graphics on;
title 'Descriptive Statistics 2';
proc corr data=Nigeria nomiss plots=matrix (histogram);
var NPL GDP Crude TBill ICR;
run;
ods graphics off;

ods graphics on;
title 'Descriptive Statistics 3';
proc corr data=Nigeria nomiss plots=matrix (histogram);
var LRate DRate M2 FX FXR;
run;
ods graphics off;

ods graphics on;
title 'Descriptive Statistics 4';
proc corr data=Nigeria nomiss plots=matrix (histogram);
var M1 Infl MPR MLR;
run;
ods graphics off;

**********Checking for Normality**********;
proc univariate data= Nigeria Normal;
var NPL GDP Crude TBill ICR LRate DRate M2 FX FXR M1 Infl MPR MLR;
run;
```

```
**********Checking for Autocorrelations on the error terms - Durbin Watson Test**********;

proc autoreg data=Nigeria;

  model NPL = GDP Crude TBill ICR LRate DRate M2 FX FXR M1 Infl MPR MLR / dw=4 dwprob;

run;


**********Checking for Stationarity**********;

proc arima data=Nigeria;

identify var= NPL stationarity=(adf=1);

run;
```

Annexure B – SAS Program: Kenya

```
data Kenya;
input NPL GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate;
datalines;
3.87 6.3 6.0 6.43 13.80 5.20 16.60 70.50 2.50 22.90 4.63 10.00
3.15 6.3 6.22 6.52 13.60 5.10 15.80 69.70 2.50 21.60 3.02 10.00
3.94 6.3 6.32 6.55 13.60 5.20 16.90 68.80 2.60 21.30 2.19 10.00
3.76 8.9 6.65 6.81 13.40 5.10 14.30 68.30 2.70 11.30 1.85 10.00
3.59 8.9 6.77 7.11 13.40 5.10 14.70 67.00 2.70 28.40 1.96 10.00
11.55 8.9 6.53 6.98 13.10 5.10 16.00 66.60 2.70 31.10 4.07 8.50
9.52 6.9 6.52 7.07 13.30 5.20 15.10 67.50 2.80 27.40 5.48 8.50
9.47 6.9 7.30 7.38 13.00 5.30 18.40 67.00 2.80 33.60 5.30 8.75
9.47 6.9 7.35 7.59 12.90 5.30 17.40 67.00 2.80 30.90 5.53 8.75
9.81 6.9 7.55 7.65 13.20 5.10 16.50 67.10 2.90 24.50 5.38 8.75
15.48 6.9 7.52 6.50 13.40 5.10 16.20 64.40 3.00 26.40 6.08 8.75
10.9 6.9 6.87 7.05 13.30 5.20 20.40 62.70 3.40 27.90 5.70 8.75
5.29 1.4 6.95 7.66 13.80 5.10 22.50 70.60 3.50 29.00 9.40 8.75
4.4 1.4 7.28 7.18 13.80 5.10 23.50 69.00 3.50 30.80 10.58 8.75
5.58 1.4 6.90 6.35 14.10 5.20 21.00 62.80 3.40 29.30 11.90 8.75
5.23 3.2 7.35 6.59 13.90 5.10 20.70 62.10 3.40 32.00 16.12 8.75
5.0 3.2 7.76 7.72 14.00 5.10 21.00 62.00 3.40 21.40 18.61 8.75
12.83 3.2 7.73 7.79 14.10 5.20 18.30 64.70 3.40 15.60 17.87 9.00
11.39 2.8 8.03 8.07 13.90 5.20 17.40 67.30 3.40 12.50 17.12 9.00
11.9 2.8 8.02 6.92 13.70 5.40 15.10 68.70 3.30 8.90 18.33 9.00
11.9 2.8 7.69 6.70 13.70 5.20 16.70 73.20 3.20 10.10 18.73 9.00
10.53 1.5 7.75 6.81 14.10 5.40 18.60 79.70 2.90 13.60 18.74 9.00
19.08 1.5 8.39 6.83 14.30 5.90 16.80 77.90 2.90 9.10 19.54 9.00
13.48 1.5 8.59 6.67 14.90 5.70 14.90 77.70 2.90 5.20 17.83 8.50
```

13.13 5.6 8.46 5.95 14.80 6.00 11.90 79.50 2.80 5.50 13.22 8.50
14.22 5.6 7.55 5.49 14.70 6.10 11.00 79.70 2.70 0.40 14.69 8.50
14.22 5.6 7.31 5.57 14.90 5.90 12.00 80.40 2.70 7.40 14.60 8.25
11.61 2.1 7.34 5.81 14.70 6.00 12.90 78.70 2.90 -3.00 12.42 8.25
9.61 2.1 7.45 5.55 14.90 6.00 12.20 78.30 2.90 -0.60 9.61 8.00
9.26 2.1 7.33 3.08 15.10 6.00 13.40 77.20 3.20 2.30 8.60 8.00
8.89 0.5 7.24 2.69 14.80 6.00 15.20 76.60 3.20 11.80 8.44 7.75
9.02 0.5 7.25 3.68 14.80 5.90 14.90 76.20 3.60 13.00 7.36 7.75
8.72 0.5 7.29 3.38 14.70 6.10 15.30 75.70 3.70 12.50 6.74 7.75
7.96 2.7 7.26 2.57 14.80 5.80 15.60 75.00 3.80 12.30 6.62 7.75
7.93 2.7 7.22 3.11 14.90 6.00 17.10 74.60 3.90 12.10 5.00 7.00
6.06 2.7 6.82 2.95 14.80 5.90 17.20 75.80 3.80 12.60 5.32 7.00
6.76 6.6 6.56 3.69 15.00 5.80 20.10 75.90 3.80 16.20 5.95 7.00
5.86 6.6 6.21 2.39 15.00 5.60 22.10 76.90 3.70 20.30 5.18 7.00
5.76 6.6 5.98 2.21 15.00 5.30 22.90 77.30 3.70 13.90 3.97 6.75
7.34 7.6 5.17 2.46 14.60 4.90 22.10 77.20 3.80 24.50 3.66 6.75
7.55 7.6 4.21 2.16 14.50 5.20 25.60 78.40 3.80 25.00 3.88 6.75
7.53 7.6 2.98 1.15 14.40 5.10 27.30 80.90 3.80 27.70 3.49 6.75
7.93 7.9 1.60 1.35 14.30 4.20 26.00 80.20 4.20 22.20 3.57 6.00
7.25 7.9 1.83 1.66 14.20 3.90 25.50 81.10 4.30 18.70 3.22 6.00
6.07 7.9 2.04 1.18 14.00 3.60 27.00 80.80 4.40 23.90 3.21 6.00
5.86 11.6 2.12 0.98 13.90 3.70 24.70 80.80 4.40 23.70 3.18 6.00
5.62 11.6 2.21 1.01 14.00 3.60 23.80 81.00 4.30 27.60 3.84 6.00
5.36 11.6 2.28 1.18 13.90 3.90 22.40 80.80 4.30 30.50 4.51 6.00
5.42 7.5 2.46 1.24 14.00 3.70 21.50 81.30 4.30 24.30 5.42 5.75
4.8 7.5 2.59 1.13 13.90 3.70 20.50 82.40 4.40 28.00 6.54 5.75
4.34 7.5 2.77 1.24 13.90 3.90 19.40 83.00 4.20 29.70 9.19 6.00
4.04 6.6 3.26 3.97 13.90 3.90 18.20 83.40 4.20 24.30 12.05 6.00
3.51 6.6 5.35 5.54 13.90 4.10 16.60 85.70 4.20 23.30 12.95 6.00
3.09 6.6 8.95 6.36 13.90 4.40 14.50 89.90 4.20 21.20 14.48 6.25
2.81 6.1 8.99 8.61 14.10 4.70 14.70 91.10 4.20 19.60 15.53 6.25
2.64 6.1 9.23 14.29 14.30 5.50 15.20 93.60 4.20 20.40 16.67 6.25
2.87 6.1 11.93 7.46 14.80 7.00 14.30 99.80 4.00 16.90 17.32 7.00
3.05 4.4 14.80 14.95 15.20 7.00 14.00 99.80 4.00 19.60 18.91 11.00
3.15 4.4 16.14 28.90 18.50 8.90 13.80 89.70 4.00 12.40 19.72 16.50
3.14 4.4 18.30 21.75 20.00 10.90 14.10 85.10 4.30 7.90 18.93 18.00
3.16 4.2 20.56 19.27 19.50 11.50 10.60 84.60 4.10 5.30 18.31 18.00
3.38 4.2 19.70 18.15 20.30 12.40 11.20 83.00 4.40 5.70 16.69 18.00
3.61 4.2 17.80 24.02 20.30 12.10 11.50 83.10 4.70 1.40 15.61 18.00
3.85 4.3 16.01 16.15 20.20 13.70 13.00 83.20 5.00 6.10 13.06 18.00
4.24 4.3 11.18 17.16 20.10 12.80 12.50 86.80 4.50 1.70 12.22 18.00
4.09 4.3 10.09 17.09 20.30 12.10 13.10 84.20 5.30 0.60 10.05 18.00
4.26 5.0 11.95 13.71 20.10 13.50 13.90 84.20 5.30 2.30 7.74 16.50
4.38 5.0 10.93 8.97 20.10 12.10 15.00 84.30 5.40 4.20 6.09 16.50
3.86 5.0 7.77 7.02 19.70 10.60 14.30 85.30 5.50 6.30 5.32 13.00
3.92 4.7 8.98 9.14 19.00 9.60 14.80 85.20 5.50 3.90 4.14 13.00
4.11 4.7 9.80 7.14 18.70 9.30 18.10 85.90 5.80 9.30 3.25 11.00
4.29 4.7 8.30 5.84 18.20 9.20 17.20 86.00 5.70 14.10 3.20 11.00
3.87 6.1 8.08 5.86 18.10 8.80 18.20 87.60 5.30 16.00 3.67 9.50
3.93 6.1 8.38 9.25 17.80 8.40 17.00 86.20 5.50 15.50 4.45 9.50

```
4.19 6.1 9.88 8.93 17.80 8.60 15.70 85.60 6.00 17.80 4.11 9.50
4.22 7.5 10.38 7.90 17.90 8.60 18.50 83.80 6.00 20.00 4.14 9.50
4.84 7.5 9.46 7.16 17.50 9.00 17.80 85.10 6.20 22.10 4.05 8.50
3.91 7.5 6.21 7.14 17.00 8.80 15.50 86.00 6.10 20.80 4.91 8.50
3.93 6.4 5.92 7.93 17.00 8.70 13.90 87.30 6.10 18.60 6.03 8.50
3.89 6.4 10.03 8.88 17.00 8.40 13.80 87.60 6.10 17.70 6.67 8.50
3.88 6.4 9.58 7.52 16.90 8.40 13.10 86.60 6.30 16.70 8.29 8.50
3.8 3.5 9.72 10.66 17.00 8.40 12.20 85.10 6.30 14.90 7.76 8.50
3.62 3.5 9.94 10.77 16.90 8.80 11.00 87.00 6.30 14.90 7.36 8.50
3.88 3.5 9.52 8.98 17.00 9.00 11.20 86.30 6.60 10.90 7.15 8.50
4.44 5.2 9.26 10.43 17.00 8.70 13.70 86.20 6.60 13.60 7.21 8.50
4.12 5.2 9.16 8.83 17.10 8.80 15.00 86.30 6.60 14.40 6.86 8.50
3.34 5.2 8.98 6.47 16.90 8.30 16.00 86.40 6.70 14.20 6.27 8.50
3.61 6.0 8.80 7.40 16.70 8.30 13.30 86.90 6.80 11.20 6.41 8.50
3.68 6.0 8.82 7.76 17.00 8.10 17.30 87.80 6.50 17.40 7.30 8.50
3.86 6.0 9.81 6.60 16.40 8.10 18.50 87.60 8.60 20.50 7.39 8.50
3.53 4.6 9.78 8.08 16.90 8.60 18.80 87.80 8.10 18.80 7.67 8.50
3.36 4.6 8.29 11.79 16.30 7.70 20.10 88.40 7.90 20.80 8.36 8.50
3.44 4.6 8.38 7.43 16.00 8.50 18.80 89.30 7.70 16.00 6.60 8.50
3.69 5.6 8.67 6.73 16.00 8.60 19.90 89.40 7.70 15.90 6.43 8.50
3.76 5.6 8.64 6.86 15.90 8.10 20.30 90.20 7.30 18.40 6.09 8.50
3.66 5.6 8.58 6.91 16.00 8.80 21.40 90.60 7.90 18.80 6.02 8.50
3.98 5.7 8.59 7.12 15.90 8.50 20.10 91.70 7.60 17.50 5.53 8.50
3.89 5.7 8.59 6.77 15.50 8.60 20.10 91.40 7.90 15.70 5.61 8.50
3.85 5.7 8.49 6.85 15.50 8.50 19.40 92.30 7.80 18.00 6.31 8.50
3.79 5.6 8.42 8.77 15.40 8.10 20.20 94.60 7.50 19.30 7.08 8.50
4.21 5.6 8.26 11.17 15.30 8.50 15.90 97.80 7.50 12.20 6.87 8.50
4.06 5.6 8.26 11.78 15.50 8.30 16.60 98.60 7.30 10.20 7.03 10.00
3.98 6.1 10.57 12.89 15.80 7.90 15.20 102.50 7.00 11.30 6.62 11.50
3.89 6.1 11.54 18.80 15.70 9.20 14.30 103.90 7.00 10.50 5.84 11.50
4.03 6.1 14.61 19.85 16.60 10.10 13.00 105.30 6.80 9.20 5.97 11.50
4.14 5.5 21.65 14.82 16.60 10.60 35.00 101.80 7.30 10.80 6.72 11.50
4.53 5.5 12.34 8.77 17.20 10.90 33.90 102.10 7.20 7.90 7.32 11.50
4.86 5.5 9.81 7.27 18.30 11.10 34.60 102.30 7.50 8.50 8.01 11.50
;
run;

proc means data=kenya;
run;

**********Ordinary Least Regression with VIFs**********;
ods graphics on;
proc reg data=Kenya plots=residuals plots=residualbypredicted;
model NPL = GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate/vif tol collin ;
run;

*****model reduced and fitted with only significant variables*****;
proc reg data=Kenya plots=residuals plots=residualbypredicted;
model NPL = Lrate FX M1/vif tol collin;
```

```
run;
ods graphics off;

**********Ordinary Least Regression post Variable Selection with VIFs**********;
ods graphics on;
proc reg data=Kenya plots=residuals plots=residualbypredicted;
model NPL = GDP M2 FX M1 Infl/vif collin;
run;

proc reg data=Kenya plots=residuals plots=residualbypredicted;
model NPL = M2 FX M1 Infl/vif collin;
run;
ods graphics off;

**********Standardising the Variables**********;
proc standard data = Kenya mean=0 std=1 out=Kenya1;
var NPL GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate;
run;

proc means data=Kenya1;
run;

**********Ridge Regression with standardized
variables**********
**********;
ods graphics on;
proc reg data=Kenya1 outvif plots=residuals plots=residualbypredicted
      outest=b1 ridge=0 to 0.2 by 0.02;
  model NPL = GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate;
run;

proc print data=b1;
run;

proc reg data= Kenya1 outvif rsquare plots=residuals plots=residualbypredicted
   outest=b2 ridge=0.12 outseb;
model NPL = Lrate FX M1/vif;
run;

proc print data=b2;
run;

ods graphics off;


**********Principal Component Analysis with standardized
```

```
variables**************************************************************************
************;
ods graphics on;

proc princomp data=Kenya1 cov out=prin;
var GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate;
run;

proc reg;
model NPL=prin1 - prin11/vif;
run;

*****reduced model***;
proc reg;
model NPL=prin1 prin2 prin3 prin6 prin9 prin11/vif;
run;

proc reg plots=residuals plots=residualbypredicted;
model NPL=prin1 prin2/vif;
run;

ods graphics off;


**************************************************************Correlations
************************************************************;

ods graphics on;
title 'Descriptive Statistics1';
proc corr data=Kenya;
var GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate;
run;
ods graphics off;

ods graphics on;
title 'Descriptive Statistics 2';
proc corr data=Kenya nomiss plots=matrix (histogram);
var GDP Tbill Icr Lrate Drate ;
run;
ods graphics off;

ods graphics on;
title 'Descriptive Statistics 3';
proc corr data=Kenya nomiss plots=matrix (histogram);
var M2 FX FXR M1 Infl Crate;
run;
ods graphics off;
```

```
**********Checking for Normality**********;
proc univariate data=Kenya normal;
var NPL GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate;
run;


**********Checking for Autocorrelations on the error terms - Durbin Watson Test**********;
proc autoreg data=Kenya;
  model NPL = GDP Tbill Icr Lrate Drate M2 FX FXR M1 Infl Crate / dw=4 dwprob;
run;

**********Checking for Stationarity**********;

proc arima data=Kenya;
identify var= NPL stationarity=(adf=1);
run;
```