DETERMINATION OF SOME FACTORS THAT INFLUENCES LOAN DEFAULT PAYMENT

CASE STUDY: CUSTOMERS FROM AKATAKYIMAN RURAL BANK LTD KOMENDA

BY

GEORGE YAW MENSAH BSC. STATISTICS (Hons.)

A Thesis submitted to the Department of Mathematics, Kwame Nkrumah University of Science and Technology. In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

(INDUSTRIAL MATHEMATICS)

INSTITUTES OF DISTANCE LEARNING

JUNE 2012

Declaration

I, George Yaw Mensah hereby declare that this submission is my own work towards the MSc Industrial Mathematics and that, to the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for the award of another degree of the University, except where due acknowledgment has been made in the text.

George Yaw Mensah (PG 3015309)	……	……..
Student Name and ID	Signature	Date

Certified by:

Nana Kena Frempong	……	……..
(Supervisor)	Signature	Date

Certified by:

Mr. F. K. Darkwah	……	……..
(Head of Dept.)	Signature	Date

Dedication

This work is dedicated

to

my loving wife, Rosemond and children Laurene and Percy and my parents who with their love and care encouraged me in my educational endeavours.

Acknowledgment

My deepest and sincere gratitude go to my supervisor Nana Kena Frempong for spending a substantial part of his time reading my manuscript, criticizing where necessary, explaining the criticism to my understanding and offering very priceless suggestions and pieces of advice.

I further wish to express my sincere gratitude and appreciation to my parents for their support and encouragement they gave me throughout my studies.

I wish to thank Mr. And Mrs. Scot Edwin Baiden for their support more especially making their printer available to me for my work.

I wish also to thank my colleagues in my department at my work place for being there for me where the need arose.

Finally, I wish to thank my study mates; Eric Afful Mensah, Joseph Edzie-Dadzie, Emmanuel Adobah and Patrick Amoako for their priceless suggestions and pieces of advice.

Abstract

Loan Default is the failure of an applicant to fulfil his/her obligation with respect to repayment of loans. Loan default lowest the financial capacity of the agency to fulfil its promises to other applicants.

This study seeks to determine some risk factors that influence loan default repayment among customers in Akatakyiman Rural Bank Ltd –Komenda. To this end, some secondary data on some variables which influenced whether a customer defaulted or not in a loan accessed, was obtained from the credit department of Akatakyiman Rural Bank Ltd –Komenda. A total of 100 observations for a period of four (4) years (2006-2010). There were eleven (11) variables in the data set. A logistic regression model was fitted to the data. It was found that among the variables that were used, Security and Type of Loan were significant to the study where as Sex, Marital Status, Age, Educational Level, Town were not significant to the study. We conclude that the risk of default for a customer who used collateral as a security in accessing the loan is less than for a customer who used personal guarantee. Taking transport loan as a reference group, the risks of a customer defaulting when given a personal loan is less than when given a transport loan, all other factors being equal.

Key words: Loan default, logistic regression, Risk

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

NPLs	-	Non-performing Loans
FCSECO	-	Farm Credit Service of East Central Oklahoma
PLS	-	Profit-and-Loss Sharing
DER	-	Debt-Equity Ratio
DTAR	-	Debt to total Assets
EM	-	Earning Multiplier
RWA	-	Lagged Risk-Weighted Asset
LLP	-	Loan Loss Provision
LTV	-	Lower Loan-to-Value Ratios
GPA	-	Grade Point Average
SPSS	-	Statistical Package for Social Sciences
OR	-	Odds Ratio

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The effect of monetary policy and the banking institution in Ghana and in Africa as a whole cannot be over emphasized. The role of the banking institution on the socio economic live of the citizenry has been very significant and worth nothing especially over the past decades. On the whole, financial institution has had an immense effect on real economic activity by affecting the supply of bank loans. Since most bank loans are important source of external finance within the present day Ghanaian economy. Crockett,(1996) asserts that disruption to loan supply might cause great changes in economic activity.

The role of rural banks in the socio-economic life of people has been seen to have a positive effect on the entire economy. People who under normal circumstances would not have had access to money for investment have been able to do so under the credit creation function of the rural banks. To this end bank customers who are very industrious have been able to expand their businesses.

In Ghana, over the past ten years, there has been an unprecedented springing up of financial institutions. The dynamics of economics is such that once interest rate begins to decline, it becomes attractive to borrow from the banks for trading purpose as the returns will far outweigh the interest rate payments. Both sole traders and corporate bodies are encouraged to borrow from the bank. However should interest rise to

astronomical levels, people are rather encourage to save in the banks or buy fixed interest bonds whose returns are higher vis-a-viz borrowing for capital investment.

The banking institutions in spite of the risk associated with giving out loans are increasingly expanding their area of their services. Currently there are institutions that employ full time personal with sole duties of selling bank loans. They move from office to office appealing to the several public/private customers to come and take loans. As a result of this, some private entrepreneurs has been able to expand their businesses and earn substantial profit. The whole process has ripple effect on the entire economy. If individuals establish businesses, they employ other people to help them in the day to day running of activities.

The banking institutions has advance loans to a wide category of customers while some is given directly to corporate institutions, large portion goes to private individual. The banks have various forms of valuation method to access the credit worthiness of the customers. This area is very critical if the banks are to be able to recover their loans when the time is due. Most banks therefore have credit department that carry out all these important functions. However, despite the laudable ideas of these banks, there are quite a large number of customers who are unable to pay the interest that accrues on the loan. This situation has been so alarming to the extent that some financial institution have adopted some unconventional means of retrieving these loans. In fact there are reported cases of situation where some workers of a bank have been held directly responsible or being given ultimatum to retrieve loans or be reprimanded although they followed all the laid down procedure in advancing such loans. Fama (1980).

Akatakyiman Rural bank limited commenced operations in August, 1983 and is considered to be one of the most innovated rural banks in Central Region. It has about five thousand four hundred and forty (5440) shareholders with its share capital being Eighteen thousand, five hundred and four (18,504.00) Ghana cedis as of May, 2011. The mission statement of the bank is to aspire to be financial services institution of preference through delivery of quality service, using innovative technology and skilled personnel to achieve sustainable growth and enhance stakeholder value. The Bank's vision is to be a leading financial services group creating sustainable value for our stakeholders.

The following services are rendered by the bank;

1. Personal banking: These are current account service, salary account and savings deposit account service, loans, local payment service, treasury bills, fixed deposits.
2. Financial advice to customers.
3. Money Transfers (Apex money transfer, Western Unions and E-Swizch)

Loan making procedure and conditions: As much as possible, applicants are to operate or maintain an account with a branch of the bank. Such application for credit shall be directed at the bank's main branch since the bank is not networked for its operation.

Factors considered in loan applications are as follows: The following are the factors to consider when applying for a loan; Applicant's background, purpose of the request, the amount of credit required, amount and source of borrower's contribution, the repayment terms of borrower, collateral security proposed, business and project location, technical and financial soundness of credit proposal.

Security requirements: The following can be used as collateral for the loan being applied for. Fixed deposit, Blocked savings/current accounts, guarantees, bills, bonds and note, life insurance, mortgage of real estates, Assignment of contract funds, Assignment of assets, stocks and inventory and assignment of account receivable .

The financial system consist of many different types of private financial institutions including banks, insurance companies, mutual funds, finance companies and investment banks. The government or an appointed body strictly regulates all these institutions, irrespective of the countries where they operate.

In recent years, the process of financial innovation has advanced enormously increasing the importance and profitability of non bank finance. Such profitability priority restricted to the non banking industry, has prompted the office of the controller of the currency (OCC) to encourage banks to explore other financial instruments, diversifying bank's business as well as improving banking economic health. Hence as the distinct financial instruments are being explored and adopted by the banking and non banking industries the distinction between different financial institutions are gradually vanishing.

1.2 Statement of the Problem

The issue of loan default has become an issue in the financial circle all over the world. Financial expert are still researching various ways of addressing this problem. Over the years there has been a debate as to which method works best. The consensus among these experts is that no one method stands out, the choice is independent on other factors such as economic stability and the effectiveness and the dependability of the national

data base. As one of vibrant rural banks in the Central Region, the Akatakyiman rural bank limited has over the years also engage in all this important role of credit creation by way of advancing loan. Since it inceptions credit, the bank has extended credit facility to a wide variety of customers, the problem with loan default is not different from Akatakyiman rural bank's operations. The bank has also been devising various means of addressing this anomaly. The researcher will attempt to use logistic regression analysis to determine some risk factors influencing loan default among customers in Akatakyiman rural bank ltd.

1.3 Objectives of the Study

The main objective of the study is to fit a logistic regression model of the repayment status of the loan customers data in Akatakyiman rural bank.

The Specific objective is;

1. To determine the risk(s) factors that have impact on repayment status of the loan customers.

1.4 Significance of the Study.

The study has been necessitated by the quest for an in depth research work in the area of banking in Ghana. The following has particularly influenced the choice of the topic.

1. Recent development in the banking sector couple with its ripple effects on the capital market.
2. The spate of new financial institution emerging the Ghanaian economy coupled with continuous reliance on bank loans to capital investment.

3. The general increase in granting of loan facilities by all banking institution in the Ghanaian economy.

1.5 Methodology of the Study

The tool used in the research analysis includes logistic regression model. SPSS will also be used in the data analysis.

1.5.1 Source of Data

The data was collected from the credit department of the bank, since they keep records of all the banks loans customers and other relevant information. It was for a period of four (4) years (i.e. 2006-2010), the relevant data for the research were Repayment Status, Age, Marital Status, Sex, Security, Town dummy, Interest Rate, Type of Loan and Educational Level of all the customers.

1.5.2 Assumption Associated with Data Collection

The loans that were not paid within the repayment period were assumed to be defaulted. Although some of them are in the long run realized within an extended period and others are written off. i.e. loss. Also, collaterals used for security such as plots of land, office properties, building properties, assignment of account receivable e.t.c were assumed to be mortgaged.

1.6 Limitations

Due to time constraint, cost and objectives of the study, secondary data instead of primary data was used. The data collection was a sample of all the loan customers of the bank. Data was collected from all the bank's five branches, since all loans were accessed

through their main office. Due to difficulty in the collection of the data, the data was obtained for only 100 customers of the bank during the period 2006-2010.

1.7 Organization of the Thesis

The purpose of the research is to model factors influencing loan default among customers in Akatakyiman rural bank.

In Chapter One, which is the introduction, deals generally with the background of the study, statement of the problem, objective of the study, significance of the study, methodology, limitation, assumptions and organization of the thesis.

Chapter Two deals with the review of related literature on loan default. Chapter Three deals with the methodology used in the study. Chapter Four, focuses on the data analysis.

Finally, Chapter Five deals with the summary, discussions and conclusions of all the findings.

CHAPTER TWO

LITERATURE REVIEW

The financial institutions generally serves as financial intermediaries. It is their functions to mobilize funds savers by issuing to them their own securities. This form of asset transformation is required to ensure that funds are moved from surplus economic units to deficits economic units within the economy. These institutions, like any other business organization, have some risks to manage before they can successfully achieve their aim and objectives, which are almost always profit oriented. Non-performing Loans (NPLs) generally refer to loans which for a relatively long period of time do not generate income; that is the principal and/or interest on these loans has been left unpaid for at least 90 days; Caprio and Klingebiel, (1999). Non-performing Loans (NPLs) could also occur when the amortization schedules are not realized as at when due resulting in over-bloated loan interest due for payments.

NPLs reduce the liquidity of banks, credit expansion; it slows down the growth of the real sector with direct consequences on the performance of banks, the firm which is in default and the economy as a whole. According to the theory of finance, there are various risks facing financial institutions. They include: credit risk, liquidity risk, market risk, operating risk, reputation risk and legal risk. The system is highly sensitive while the activities of the operators need to be conducted within the laid down and agreed rules and procedures, in order to achieve a reasonable level of efficiency. (Ibid)

NPLs have become contemporary issues in credit management and undoubtedly the new frontier in finance the accumulation of NPLs is generally attributable to a number of factors, including economic down turns and macroeconomic volatility, terms of trade

deterioration, high interest rates, excessive reliance on overly high-priced inter-bank borrowings, insider lending and moral hazard Goldstoin and Turner, (1996).

deServigny and Renault, (2004) submitted that NPLs has taken a new dimension in finance just as interest rate and asset and liability management were 15 years ago. Because of mounting pressure of NPLs on bank's balance sheets and incessant bank failures, the Central Bank of Nigeria's Prudential Guidelines (1990) and subsequent reviews subsume credit facilities into loans, advances, overdrafts, commercial papers, banker's acceptances, bills discounted, leases, guarantees, and other loss contingencies connected with a bank's credit risks. The activities of these credits in terms of frequency of repayment or inability to repay same have further made it possible to group them into performing and non-performing credit facilities.

From the view of Elaine, (2007), NPLs or credit risk encapsulates the potential loss in the event of credit deterioration or default of a borrower. Thus a sound credit appraisal of loans is very important to the creditor. As argued by Dorfman, (1998), bankers required an understanding of credit standards, the process by which credit worthiness and credit structure are analyzed, decision-making techniques, negotiation, follow-up and problem resolution, in order to effectively manage credit risk. Abolo, (1999) supported Dorfman's assertion and presented his own principles of lending under three headings, that is, safety, suitability and profitability of credit, which equally compel bankers to follow the lending rules. Although credit depends on good faith, and no matter the amount of confidence that parties have on each other, it does not reduce the importance of scrutiny of these loan portfolios where good faith has been violated either

deliberately or inadvertently. Thus, the lenders must search for and avoid dishonest borrowers.

This involves sound credit analysis, which Nwankwo, (1991) describes as the process of assessing the risk of lending to a business or individual against the benefits to accruable from such investment. The benefits can be direct, such as interest earnings and possibly deposit balances required as a condition of the loan or indirect, such as initiation or maintenance of a relationship with the borrower, which may provide the bank with increased deposits and with demand for a variety of bank services. He argues further that credit risk assessment has two aspects. One is qualitative, and generally the more difficult; and the other is quantitative. To evaluate the qualitative risk, the loan officer has to gather and appraise information on the borrower's record of financial responsibility, determine his true or correct need for borrowing, identify the risks facing the borrower's business under current and prospective economic and political situations, and estimate the degree of his commitment regarding the payment. To estimate the financial viability of a portfolio, banks should not only limit their analysis to project evaluation techniques alone, but also by evaluating all credit risks that could become threats to the overall performance of such a portfolio.

Schall and Halley, (1980) outlined the key indicators for loan analysis as capacity, collateral, capital, condition and character. He concludes that lending involves the creation and management of risk assets and is an important task of bank management. While being the highest earning asset, the loan portfolio is also the most illiquid and most risky of banks' operation. The fiscal costs of these impaired loans are important as well, and vary with the scope and length of the crisis; Cortavarria Luis, et al. (2000). NPLs are the most common causes of bank failures. This has made all regulatory

institutions to prescribe minimum standards for credit risk management. The basis of sound credit risk management is the identification of the existing and potential risks inherent in lending activities. Measures to counteract these risks normally comprise clearly defined policies that express the bank's credit risk management philosophy and the parameters within which credit risk is to be controlled.

deServigny and Renault, (2004) opined that specific credit risk management measures typically include three kinds of policies. One set of policies include those aimed to limit or reduce credit risk, such as policies on concentration and large exposure, adequate diversification, lending to connected parties, or over-exposure. The second set includes policies of asset classification which expose a bank to credit risk. The third set include policies of loss provisioning or the making of allowances at a level adequate to absorb anticipated loss-not only on the loan portfolio, but also on all other assets that are sensitive to losses.

In a research submitted by Jorgensen, (2007) prepared for American Agricultural Economics Associations undergraduate research paper competition on the default of loans granted to farm credit customers, the researcher was interested in whether customers default because farm credit customers prefer lower interest rates or higher patronage payments. A farm credits service of East Central Oklahoma (FCSECO) is part of a nation-wide cooperative that supplies financing for full time and part time farmers. FCSECO not only makes loans to farmers but because it is a cooperative, its members/borrowers also benefit from what is known as the patronage payment. The patronage payment is a way of distributing farm credit's benefits to its members/

borrowers. Since FCSECO is customer focused and customer-driven it is essential that the FCSECO Board of Directors knows their customer base and what they desire as a customer. It would benefit FCSECO to determine the sustainability between patronage payment and fixed interest rates.

A conjoint survey was conducted on random FCSECO customers. After performing an OLS regression analysis, the results illustrated that the average FCSECO customer values a higher patronage payment more than a lower fixed interest rate on a given loan. This information is valuable to the FCSECO Board of Directors because it shows which attribute the average FCSECO customer has a preference towards. Since the average FCSECO customer greatly values the patronage payment, the FCSECO Board of Directors could use the patronage payment to its advantage in securing new loans. This study uses conjoint analysis to determine the trade-off between these two attributes; Hudson, (2007).

Within the survey that was used to determined the substitutability between fixed interest rate and patronage payment, FCSECO customers were able to rate their desirability concerning these two attributes. Regression analysis was then used to determine the relative importance of the two attributes, Mankiw, (2003) fixed interest rate and patronage payment. The regression analysis results will show the substitutability between the two attributes based on the preference of the average FCSECO customer. Furthermore, the result will show if the average FCSECO customer prefers a higher patronage payment or a lower fixed interest rates on real estate loans.

$$Y_{i,m} = \alpha + \beta_1 (INT)_{i,m} + \beta_2 (PAT)_{i,m} + e \; \forall_{i=1,2,\ldots 174;\, m=1,2\ldots,9}$$

The equation above is the regression model used to determine the substitutability between fixed interest rates and patronage payment. The Y represents the predicted

utility of the average FCSECO customer. The INT stands for the fixed interest rates variable and PAT stands for the patronage payment variable. The letter i represent the number of surveys used and the letter m represent the number of questions the customer were asked to rank their desirability. Find the predicted utilities using different variables of fixed interest rates and patronage payment will show which attributes is more desirable. Using the data from the desirability question will allow the FCSECO Board of Directors to see which attributes the average FCSECO customer desires more.

A research to Islamic financial institutions in 28 countries by Khan and Ahmed, (2001) find that credit risk is found highest in Musharakah (3.69 from a score of 5) followed by Mudarabah (3.25). Their findings highlights that the bankers perceive profit-and –loss sharing (PLS) modes to have higher credit risk. Mark-up risk is found highest in product- deferred contracts of Istina (3.57).

Sundararajan and Errico, (2002) opine that while PLS modes may shift the direct credit risk of Islamic banks to their investment depositors, they may also increase the overall degree of risk of the asset side of banks' balance sheet since the assets under this mode are uncollaterised. Their deductive intuition is that in principles, the ratio of riskier assets to total assets should typically be higher in an Islamic bank than in conventional bank.

Samad and Hasan, (1999) study on Malaysian Islamic banking reveals that Bank Islam performance of risk from 1984-1997 in risky business measured by debt-equity ratio (DER), debt to total Assets (DTAR) and Earning Multiplier (EM) increased over the

years. DER and EM are significantly related to profitability. In comparison with two conventional banks; Bank Pertanian and Perwira Affin Bank, Bank Islam risk indicators are lower. The reason for low risk of the Islamic bank is that its investment in government securities is much larger than the conventional banks.

In a study over 1984-1994 period, Makiyan, (2003) found that in the Iranian Islamic banking system, the supply of loan is significantly dependant on the changes in total deposits, the changes in the rate of inflation and the changes the time lags of the variables but it is not related to the changes in the expected rate of return on loans assigned to various economic sectors.

As for conventional banks, (Brewer, Jackson and Mondschean, 1996) found that loan sectors are associated with risk. Fixed-rate mortgage loans, investment in service corporations and real estate loans are found to be significant but negatively related to risk. Non-fixed rate mortgage loan is however, significant and positively related to risk. (Berger and DeYoung, 1997) find lagged risk-weighted asset (RWA) is significantly and positively related to credit risk measured by NPL to total loans. They rationalized that a relatively risky loan portfolio will result in higher NPLs. Lagged Capital measured by equity capital to total assets shows mixed results. For thinly capitalized banks, lagged Capital coefficient estimate is significantly but negatively related to risk. This finding supports the moral hazard hypothesis, and suggests that, on an average, thinly capitalized banks take more risky loans, which potentially could lead to higher NPLs

LLP (loan loss provision to average loans outstanding) has been identified in banking literature as a proxy for credit risk Rose, 1(996). Ahmed, (1998) found LLP to be positive and is significantly associated with NPL. Hence, a higher LLP indicates an increase in risk and deterioration in loan quality.

On the other hand, Lending is a risky enterprise because repayment of loans can seldom be fully guaranteed. Generally, in spite of the importance of loan in agricultural production, its acquisition and repayment are fraught with a number of problems especially in the small holder farming Awoke, (2004). It is reported in empirical studies that large rate of default has been a perennial problem in most agricultural credit schemes organized or supported by governments. Most of the defaults arose from poor management procedures, loan diversion and unwillingness to repay loans. For this reason, lenders devise various institutional mechanisms aimed at reducing the risk of loan default (pledging of collateral, third-party credit guarantee, use of credit rating and collection agencies, etc.). In the context of providing credit to the rural asset-poor, what is required is institutional innovation that combines prudent and sustainable banking principles with effective screening and monitoring strategies that are not based on physical collateral (such as land).

Koopahi and Bakhshi, (2002) used a discriminant analysis to identifying defaulter farmers from Non-defaulters of agricultural bank recipients in Iran. Results showed that use of machinery, length of repayment period, bank supervision on the use of loan had significant and positive effect on the agricultural credit repayment performance. On the other hand incidence of natural disasters, higher level of education of the loan recipient

and length of waiting time for loan reception had a significant and negative effect on dependent variable.

Deng et al., (1996) developed an empirical, option-based model of homeowner's default behavior, in a proportional hazard framework. These authors simulate probabilities of default and default costs on zero-down payment loans and then compare the results with conventional underwriting standards. They estimate that, if low-income borrowers are enticed by zero-down payment requirements and if no adjustment for the higher default rates is made, the cost of the implicit subsidy would amount from $74,000 to $87,000 per million dollars of lending.

Quercia et al., (1995) show that a lower loan-to-value (LTV) ratio at the time of origination (i.e., higher down payment) leads to lower default rates for rural, low-income borrowers. These authors focused on the 1981 Farmers Home Administration Section 502 program and show that, while contemporaneous equity value in rural low-income mortgage loans is not associated with default, crisis events are; Van Order et al., (2000) find, however, that the default behavior of both low- income and average-income groups is responsive to negative contemporaneous equity, while default rates and default losses are higher for low-income borrowers. Moreover, the influence on credit risk of individual and neighborhood income is small for LTV less than 80 percent, but it ranges from 15 up to 50 basis points for very high LTV ratios. Enticing low-income mortgage borrowers with lower down payment requirements thus increases the risk of default.

Oladeebo, (2008) examined socio-economic factors influencing loan repayment among small scale farmers in Ogbomoso agricultural zone of Oyo State of Nigeria. Results of multiple regression analysis showed that amount of loan obtained by farmers; years of farming experience with credit use and level of education were the major factors that positively and significantly influenced loan repayment. A main strategy of governments in developing countries like Iran is help to develop the rural areas and increase agricultural production through investment in the sector, so farmer's access to credit and direct to productive investment projects seems to be required. One of the Iranian financial institutes that play an important role in financing agriculture sector is Agricultural Bank. This bank is the main institutions of formal agricultural credit supply in Iran that can direct agricultural credit flow such that helps general economic policies of government. So duty of agricultural bank includes financing farmers and related industries and participation in activities which private sector can't invest in it. A main part of financial resources of Agricultural Bank comes through recovery of overdue granted credits while lending activity for banking system is accompanied with some risks and problems. Although in Khorasan-Razavi province of Iran, 64 percent of total credit demand of farmers in 2006 is covered by agricultural bank but it is not investigated how received credit has been repaid and which factors influencing on repayment behavior of farmers. Thus in this study, in order to adopt further proportional policies, the role of socio-economic factors in repayment behavior of farmers for last received loan from agricultural bank has been identified.

In-person exit counseling is strongly related to default behavior. Borrowers at Texas A&M who receive exit counseling through in-person contact with a counselor have a 1.3 percent default rate, while borrowers who do not receive in-person counseling have an

11.1 percent default rate. However, in-person exit counseling might owe much of its association with default to the fact that nearly everyone who graduates receives in-person exit counseling, but few borrowers who do not graduate receive it; Steiner and Teszler (2003).

According to Oni O.A et al., (2005) study on factors influencing loan default among poultry farmers in Ijebu Ode Local Government Area of Ogun State Nigeria; the result from the probit model revealed that flock size of the farmers significantly influence default in loan repayment at ($P < 0.10$) level. Age of the farmers significantly influence default in loan repayment at ($P < 0.01$) level, while Educational level and Income of the farmers also significantly influence default in loan repayment at ($P < 0.05$) level.

A study of University of Texas at Austin borrowers found that the highest degree attained accounted for 27 percent of the variation in default behavior in the study, the most of any variable in the study. The variable with the second greatest impact on defaults – number of credit hours failed – accounted for 21 percent of the variation in default behavior; Thein and Herr, (2001).

In the study conducted by Hooman Mansoori, (2009) on factors affecting loan repayment performance of farmers in Khorasan-Razavi Province of Iran, he found that using a logit model and a cross sectional data of 175 farmers of Khorasan-Razavi province in 2008, the results showed that loan interest rate is the most important factor affecting repayment of agricultural loans. Farming experience and total application costs are the next factors, respectively.

Researchers speculate that GPA may serve as a proxy for ability and motivation, traits associated with success in later life as well as in college; Volkwein and Szelest, (1995). College experience and success variables are those that occur in college and which the college, the borrower, or both have some ability to affect. These characteristics include college major, academic achievement, transfer status, educational goals of the student, financial support, and degree completion; Volkwein et al., (1998).

From Yegorova et al., (2000) research on a successful loan default prediction model for small business, a total of 117 variables representing loan characteristics were initially examined, and a series of practical screening methods were used to isolate the more statistically relevant variables for predicting loan default. Only the most statistically significant variables with an economically "correct" sign were then used to build a binary logistic regression model. Three ratios, the current liabilities/current assets, the sales/gross margin, and the equity/working capital were found to be highly significant in predicting loan default. The resulting model correctly predicted 87% of bad loans.

The reason for the correlation between college success and default behavior is unknown; however, it is possible that the hard work and responsibility that result in college success are established habits that carry over to other responsibilities in students' lives, such as loan repayment. Also, borrowers who achieve success in college will most likely obtain better positions in the job market and be in a better position to repay their loans after they leave school; Steiner and Teszler, (2003).

In a study of Texas A&M University students, borrowers who did not graduate had a nearly 14 percent default rate while borrowers who did graduate had less than a 2 percent default rate. The study further indicates that borrowers who obtain degrees have low default rates no matter what type of degree (Bachelor of Science, Bachelor of Arts, etc.) they get; Steiner and Teszler, (2003).

In a study of California borrowers, failure to complete the academic program was one of the strongest predictors of default among all types of students; Woo, (2002). Flint's study, which was national, found that among student academic characteristics, only GPA was related to repayment, such that higher GPAs are associated with avoidance of default; Flint, (1997).

A study of non-federally guaranteed loans extended to law school students in the early 1990s challenges the notion that there are institutional as well as borrower explanations for default. In this study, variables associated with borrower characteristics, such as ethnicity and family income, were entered first into the model followed by institutional variables. The study found that, after taking into account the characteristics a student brought with him or her to postsecondary study, very little predictiveness was added to the model by also taking into account the characteristics and practices of the school the borrower attended. That is to say, this study found default is primarily related to borrower willingness and ability to repay, not to anything the institution is doing; Monteverde, (2000).

Despite earlier studies to the contrary, there is little evidence that institutional characteristics have an impact on default. Rather, loan repayment and default behavior can mostly be predicted by the characteristics of individual borrowers, including choice of major, performance in college, and subsequent post college achievement and behavior. Staying in college, earning good grades, completing a degree, getting and staying married, and not having dependent children are all actions that lower the likelihood of default; Volkwein and Szelest, (1995).

CHAPTER THREE

METHODOLOGY

The main objective of the study is to fit a logistic regression model of some factors that influences loan default payment among customers in Akatakyiman Rural Bank limited. In this study, secondary data was collected from the credit department of the bank on customers who either defaulted or not in the loan facilities they had accessed (Yes/No) thus, Repayment Status. This variable will be used as the dependent variable in the analysis. The set of predictors (independent variables) includes sex, age, marital status, security used as collateral, town dummy educational level, Interest Rate and Type of Loan.

Due to the importance of the techniques to the analysis, this chapter is devoted to a brief review of logistic regression analysis. Further details of this technique can be seen in Applied Logistic Regression 2nd Edition by Hosmer and Lemeshow, (2000), Introduction to Categorical Data Analysis 2nd Edition by Alan Agresti, (2007) and Linear Statistical Inference 2nd Edition by Rao C.R , (1973).

3.1 SIMPLE LOGISTIC REGRESSION MODEL

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. Over the last decades the logistic regression model has become in many fields, the standard method of analysis in this situation. Before beginning a study of logistic regression it is important to understand that the goal of an analysis using this method is the same as that of any model-building technique used in statistics; to find the

best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. These independent variables are often called covariates. The most common example of modeling and one assumed to be familiar is the usual linear regression model where the outcome is assumed to be continuous.

What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. This difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression. Thus the technique used in linear regression analysis will motivate our approach to logistic regression.

The first difference concerns the nature of the relationship between the outcome and independent variables. In any regression problem the key quantity is the mean value of the outcome variable given the value of the independent variable. This quantity is called the conditional mean and will be expressed as "$E(y/x)$" where y denote the outcome variable and x denote a value of independent variable. The quantity $E(y/x)$ is read "the expected value of y given the value of x". In linear regression it can be assume that this mean may be expressed as an equation linear in x (or some transformation of x or y) such as

$$E(y/x) = \beta_\circ + \beta_1 x \qquad (3.1)$$

This expression implies that it is possible for $E(y/x)$ to take on any value as x ranges between$-\infty \; and \; +\infty$.

With dichotomous data, the conditional mean must be greater than or equal to zero and less than or equal to 1. (i.e. $0 \leq E(y/x) \leq 1$).

It means that this mean approaches zero and 1 "gradually". The change in the $E(y/x)$ per unit change in x becomes progressively smaller as the conditional mean gets closer to zero and 1. An S-shaped curve is produced by the function. It also resembles a plot of a cumulative distribution of a random variable. It should not seem surprising that some well-known cumulative distributions have used to provide a model for $E(y/x)$ in the case when y is dichotomous. The model that will be used is the logistic distribution.

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable. Cox and Snell, (1989) discussed some of these issues in literature. There are two primary reasons for choosing the logistic distribution. First, from a mathematical point of view, it is an extremely flexible and easily used function and second, it lends itself to a clinically meaningful interpretation.

In order to simplify notation, the quantity $\pi(x) = E(y/x)$ is used to represent the conditional mean of y given x when the logistic distribution is used. The specific form of the logistic regression model used is;

$$\pi(x) = \frac{e^{\beta_\circ + \beta_1 x}}{1 + e^{\beta_\circ + \beta_1 x}} \qquad (3.1)$$

A transformation of $\pi(x)$ that is central to our study of logistic regression is the logit transformation. This transformation is defined in terms of $\pi(x)$ as;

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_\circ + \beta_1 x \qquad 3.1a$$

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit $g(x)$, is linear in its parameters, may be continuous and many range from $-\infty \; and \; +\infty$ depending on the range of x.

In the linear regression model an assumption that an observation of the outcome variable may be expressed as $y = E(y/x) + \varepsilon$ is made. The quantity ε is called the error and expresses an observation's deviation from the conditional mean. The most common assumption is that ε follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that the conditional distribution of the outcome variable given x will be normal with mean $E(y/x)$, and variance that is constant. This is not the case with a dichotomous outcome variable. In this situation the value of the outcome variable given x may expressed as $y = \pi(x) + \varepsilon$. Here the quantity ε may assume one of the two possible values. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus, ε has a distribution with mean zero and variance equal to $\varepsilon = \pi(x)[1 - \pi(x)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean, $\pi(x)$.

3.1.1 FITTING THE SIMPLE LOGISTIC REGRESSION MODEL

Assuming a sample of n independent observations of the pair $(x_i, y_i), i = 1,2, \dots, n$, where y_i denotes the value of a dichotomous outcome variable and x_i is the value of the independent variable for the i^{th} subject. Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of a characteristic,

respectively. To fit the logistic regression model in equation (3.1) to a set of data requires the values of $\beta_\circ \; and \; \beta_1$, the unknown parameters are estimated.

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is called maximum likelihood. This method will provide the foundation for our approach to estimation with the logistic regression model. In a very general sense the method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method, firstly, a function, called the likelihood function must be constructed. This function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of these parameters are chosen to be those values that maximize this function. Thus the resulting estimators are those which agree most closely with the observed data. Now the description of how to find these values from the logistic regression model is necessary.

If y is coded as 0 or 1 when the expression for $\pi(x)$ given in equation (3.1) provides (for an arbitrary value of $\beta = (\beta_\circ, \beta_1)$, the vector of parameters) the conditional probability that y is equal to 1 given x. This will be denoted as $P(y = 1|x)$. It follows that the quantity $1 - \pi(x)$ gives the conditional probability that y is equal to zero given x, as $P(y = 0|x)$. Thus, for those pairs$(x_i y_i)$, where $y_i = 1$, the contribution to the likelihood function is $\pi(x)$, and those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x)$, where the quantity $\pi(x_i)$ denotes the value of $\pi(x)$ computed at x_i. A convenient way to express the contribution to the likelihood function for the pair$(x_i y_i)$ is through the expression

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad (3.2)$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation (3.2) as follows:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad (3.3)$$

The principle of maximum likelihood states that we use as our estimate of β the value which maximizes the expression in equation (3.3). However, it is easier mathematically to work with the log of equation (3.3). This expression, the log likelihood, is defined as

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \qquad (3.4)$$

To find the value of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to $\beta_\circ \, and \, \beta_1$ and set the resulting equations to zero. These equations, known as the likelihood equations, are:

$$\sum [y_i - \pi(x_i)] = 0 \qquad (3.5)$$

And

$$\sum x_i \, [y_i - \pi(x_i)] = 0 \qquad (3.6)$$

In equations (3.5) and (3.6), the summation is over the *i* varying from 1 to n.

For logistic regression the expressions in equations (3.5) and (3.6) are nonlinear in $\beta_\circ \, and \, \beta_1$, and thus require special method for their solution. These methods are iterative in nature and have been programmed into available logistic software. For the moment these iterative methods needs not to be concerned about and will be viewed as a computational detail taken care of for us.

The value of β given by the solution to equations (3.5) and (3.6) is called the maximum likelihood estimate and will be denoted as $\widehat{\beta}$. In general, the use of the symbol ^ denotes the maximum likelihood estimate of the respective quantity. For example, $\hat{\pi}(x_i)$ is the maximum likelihood estimate of $\pi(x_i)$. This quantity provides an estimate of the conditional probability that y is equal to 1, given that x is equal to x_i

3.1.2 TESTING FOR THE SIGNIFICANCE OF THE LOGISTIC REGRESSION COEFFICIENTS

After estimating the coefficients, our first look at the fitted model commonly concerns an assessment of the significance of the variables in the model. This usually involves formulation and testing of a statistical hypothesis to determine whether the independent variables in the model are significant. The method for performing this test is quite general and differs from one type of model to the next only in the specific details. We begin by discussing the general approach for a single independent variable. The multivariate case will be discussed in the subsequent sections.

One approach to testing for the significance of the coefficient of a variable in any model relates to the following question. Does the model that includes the variable in question tell us more about the outcome (or response) variable than a model that does not include that variable? This question is answered by comparing the observed values of the response variable to those predicted by each of the two models; first with and the second without the variable in question. The mathematical function used to compare the observed and predicted values depends on the particular problem. If the predicted values

with the variable in the model are better, or more accurate in some sense, than when the variable is not in the model, then we feel that the variable in question is "significant". It is important to note that we are not considering the question of whether the predicted values are an accurate representation of the observed values in an absolute sense (this would be called goodness-of-fit). Instead, our question is posed in a relative sense.

Compare observed values of the response variable to predicted values obtained from models with and without the variable in question. In logistic regression comparison of the observed to predicted values is based on the log likelihood function defined in equation (3.4). To better understand this comparison, it is helpful conceptually to think of an observed value of the response variable as being a predicted value resulting from a saturated model. A saturated model is one that contains as many parameters as there are data points.

The comparison of the observed to predicted values using the likelihood function is based on the following expression:

$$D = -2ln\left[\frac{(likelihood\ of\ the\ fitted\ model)}{(likelihood\ of\ the\ saturated\ model)}\right] \qquad (3.7)$$

The quantity inside the large bracket in the equation (3.7) above is called the likelihood ratio. Using the minus twice it log is necessary to obtain a quantity whose distribution is known and can therefore be used for hypothesis testing purposes. Such a test is called the likelihood ratio test. Using equations (3.4) and (3.7) becomes

$$\mathrm{D} = -2\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[\mathrm{y_i ln}\left(\frac{\hat{\pi}_\mathrm{i}}{\mathrm{y_i}}\right) + (1-\mathrm{y_i})\mathrm{ln}\left(\frac{1-\hat{\pi}_\mathrm{i}}{1-\mathrm{y_i}}\right)\right] \qquad (3.8)$$

Where $\hat{\pi}_i = \hat{\pi}(x_i)$.

The statistic D, in equation (3.8) is called the deviance by some authors and plays a central role in some approaches to assessing goodness-of-fit.

Furthermore, in a setting where the values of the outcome variables are either 0 or 1, the likelihood of the saturated model is 1. Specifically, it follows from the definition of a saturated model that $\hat{\pi}_i = y_i$ and the likelihood is

$$l(saturated\ model) = \prod_{i=1}^{n} {y_i}^{y_i} \times (1 - y_i)^{(1-y_i)} = 1$$

Thus it follows from equation (3.7) that the deviance is

$$D = -2\ln(likelihood\ for\ the\ fitted\ model) \qquad (3.9)$$

For purposes of assessing the significance of an independent variable the value of D with independent variable is compared with the value of D without the independent variable in the equation. The change in D due to the inclusion of the independent variable in the model is obtained as:

$$G = D(model\ without\ the\ variable) - D(model\ with\ the\ variable).$$

Because the likelihood of the saturated model is common to both values of D being differenced to compute G, it can be expressed as

$$\mathrm{G} = -2\ln\left[\frac{(\text{model without the variable})}{(\text{model with the variable})}\right] \qquad (3.10)$$

For specific case of a single independent variable, it is easy to show that when the variable is not in the model, the maximum likelihood estimate of $\beta_\circ = \ln(n_1/n_\circ)$ where $n_1 = \sum y_i$ and $n_\circ = \sum(1 - y_i)$ and the predicted value is constant, n_1/n. In this case, the value of G is:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} {y_i}^{y_i} \times (1 - y_i)^{(1-y_i)}}\right] \qquad (3.11)$$

Or

$$G = 2\left\{\sum_{1}^{n}[y_i \ln(\hat{\pi}_i) + (1 - y_i)\ln(1 - \hat{\pi}_i) - n_1 \ln(n_1) + n_0 \ln(n_0) - nln(n)]\right\} \qquad (3.12)$$

Under the hypothesis that β_1 is equal to zero, the statistic G follows a chi-square distribution with 1 degree of freedom. Additional mathematical assumptions are also needed; however, for the above case they are rather nonrestrictive and involve having a sufficiently large sample size, n. The symbol $x^2(v)$ is use to denotes a chi-square random variable with v degrees of freedom.

The calculation of the log likelihood and the likelihood ratio test are standard features of all logistic regression software. This makes it easy to check for the significance of the addition of new terms to the model. In the simple case of a single independent variable, a model containing only the constant term is first fit, then a model containing the independent variable along with the constant is next fit. This gives rise to new log likelihood. The likelihood ratio test is obtained by multiplying the difference between these two values by -2.

Two other similar, statistically equivalent tests have been suggested in literature. These are the Wald test and the Score test. The assumptions needed for these tests are the same as those of the likelihood ratio test in equation (3.11)

The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter,$\hat{\beta}_1$, to an estimate of its standard error. The resulting ratio, under the

hypothesis that $\beta_1 = 0$, will follow a standard normal distribution. While we have not yet formally discussed how the estimates of the standard errors of the estimated parameters are obtained, they are routinely printed out by computer software. Thus Wald test is computed as;

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \qquad (3.13)$$

The likelihood ratio statistic and its corresponding squared Wald statistic give approximately the same value in very large samples; so if one's study is large enough, it will not matter which statistic is used.

Nevertheless, in small to moderate samples, the two statistics may give very different results. Statisticians have shown that the likelihood ratio statistic is better than the Wald statistic in such situations. So, when in doubt, it is recommended that the likelihood ratio statistic be used. However, the Wald statistic is somewhat convenient to use because only one model, the full model, needs to be fit.

A test for the significance of a variable which does not require these computations is the Score test. Proponents of the Score test cite this reduced computational effort as its major advantage. Use of the test is limited by the fact that it cannot be obtained from some software packages. The Score test is based on the distribution theory of the derivatives of the log likelihood. In general, this is a multivariate test requiring matrix calculations which will be discussed in subsequent sections.

In the univariate case, this test is based on the conditional distribution of the derivative in equation (3.6), given the derivative in equation (3.5). In this case, we can write down an expression for the score test. The test uses the value of equation (3.6), computed

using $\beta_\circ = ln(n_1/n_\circ)$ and $\beta_1 = 0$. As noted earlier, under the parameter values, $\hat{\pi} = n_1/n = \bar{y}$. Thus, the left-hand side of equation (3.6) becomes $\sum x_i\,(y_i - \bar{y})$. It may be shown that the estimated variance is $\bar{y}(1-\bar{y})\sum(x_i - \bar{x})^2$. The test statistic for the Score test (ST) is

$$ST = \frac{\sum_{i=1}^{n} x_i\,(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

3.1.3 CONFIDENCE INTERVAL ESTIMATION

An important adjunct to testing for significance of the model, discussed earlier is calculation and interpretation of confidence intervals for parameters of interest. As is the case in linear regression we can obtain these for the slope, intercept and the "line" (i.e. the logit). In some settings it may be of interest to provide interval estimates for the fitted values (i.e. the predicted probabilities)

The basis for construction of the interval estimators is the same statistical theory we used to formulate the tests for significance of the model. In particular, the confidence interval estimators for the slope and intercept are based on their respective Wald tests. The endpoints of a

$100(1-\alpha)\%$ Confidence interval for the slope coefficient is

$$\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1) \qquad (3.14)$$

and for the intercept they are

$$\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_0) \qquad (3.15)$$

Where $z_{1-\frac{\alpha}{2}}$ is the upper $100(1-\alpha)\%$ point from the standard normal distribution and $\widehat{SE}(.)$ denotes a model-based estimator of the standard error of the respective parameter estimator.

The logit is the linear part of the logistic regression model and as such is most like the fitted line in a linear regression model. The estimator of the logit is

$$\hat{g}(x) = \hat{\beta}_\circ + \hat{\beta}_1 x \qquad (3.16)$$

The estimator of the variance of the estimator of the logit requires obtaining the variance of a sum. In this case it is

$$\hat{var}[\hat{g}(x)] = \hat{var}(\hat{\beta}_\circ) + x^2 \hat{var}(\hat{\beta}_1) + 2x\, \hat{cov}(\hat{\beta}_\circ \hat{\beta}_1) \qquad (3.17)$$

In general the variance of a sum is equal to the sum of the variance of each term and twice the covariance of each possible pair of terms formed from the components of sums. The endpoints of a $100(1-\alpha)\%$ Wald-based confidence interval for the logit are

$$\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} \widehat{SE}\big(\hat{g}(x)\big) \qquad (3.18)$$

Where $\widehat{SE}\big(\hat{g}(x)\big)$ is the positive square root of the variance estimator in (3.17)

3.2 THE MULTIPLE LOGISTIC REGRESSION MODEL

Consider a collection of p independent variables denoted by the vector $x' = (x_1, x_2 \ldots x_p)$. For the moment we will assume that each of these variables is at least scale. Let the conditional probability that the outcome is present be denoted by $P(Y = 1|x) = \pi(x)$. The logit of the multiple logistic regression model is given by the equation

$$g(x) = \beta_\circ + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (3.19)$$

in which case the logistic regression model is:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \qquad (3.20)$$

If some of the independent variables are discrete, nominal scale variables such as race, sex, marital status and so forth, it is inappropriate to include them in the model as if they were interval scale variables. The number used to represent the various levels of these nominal scale variables are merely identifiers and have no numeric significance. In this situation the method of choice is to use a collection of design variables (dummy variable). For instance if one of the independent variables is Educational level which has been coded as "High", "Low" and "Other". In this case, two design variables are necessary. One possible coding strategy is that when the respondent is "High" the two design variables D_1 and D_2, would both be set equal to zero, when the respondent is "Low", D_1 would be set equal to 1 while D_2 would still equal to 0; when the level of education of the respondent is "Other", we would use D_1= 0 and D_2=1.

Most logistic regression software will generate design variables and some programmes have a choice of several different methods.

In general, if a nominal scaled variable has *k* possible values, then *k-1* design variables will be needed. This is true since, unless stated otherwise, all of our models have a constant term. To illustrate the notation used for design variables, suppose that the j^{th} independent variable x_j has k_j levels. The k_j*-1* design variables will be denoted as D_{jl} and the coefficients for these design variables will be denoted as β_{jl}, $l = 1,2, \dots, kj - 1$. Thus, the logit for a model with p variables and the j^{th} variable being discrete would be

$$g(x) = \beta_\circ + \beta_1 x_1 + \sum_{l=1}^{kj-1} \beta_{jl} D_{jl} + \beta_p x_p$$

3.2.1 FITTING THE MULTIPLE LOGISTIC REGRESSION MODEL

Assume that we have a sample of n independent observations $(x_i, y_i), i = 1,2, \dots, n$. As in the univariate case, fitting the model requires that we obtain estimates of the vector $\boldsymbol{\beta}' = (\boldsymbol{\beta}_\circ, \boldsymbol{\beta}_{\mathbf{1},\dots,}\boldsymbol{\beta}_{\boldsymbol{p}})$. The method of estimation used in the univariate situation will be employed in the multivariate case – maximum likelihood. The likelihood function is nearly identical to that given in equation (3.3) with the only change being that $\pi(x)$ is defined as in equation (3.20). There will be $p + 1$ likelihood equations that are obtained by differentiating the log likelihood function with respect to the $p + 1$ coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0 \; for \; j = 1,2, \dots, p$$

As in the univariate model, the solution of the likelihood equations requires special software that is available in most, if not all, statistical packages. Let $\widehat{\boldsymbol{\beta}}$ denote the solution to these equations. Thus, the fitted values for the multiple logistic regression model are $\hat{\pi}(x_i)$, the value of the expression in equation (3.20) computed using $\widehat{\boldsymbol{\beta}}$ and x_i.

In the previous section only a brief mention was made of the method for estimating the standard errors of the estimated coefficients. Now that the logistic regression model has been generalized both in concept and notation to the multivariate case, we consider estimation of the standard errors in more detail.

The method of estimating the variance and covariance of the estimated coefficients follows from well-defined theory of maximum likelihood estimation (Rao (1973). This theory states that the estimators are obtained from the matrix of second partial derivatives of the log likelihood function. These partial derivative have the following general form

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \qquad (3.21)$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \qquad (3.22)$$

$for\ j, l = 0,1,2, \dots, p$ where π_i denotes $\pi(x_i)$. Let the $(p+1) \times (p+1)$ matrix containing the negative of the terms given in equations (3.21) and (3.22) be denoted as$\mathbf{I(\beta)}$. This matrix is called the observed information matrix. The variances and covariances of the estimated coefficients are obtained from the inverse of this matrix which we denote as $Var(\beta) = \mathbf{I^{-1}(\beta)}$. Except in very special cases it is not possible to write down an explicit expression for the elements in this matrix. Hence, we will use the notation $Var(\beta_j)$ to denote the j^{th} diagonal element of this matrix, which is the variance of $\hat{\beta}_j$, and covariance $Cov(\beta_j, \beta_l)$ to denote an arbitrary off-diagonal element, which is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_l$. The estimators of the variances and covariances, which will be denoted by $\hat{Var}(\hat{\beta})$ are obtained by evaluating $\boldsymbol{Var(\beta)}\ at\ \boldsymbol{\hat{\beta}}$. We will use $\boldsymbol{\hat{Var}(\hat{\beta}_j) and\ Cov(\hat{\beta}_j, \hat{\beta}_l)}$, $j, l = 0,1,2, \dots, p$ to denote the values in this matrix.

For the most part, we will have occasion to use only the estimated standard errors of the standard coefficients, which we will denote as

$$\hat{SE}(\hat{\beta}_j) = \sqrt{[\widehat{\boldsymbol{Var}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}})]} \qquad (3.23)$$

for j = 0,1,2, … , p. We will use this notation in developing methods for coefficient testing and confidence interval estimation.

A formulation of the information matrix which will be useful when discussing model fitting and assessment of fit is $\hat{I}(\hat{\beta}) = X'VX$ where X is an n *by* $p + 1$ matrix containing the data for each subject, and V is an n by n diagonal matrix with general element $\hat{\pi}_i(1 - \hat{\pi}_i)$. That is, the matrix X is

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

and the matrix V is

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

3.2.2 TESTING FOR THE SIGNIFICANCE OF THE MODEL

Once we have fit a particular multiple (multivariable) logistic regression model, we begin the process of model assessment. As in the univariate case presented in section 3.1, the first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the p coefficients for the independent variables in the model is performed in exactly the same manner as in the univariate case. The test is based on the statistic G given in equation (3.10). The only difference is that the fitted values $\hat{\pi}$, under the model are based on the vector containing

$p+1$ parameters $\widehat{\boldsymbol{\beta}}$. Under the null hypothesis that the p "slope" coefficients for the covariates in the model are equal to zero, the distribution of G will be chi-square with p degrees of freedom.

Before concluding that any or all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics,

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}$$

The multivariable analog of the Wald test is obtained from the following vector-matrix calculation:

$$\mathrm{W} = \hat{\beta}'\left[\hat{\mathrm{V}}\mathrm{ar}(\hat{\beta})\right]^{-1}\hat{\beta}$$

$$= \hat{\beta}'(\mathrm{X}'\mathrm{V}\mathrm{X})\hat{\beta},$$

which will be distributed as chi-square with $p+1$ degrees of freedom under the hypothesis that each of the $p+1$ coefficients is equal to zero.

Then multivariable analog of the Score test for the significance of the model is based on the distribution of the p derivatives of $L(\beta)$ with respect to β. The computation of this test is of the same order of complication as the Wald test.

3.2.3 CONFIDENCE INTERVAL ESTIMATION

We discussed confidence interval estimators for the coefficients, logit and logistic probabilities for the simple logistic regression model in subsection 3.1.3. The method used for confidence interval estimators for a multiple variable model is essentially the same.

The confidence interval estimator for the logit is a bit more complicated for the multiple variable model than the result presented in equation (3.18). The basic idea is the same, only there are now more terms involved in the summation. It follows from (3.19) that a general expression for the estimator of the logit for a model containing p covariates is

$$\hat{g}(x) = \hat{\beta}_\circ + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \qquad (3.24)$$

An alternative way to express the estimator of the logit in (3.24) is through the use of vector notation as $\hat{g}(x) = x'\hat{\beta}$, where the vector $\beta' = (\hat{\beta}_\circ, \hat{\beta}_{1,\ldots,}\hat{\beta}_p)$ denotes the estimator of the $p+1$ coefficients and the vector $x' = x_0, x_1, x_{2,\ldots,}x_p$ represents the constant and a set of values of the p – covariates in the model, where $x_0 = 1$. It follows from (3.17) that an expression of the variance of the estimator of the logit in (3.24) is

$$\hat{var}[\hat{g}(x)] = \sum_{j=0}^{p} x_j^2 \, \hat{var}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \, \hat{cov}(\hat{\beta}_j \hat{\beta}_k) \qquad (3.25)$$

We can express this result much more concisely by using the matrix expression for the estimator of the variance of the estimator of the coefficients. From the expression for the observed information matrix, we have that

$$\hat{var}(\hat{\beta}) = (\mathrm{X}'\mathrm{VX})^{-1} \qquad (3.26)$$

It follows from (3.26) that an equivalent expression for the estimator in (3.25) is

$$\hat{var}[(\hat{g}(x))] = x' \hat{var}(\hat{\beta}) x \qquad (3.27)$$

3.3 INTERPRETATION OF THE FITTED LOGISTIC REGRESSION MODEL

After fitting a model the emphasis shifts from the computation and assessment of significance of the estimated coefficients to the interpretation of their values. Strictly

speaking, an assessment of the adequacy of the fitted model should precede any attempt at interpreting it.

The interpretation of any fitted model requires that we be able to draw practical inferences from the estimated coefficients in the model. The question being addressed is: What do the estimated coefficients in the model tell us about the research questions that motivated the study?

For most models this involves the estimated coefficients for the independent variable in the model. On occasion, the intercept coefficient is of interest but this is the exception, not the rule. The estimated coefficients for the independent variable represent the slope (i.e. rate of change) of a function of the dependent variable per unit of change in the independent variable. Thus interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable and appropriately defining the unit of change for the independent variable.

The first step is to determine what function of the dependent variable yields a linear function of the independent variables. This is called the link function.

In the logistic regression model the link function is the logit transformation $g(x) = ln\{\pi(x)/[1-\pi(x)]\} = \beta_0 + \beta_1 x$.

In the logistic regression model, the slope coefficient represents the change in the logit corresponding to a change of one unit in the independent variable (i.e. $\beta_1 = g(x+1) - g(x)$. Proper interpretation of the coefficient in a logistic regression model depends on being able to place meaning on the difference between two logit.

3.3.1 DICHOTOMOUS INDEPENDENT VARIABLE

We begin our consideration of the interpretation of logistic regression coefficients with the situation where the independent variable is nominal scale and dichotomous (i.e. measured at two levels). This case provides the conceptual foundation for all the other situations.

We assume that the independent variable, x, is coded as either zero or one. The difference in the logit for a subject with x = 1 and x = 0 is $g(1) - g(0) = \beta_0 + \beta_1 - \beta_0 = \beta_1$.

The algebra shown in this equation is rather straightforward. We present it in this level of detail to emphasize that the first step in interpreting the effect of a covariate in a model is to express the desired logit difference in terms of the model. In this case the logit difference is equal to β_1. In order to interpret this result we need to introduce and discuss measure of association termed the odds ratio.

The possible values of the logistic probabilities may be conveniently displayed in a 2×2 as shown in Table 3.1.

Table 3.1 Values of the Logistic Regression Model When the Independent Variable Is Dichotomous

	Independent Variable (X)	
Outcome Variable (Y)	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$
Total	1.0	1.0

The odds of the outcome being present among individuals with $x = 1$ is defined as $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with $x = 0$ is defined as $\pi(0)/[1 - \pi(0)]$. The odds ratio, denoted OR, is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by the equation

$$OR = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \qquad (3.28)$$

Substituting the expression for the logistic regression model shown in Table 3.2 into (3.28) we obtain:

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{(\beta_0+\beta_1)-\beta_0}$$

$$= e^{(\beta_1)}$$

Hence, for logistic regression with dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and the regression coefficient is

$$OR = e^{\beta_1} \qquad (3.29)$$

Nevertheless, if the coding scheme is different from the (0, 1) then the odds ratio formula needs to be modified, but for the purpose of this study all the dichotomous variables will be coded using the (0, 1) coding scheme.

The simple relationship between the coefficient and the odds ratio is the fundamental reason why logistic regression has proven to be such a powerful analytic research tool. The odds ratio is a measure of association which has found a wide use, especially in

epidemiology, as it approximates how much more likely (or unlikely) it is for the outcome to be present among those with $x = 1$ than among those with $x = 0$.

The interpretation given for the odds ratio is based on the fact that in many instances it approximates a quantity called the relative risk. This parameter is equal to the ratio $\frac{\pi(1)}{\pi(0)}$.

It follows from (3.28) that the odds ratio approximates the relative risk if $[1 - \pi(0)]/[1 - \pi(1)] \approx 1$. This holds when $\pi(x)$ is small for both $x = 1$ and 0.

A $100(1 - \alpha)\%$ confidence interval (CI) estimate for the odds ratio is obtained by first calculating the endpoint of a confidence interval for the coefficient, β_1, and then exponentiating these values. In general, the endpoints are given by the expression

$$exp\left[\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \times S\hat{E}(\hat{\beta}_1)\right]$$

Because of the importance of the odd ratio as a measure of association, many software packages automatically provide point and confidence interval estimates based on the exponentiation of each coefficient in a fitted logistic regression model. These quantities provide estimates of odds ratios of interest in only few special cases (e.g. a dichotomous variable coded zero or one that is not involved in any interactions with other variables).

3.3.2 CONTINUOUS INDEPENDENT VARIABLE

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficient depends on how it is entered into the model and the particular units of the variable. For purposes of developing the method to interpret the coefficient for continuous variable, we assume that the logit is linear in the variable.

Under the assumption that the logit is linear in the continuous covariate, x, the equation for the logit is $g(x) = \beta_0 + \beta_1 x$. It follows that the slope coefficient, β_1 gives the change in the log odds for an increase of "1" unit in x, that is $\beta_1 = g(x+1) - g(x)$ for any value of x. Most often the value of "1" is not clinically interesting. Hence, to provide a useful interpretation for a continuous scale covariate we need to develop a method for point and interval estimation for an arbitrary change of "c" uints in the covariate. The log odds ratio for a change of c units in x is obtained from the logit difference $g(x+c) - g(x) = c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference, $OR_{(c)} = OR(x+c, x) = \exp(c\beta_1)$. An estimate may be obtained by replacing β_1 with its maximum likelihood estimate $\hat{\beta}_1$. An estimate of the standard error needed for confidence interval estimation is obtained by multiplying the estimated standard error of $\hat{\beta}_1$ by c. Hence the endpoints of the $100(1-\alpha)\%$ confidence interval (CI) estimate of $OR_{(c)}$ are

$$exp\left[c\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \times c\hat{SE}(\hat{\beta}_1)\right]$$

Since both the point estimate and endpoints of the confidence interval depends on the choice of c, the particular value of c should be clearly specified in all tables and calculations

In summary, the interpretation of the estimated coefficient for a continuous variable is similar to that of nominal scale variables: an estimated log odds ratio. The primary difference is that a meaningful change must be defined for the continuous variable.

3.4 PEARSON CHI-SQUARE GOODNESS OF FIT TEST OF INDEPENDENCE

The Chi-square test is a statistical test that can be used to test the hypothesis of no association between two variables. The Chi-square test statistic is given by

$$\chi^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(o_{ij}-e_{ij})^2}{e_{ij}} \qquad (3.30)$$

where $e_{(i,j)}$ is the expected cell frequency for the $(ij)^{th}$ cell. It can be shown that,

$$e_{(i,j)} = \frac{(R_i \times C_j)}{n} \qquad (3.31)$$

With the $(r-1)(c-1)$ degrees of freedom.

It is also assumed that when the observations in a cell is less than 5, the Chi-square test might lose it strength.

3.5 COCHRAN ARMITAGE TREND TEST

The Cochran-Armitage test for trend, named for William Cochran and Peter Armitage, is used in categorical data analysis when the aim is to assess for the presence of an association between a variable with two categories and a variable with k categories. It modifies the chi-square test to incorporate a suspected ordering in the effects of the k categories of the second variable.

The trend test is applied when the data take the form of a $2 \times k$ contingency table.

The trend test statistic is given by;

$$T \equiv \sum_{i=1}^{k} t_i(N_{1i}R_2 - N_{2i}R_1) \qquad (3.32)$$

where the t_i are weights, and the difference $N_{1i}R_2 - N_{2i}R_1$ be seen as the difference between N_{1i} and N_{2i} after reweighting the rows to have the same total.

The hypothesis of no association (the null hypothesis) can be expressed as:

$$\Pr(A = 1|B = 1) = \cdots \Pr(A = 1|B = k)$$

Assuming this holds, then, using iterated expectation,

$$E(T) = E(E(T|R_1, R_2) = E(0) = 0$$

The variance can be computed by decomposition, yielding

$$Var\ (T) = \frac{R_1 R_2}{N}\left(\sum_{i=1}^{k} t_i^2 C_i(N - C_i\right) - 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} t_i t_j\ C_i C_j \qquad (3.33)$$

and as a large sample approximation,

$$\cdot\frac{T}{\sqrt{Var\ (T)}} \sim N(0,1)$$

The weights t_i can be chosen such that the trend test becomes locally most powerful for detecting particular types of associations.

Interpretation and role

The trend test will have higher power than the chi-square test when the suspected trend is correct, but the ability to detect unsuspected trends is sacrificed. The trend test exploits the suspected effect direction to increase power, but this does not affect the sampling distribution of the test statistic under the null hypothesis. Thus, the suspected trend in effects is not an assumption that must hold in order for the test results to be meaningful.

3.6 SUMMARY

In summary, we described how to fit simple and multiple logistic regression models, described the three test procedures, the likelihood ratio test, the Wald test and the Score test. We have also described how to compute the odds ratio for an arbitrarily coded single exposure variable that may be dichotomous.

We have also shown how to obtain interval estimates for odds ratios obtained from a logistic regression. In particular, we have described confidence interval formula. We have also described how to interpret the fitted logistic regression model for both dichotomous independent variable and continuous independent variable

Finally, we have described the Pearson Chi-square goodness of fit test of independence and the Cochran Armitage trend Test.

CHAPTER FOUR

ANALYSIS

This chapter seeks to analyze the factors that influence customer's ability to repay loans from data in appendix 2. This will be done by exploring relationship between some variables and also by using binary logistic regression analysis with the help of SPSS.

An important objective of the study is to model repayment status. That is to regress repayment status on some predictor variables. Repayment Status is a categorical variable, therefore the ordinary regression approach is not appropriate. For this reason we resort to binary regression technique. This technique, as discussed earlier in chapter three (3) is appropriate when the response variable is categorical.

Repayment Status is categorical because it comprises Yes/No, thus whether respondents defaulted or not the loan accessed. As a result the response variable is suitable for not just any logistic regression but a binary logistic regression.

Using SPSS, the result of the binary regression of repayment status on type of security, age, marital status, city dummy and education level are displayed in appendix. An extract of this is shown.

Exploring the Relationship between Amount Approved and the Type of Loan.

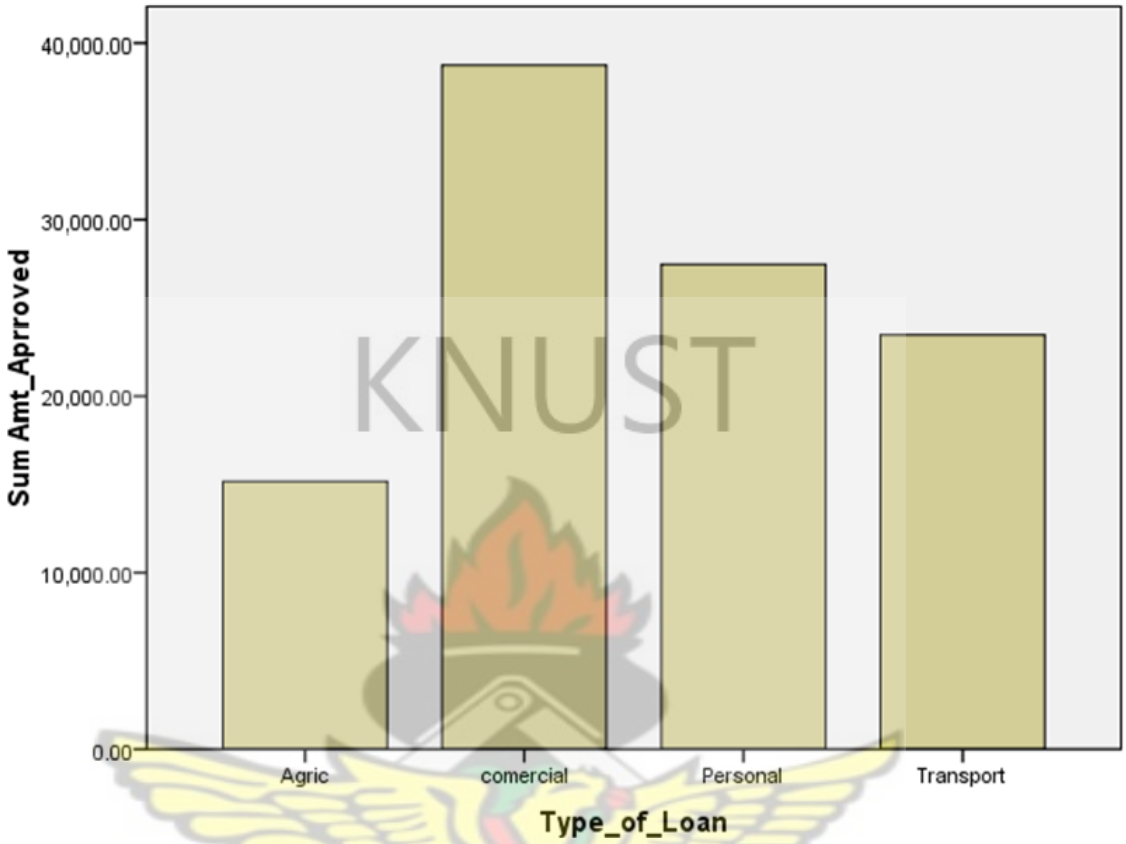

Fig.4.1: A Bar Graph Showing the Sum of Amount Approved Against Type of Loan

From Figure 4.1 above, it shows the distribution of Amount Approved to customers by the Type of Loan Accessed. It can be seen that among the types of loans, commercial loan was given the highest total amount approved to customers which is GH¢ 38810.00, followed by personal loan which also recorded GH¢ 27,500.00. Transport loan was the third highest with a total amount approved as GH¢ 23,520.00. Finally, Agric loan recorded the least amount approved which was GH¢ 15, 170.

Table 4.1: Frequency distribution of Variables

Variable		Frequency
Loan Type	Agric	16
	Commercial	41
	Personal	25
	Transport	18
Interest	24%	36
Rate	30%	39
	35%	17
	38%	8
Security	MO	34
	PG	66
Marital	M	92
Status	N	8
Educ.	High	24
Level	Low	76
Town-	K	58
Dummy	N	42
Sex	F	42
	M	58

From table 4.1: we observed 16% of the loan customers applied for Agric Loan, 41% applied for Commercial Loan, 25% applied for Personal Loan and 18% applied for Transport Loan. About 36% of the customers were given Loan with an interest rate of 24%, 39% were given loan with an interest rate of 30%, 17% were given loan with an interest rate of 35% and only 8% were given loan with an interest rate of 38%. About 34% customers used personal guarantee as security for their loans whiles 66% of the customers used mortgages as collaterals for their loans. About 58% of the customers

were regarded to be descendents of Komenda where the bank's main office is located and 42% of the customers were from the other four branches of the bank which is outside Komenda. About 58% of the customers were males while 42% were females. Finally, it is seen, about 92% of the customers were married whiles the remaining 8% were unmarried

From appendix I, it is seen that, among the customers who accessed loan during the 2006-2010 fiscal year, the youngest customer was 21 years whiles the oldest customer was 60 years. And on the average customers who were about 42 years accessed most of the loan. Taking into consideration the number of years the customer has been operating with the bank, it was realize that the average years customers have been operating with the bank is 10 years with the minimum years being 2 years whiles the maximum years customers has operated with bank is 20 years.

Chi-square Test for Independence

Table 4.2: Type of Loan versus Repayment Status

		Repayment Status		
		No	Yes	Total
Loan Type	Agric	2	14	16
	Commercial	20	21	41
	Personal	19	6	25
	Transport	7	11	18
Total		48	52	100

From Table 4.1, in the test of independence between the Loan Type and Repayment Status, the Pearson Chi-square test of Independence value obtained is 16.450 with

degree of freedom 3. P-value = 0.0001 which is less than 0.05. This means that the ability of a customer to default or otherwise of a loan depends on the Loan Type applied for.

COCHRAN-ARMITAGE TREND TEST

[Sum of scores from population <no >]

Min	Max	Mean	Std-dev	Observed	Standardized
12.00	81.00	46.56	4.627	***50.00***	0.7435

Exact Inference:

One-sided p-value: Pr { Test Statistic .GE. Observed} = **0.2632**

Pr { Test Statistic .EQ. Observed} = 0.0653

Two-sided p-value: Pr {| Test Statistic - Mean |

.GE. | Observed - Mean | = 0.5183

Two-sided p-value: 2*One-Sided = 0.5264

Objective: To determine if increase in interest rate places an increasing risk of loan default repayment. Using an exact trend test.

From the Cochran Armitage Trend Test output, the P-value = 0.2632 which is greater than 0.05. This means that an increase in interest rate given on a customer's loan does not increase the risk of loan default payment.

Table 4.3: Security versus Repayment Status

		Repayment Status		
		No	Yes	Total
Security	MO	22	12	34
	PG	26	40	66
Total		48	52	100

From Table 4.3, in the test of independence between the Security and Repayment Status, the Pearson Chi-square test of Independence value obtained is 5.760 with degree of freedom 1. The P-value = 0.0016 which is less than 0.05. This means that the ability of a customer to default or otherwise of a loan depends on the type of security offered as collateral .

From appendix I, it is seen that the Chi-square test of Independence between Town-Dummy and Repayment Status, Marital Status and Repayment Status, Sex and Repayment Status and Educational Level and Repayment were not significant. This means that the ability of a customer to default or otherwise of a loan applied for does not depend on the Town-dummy, Marital Status, Sex and Educational Level of the customer.

TABLE 4.4: OMNIBUS TEST OF MODEL COEFFICIENTS

	Chi-square	df	P-value
Step	31.345	14	0.001
Block	31.345	14	0.001
Model	31.345	14	0.001

Considering the table 4.4 above, where the model (set of predictor variables) is tested. The Omnibus Test of Model Coefficients gives us an overall indication of how well the model performs, over and above the results obtained for Block 0, with none of the predictors entered into the model. This is referred to as the 'goodness of fit' test. For this set of result we want a highly significant value (the Sig. value should be less than 0.05).

For the Hosmer and Lemeshow Test for goodness of fit table in appendix I, a Chi-square value 6.800 a P- value of 0.558 was reported. This test indicates that the model is good.

TABLE 4.5: CLASSIFICATION TABLE

		Predicted		
		Repayment Status		percentage
		No	Yes	Correct
Repayment Status	No	37	14	77.1
	Yes	18	34	65.0
Overall Percentage				71.0

From the result in table 4.5, indicates how well the model is able to predict the correct category (default/no default) for each case. Thus the model correctly classified 71.0 per cent of cases overall. Here, the model correctly classified 65.0 per cent of the customers who did default in loan repayment. The specificity of the model is the percentage of the group without the characteristics of interest (no default in loan repayment) that is correctly identified. Here the specificity is 77.1 per cent (customers with no default in loan repayment correctly predicted not to have defaulted by the model).

Table 4.6: PARAMETER ESTIMATES

	Estimates	Std. error	Wald	df	P-value	e^{β}
Intercept	2.975	0 .981	2.254	1	0.133	19.589
Age	-0.027	0.039	0.483	1	0.487	0.973
Sex(1)	-0.097	0.600	0.026	1	0.872	0.908
Marital_status	-0.514	0.884	0.338	1	0.561	0.598
Educ_level	-0.360	0.690	0.272	1	0.602	0.698
Security	-2.927	1.104	7.033	1	0.008	0.054
Years	0.000	0.073	0.000	1	0.998	1.000
Town	-0.729	0.615	1.404	1	0.236	0.482
Loan Type(1)	1.554	0.987	2.476	1	0.116	4.729
Loan Type(2)	1.075	1.008	1.139	1	0.286	2.931
Loan Type(3)	-1.451	0.746	3.784	1	0.052	0.234

From table 4.6, the fitted logistic regression equation is

$$\hat{\pi}(Ps = 1/x) = \frac{e^{(2.975-2.927S+1.554\ LT\ (1)+1.075\ LT(2)-1.451LT(3))}}{1+e^{(2.975-2.927S+1.554\ LT\ (1)+1.075\ LT(2)-1.451LT(3))}}$$

And the logit model;

$$logit\ (Ps = 1/x) = 2.975 - 2.927S + 1.554\ LT\ (1) + 1.075\ LT(2) - 1.451LT(3)$$

Where

Ps - Repayment Status S - Security

LT - Loan Type

Table 4.6 gives us information about the contribution or importance of each of predictor variables. The test that is used here is known as the Wald test, and the test statistic for each predictor variable is seen in the column labeled Wald. These are variables that contribute significantly to the predictive ability of the model. It is seen that two of the variables were significant the Type of security used as collateral for the loan and the Type of Loan assessed. The factors influencing whether a customer defaulted a loan repayment are Security and the Type of Loan assessed. Marital Status and Sex, Age, Educational Level, Town, Years the customer has been operating with the bank did not contribute significantly to the model.

Another useful piece of information in the parameter estimates table is provided in the e^{β} column. These values are the odds ratios (OR) for each of the independent variables. Taking Transport loan as the reference group, the odds of a customer defaulting in a loan repayment is 0.234 times higher for a customer who was given Personal loan than for a customer who was given a Transport Loan, all other factors being equal.

To conclude, the odds of a customer defaulting in a loan repayment is 0.054 times higher for a customer who used mortgage as collateral than for a customer who used personal guarantee, all other factors being equal.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary

The findings from the data available to the researcher for the purpose of the study indicates that, among the various types of loans accessed, Commercial loan was given the highest amount approved, it was followed by Personal loan, then Transport loan and Agric loan was given the lowest amount approved for the period 2006-2010 fiscal year.

It was observed that from the Pearson Chi-square test of independence, Type of Loan and Security were dependent on the Repayment Status whiles Sex, Marital Status, Educational Level and Town Dummy were independent on Repayment Status of the customer. The Cochran Armitage Trend Test which was performed to determine if increase in interest rate places an increasing risk of loan default repayment showed that an increase in interest rate does not place an increasing risk of loan default repayment.

It was clear that the from the interpretation of the SPSS output of the binary logistic regression model, Security and Type of Loan were the factors which significantly influenced whether a customer defaulted a loan repayment whiles Sex, Marital Status, Educational Level, years the customer has been operating with the bank and Town did not contribute significantly to the model.

The role of the banking institution as a financial intermediaries in the Ghana financial sector cannot be overemphasized. Over the past decade the proliferation of foreign banks into the economy is an attestation to the fact that Ghana has become an important

player in the financial circles. These banks play numerous roles, the most prominent being credit creation. Despite the noble idea of facilitating the growth of the economy these banks are fraught with the non-payment of loans that turn to cripple their operations.

Akatakyiman Rural Bank Ltd is one of the many rural banks operating in Ghana and are advancing loans of various types to their customers with a special dispensation to the rural sector. The bank has been able to perform this responsibility creditably.

Record from the banking industry indicate that there have been apparent and inherent difficulties with the repayment of loans by customers once the loan have been advanced. Over the years the loan default rate has been on the ascendancy with all banks being victims of these seemingly bad practice.

With Akatakyiman Rural Bank Ltd as the main focal point the objective of this study was to determine some factors that influences loan repayment which is a canker to the operations of the bank. Therefore with the aid of statistical tool especially binary logistic regression analysis, the aim was to model the repayment status of the bank loans. The result obtained were indeed quite enlightening.

According to Oni O.A et al, (2005) a study on factors influencing loan default among poultry farmers in Ijebu Ode Local Government Area of Ogun State; the result from the probit model revealed that flock size of the farmers significantly influence default in loan repayment at ($P < 0.10$) level. Age of the farmers significantly influence default in loan repayment at ($P < 0.01$) level, while Educational level and Income of the farmers

also significantly influence default in loan repayment at ($P < 0.05$) level. These really shows that all the variables which were used can somehow influence loan default.

The idea was to study the effects that other variable under listed had on customer's ability to repay the loan taken up. The variables used were as follows: Type of loan, Security, Age, Sex, Marital Status, Town and Education Level.

Conclusion

It was found that among the variables that were used, Security and Type of Loan were significant to the study where as Sex, Marital Status, Age, Educational Level, Town were not significant to the study. We conclude that the risk of loan default for a customer who used collateral as a security in accessing the loan is less than for a customer who used personal guarantee. Taking transport loan as a reference group, the risks of a customer defaulting when given a personal loan is less than when given a transport loan, all other factors being equal.

Recommendations

We recommend that more variables be added or different variables should be used in exploring the variation in repayment status.

With the influx of financial institutions into the Ghanaian economy as a result of its astronomical growth and expansion, the issue of loan repayment will continue, to be an issue for all financial institutions, a further study is recommended with entirely different approach and variables. This study was to serve as a preparatory grounds for further analysis into the subject matter.

BIBLIOGRAPHY

1. Abolo, E. M. (1999), "General Leading Principles and Policies" First Bank of Nigeria Monthly Business and Economic Report. September. Provisioning, and Macroeconomic Linkages." IMF Working Paper, WP/00/195.

2. Ahmad, Nor Hayati, (2003). "Formation of credit risk, price Effect of Regulatory changes and the path Linking credit Risk and total Risk," PhD Thesis; University Utara Malaysia.

3. Ahmed, A.S. (1998)"Bank Loan Loss Provision: A re-examination of Capital Management, Earnings management and signaling Effects" Syracuse University, Syracuse 1-37.

4. Alan Agresti (2007). Introduction to categorical data analysis 2nd edition John Wiley and Sons.

5. Awoke M.U (2004). Factors affecting Acquisition and Repayment Patterns of smallholder farmers in the North East of Delta State, Nigeria. Journal of sustainable Tropical Agricultural Research 9:61-64.

6. Brewer, E.M, Jackson, W.E. III and Mondschean T.S., (1996). " Risk Regulation and S & L diversification into nontraditional assets" Journal of banking and Finance 20: 723-744.

7. Caprio G and Kilngebiel D (2002), "Episodes of Systemic and Borderine Financial Crisis", The World Bank (Unpublished).

8. Central Bank of Nigerian (1990), "Prudential Guidelines for Licensed Bank" Banking Supervision Department, (CBN).

9. Cortavarria Luis, C. Dziobek, A. Kanaya and I. Song (2000), "Loan Review, Provisioning, and Macroeconomic Linkages." IMF Working Paper, WP/00/195.

10. Cox, D.R and Snell E.J (1989). Analysis of Binary Data, 2nd Edition Chapman & Hall

11. Crockett, A. (1996). The theory and Practice Financial Stability, Economist, Vol. 144;4:531-568.

12. Deng, Y.J, J.M Qyigley, R. Van Order. (1996). Mortgage Default and Low Down Payment Loans. The cost of public subsidy. Regional Science and Urban Economics, 26, 263-285.

13. deServigny, A. and Renault, O. (2004), "Measuring and Managing Credit Risk" Mc GrawHill. New York.

14. Dorfman, P. M. (1998), "A Lenders' Guide to Lending Excellence (part 1) Credit Risk and Lenders" Deskmate. Vol No. 3. June-August. (Lagos: Capital).

15. Elaine, D. (2007), Risk Management: Bringing the Middle Officer to the Front. Zenith Economic Quarterly. Vol. 2, No. 10, April-June. (Lagos).

16. Fama, E. (1980). Banking in the theory of Finance, Journal of monetary Economics, Vol. 6;1:39-57.

17. Flint, Thomas A. (1997). Predicting Student Loan Defaults. Journal of Higher Education 68 (3): 322-54.

18. Goldstein, M. and Turner, P. (1996), "Banking Crises in Emerging Economics: Origins and Policy Options", BIS Economic Paper, No. 46. (Basel: Bank for International Settlements).

19. Hooman Mansori (2009), "Factors Affecting on loan Repayment Performance of farmers in Khorosan-Razavi province of Iran. University of Hamburg, Oct 6-8, 2009 conference on International Research on food security, natural Resource management and Rural Development.

20. Hosmer, D and Lemeshow, S., (2000). Applied Logistic Regression 2nd Edition. John Wiley and Sons.

21. Hudson, D. (2007). Agricultural markets and Prices 1st edition. Blackwell publishing.

22. Khan, T. and Amhed, H. (2001) "Risk management-An analysis of Issues in Islamic Financial industry" Islamic Development Bank-Islamic Research and Training Institute, Occasional Paper (No. 5) Jeddah.

23. Koopahi, M and Bakhshi, M.R (2002). Factors affecting Agricultural Credit Repayment Performance: (case study in Birjand district) Iranian journal of agricultural sciences 33, 1, 11-19.

24. Makiyan, S.N, (2003). "Role of Rate of Return on Loans in the Islamic Banking System of Iran". Managerial Finance, 25 (7).

25. Mankiv, N.G (2003) macroeconomics, 5th Edition, worth publishers

26. Monteverde, Kirk. (2000). Managing Student Loan Default Risk: Evidence from a Privately Guaranteed Portfolio. Research in Higher Education 41(3): 331-352.

27. Nwankwo, G. O. (1991), "Risk Analysis and Management, Principles and Practice".Malt House Press Ltd. Lagos.

28. Oladeebo J.O, Oladeebo O.E (2008). Determinant of loan Repayment among smaller holder farmers in Ogbomoso Agricultural zone of Oye Nigeria. Journal 8iof social science 17(1):59-62.

29. Oladeebo O.E (2003), Socio economic factors influencing Loan repayment among scale farmers in Ogbomoso. Agricultural zone of the Oye state, Nigeria. Diploma thesis unpublished. Ogbomoso: Ladoka A Kintola University.

30. Oni O.A, Oladele, O.I and Oyewole, I.K (2005). Analysis of factors influencing farmers in Ogum state Nigeria. Dept of Agricultural Economics, University of Ibadan Nigeria.

31. Quercia, R.G, Macarthy, G.W and Stegman M.A (1995). Mortgage default among Rural, Low-income Borrowers. Journal of Housing Research 6 (2), 349-369.

32. Rao, C.R (1973). Linear statistical inference and its Applications 2nd Edition. John Wiley and Sons

33. Rose, Peter S., (1996). "Commercial Bank management" McGraw Hill Cos. Inc USA; 196-190.

34. Samad, A. and Hasan, M.K, (1999). "The performance of Malaysia Islamic bank During 1984-1997; An Exploratory Study". International Journal of

Islamic Financial Services, 1 (3).

35. Schall, C. D. and Halley, C. (1980), "Introduction to Financial Management" New York. McGraw Hill Book Company, pp. 494

36. Steiner, Matt, and Natali Teszler. (2003). The Characteristics Associated with Student Loan Default at Texas A&M University. Produced by Texas Guaranteed in Association with Texas A&M University.

37. Sundararajain, V and Errico, L. (2002). "Islamic Financial Institutions and products in the Global Financial System: Key Issues in Risk management and challenges Ahead: IMF paper WP/01/192.

38. Thein, Tim, and Elizabeth Herr. (2001). Loan Default Model. Produced by Education First Marketing.

39. Van Order R. and Zorn, P. (2000). Income Location and Default. Some implications for community lending Real Estate Economics 28 (3) 385-404.

40. Volkwein, J. Fredericks et al. (1998). Factors Associated with Student Loan Default Among Different Racial and Ethnic Groups. The Journal of Higher Education 69 (2): 206-37.

41. Volkwein, J. Fredericks, and Alberto F. Cabrera. (1998). "Who Defaults on Student Loans? The Effects of Race, Class, and Gender on Borrower Behavior." in Condemning 18 Students to Debt: College Loans and Public Policy. ed. Richard Fossey and Mark Bateman.

42. Volkwein, J. Fredericks, and Bruce P. Szelest. (1995). Individual and Campus Characteristics Associated with Student Loan Default. Research in Higher Education 36 (1): 41-72.

43. Woo, Jennie H. (2002). Factors Affecting the Probability of Default: Student Loans in California. Journal of Student Financial Aid 32 (2): 5-25.

APPENDIX 1

LOGISTIC REGRESSION OUTPUT

```
DATASET NAME DataSet0 WINDOW=FRONT.
CROSSTABS
  /TABLES=Sex BY Repay_status
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT

  /COUNT ROUND CELL.
```

Crosstabs

[DataSet1] C:\Users\GOARGE\Documents\GEORGE DATA.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Sex * Repay_status	100	100.0%	0	.0%	100	100.0%

Sex * Repay_status Crosstabulation

Count

		Repay_status		
		No	Yes	Total
Sex	F	23	19	42
	M	25	33	58
Total		48	52	100

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.327[a]	1	.249		
Continuity Correction[b]	.901	1	.343		
Likelihood Ratio	1.328	1	.249		
Fisher's Exact Test				.312	.171
N of Valid Cases[b]	100				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 20.16.

b. Computed only for a 2x2 table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Marital_Status * Repay_status	100	100.0%	0	.0%	100	100.0%

Marital_Status * Repay_status Crosstabulation

Count

		Repay_status		
		No	Yes	Total
Marital_Status	M	44	48	92
	N	4	4	8
Total		48	52	100

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.014[a]	1	.906		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.014	1	.906		
Fisher's Exact Test				1.000	.596
N of Valid Cases[b]	100				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 3.84.

b. Computed only for a 2x2 table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Educ_Level * Repay_status	100	100.0%	0	.0%	100	100.0%

Educ_Level * Repay_status Crosstabulation

Count

		Repay_status		
		No	Yes	Total
Educ_Level	HIG	14	10	24
	LOW	34	42	76
Total		48	52	100

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.351[a]	1	.245		
Continuity Correction[b]	.861	1	.353		
Likelihood Ratio	1.353	1	.245		
Fisher's Exact Test				.349	.177
N of Valid Cases[b]	100				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.52.

b. Computed only for a 2x2 table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Security * Repay_status	100	100.0%	0	.0%	100	100.0%

Security * Repay_status Crosstabulation

Count

		Repay_status		
		No	Yes	Total
Security	MO	22	12	34
	PG	26	40	66
Total		48	52	100

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.760[a]	1	.016		
Continuity Correction[b]	4.791	1	.029		
Likelihood Ratio	5.817	1	.016		
Fisher's Exact Test				.021	.014
N of Valid Cases[b]	100				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.32.

b. Computed only for a 2x2 table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
town_dumy * Repay_status	100	100.0%	0	.0%	100	100.0%

town_dumy * Repay_status Crosstabulation

Count

		Repay_status		
		No	Yes	Total
town_dumy	K	29	29	58
	N	19	23	42
Total		48	52	100

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.221[a]	1	.638		
Continuity Correction[b]	.072	1	.789		
Likelihood Ratio	.221	1	.638		
Fisher's Exact Test				.688	.395
N of Valid Cases[b]	100				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 20.16.

b. Computed only for a 2x2 table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Loan_Type * Repay_status	100	100.0%	0	.0%	100	100.0%

Loan_Type * Repay_status Crosstabulation

Count				
		Repay_status		
		No	Yes	Total
Loan_Type	AGRIC	2	14	16
	COMME	20	21	41
	PERSO	19	6	25
	TRANS	7	11	18
Total		48	52	100

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.540[a]	3	.001
Likelihood Ratio	17.988	3	.000
N of Valid Cases	100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.68.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.540[a]	3	.001
Likelihood Ratio	17.988	3	.000
N of Valid Cases	100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.68.

Descriptive Statistics.

	N	Minimum	Maximum	Mean	Std Deviation
Age	100	21	6	41.99	9.315
Years With bank	100	2	20	10.17	6.358

```
LOGISTIC REGRESSION VARIABLES Repay_status
  /METHOD=ENTER Age Sex Marital_Status Educ_Level Security years town_dumy Loan_Type
  /CONTRAST (town_dumy)=Indicator
  /CONTRAST (Sex)=Indicator
  /CONTRAST (Educ_Level)=Indicator
  /CONTRAST (Security)=Indicator
  /CONTRAST (Marital_Status)=Indicator
  /CONTRAST (Loan_Type)=Indicator
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Logistic Regression

[DataSet1] C:\Users\GOARGE\Documents\GEORGE DATA.sav

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)
Loan_Type	AGRIC	16	1.000	.000	.000
	COMME	41	.000	1.000	.000
	PERSO	25	.000	.000	1.000
	TRANS	18	.000	.000	.000
Marital_Status	M	92	1.000		
	N	8	.000		
Educ_Level	HIG	24	1.000		
	LOW	76	.000		
Security	MO	34	1.000		
	PG	66	.000		
town_dumy	K	58	1.000		
	N	42	.000		
Sex	F	42	1.000		
	M	58	.000		

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Repay_status		Percentage Correct
	Observed		No	Yes	
Step 0	Repay_status	No	0	48	.0
		Yes	0	52	100.0
	Overall Percentage				52.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.080	.200	.160	1	.689	1.083

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Age	.006	1	.939
		Sex(1)	1.327	1	.249
		Marital_Status(1)	.014	1	.906
		Educ_Level(1)	1.351	1	.245
		Security(1)	5.760	1	.016
		years	1.458	1	.227
		town_dumy(1)	.221	1	.638
		Loan_Type(1)	9.617	1	.002
		Loan_Type(2)	.017	1	.896
		Loan_Type(3)	10.470	1	.001
	Overall Statistics		27.149	10	.002

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	31.345	10	.001
	Block	31.345	10	.001
	Model	31.345	10	.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.124[a]	.269	.359

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.800	8	.558

Classification Table[a]

Observed			Predicted		
			Repay_status		Percentage Correct
			No	Yes	
Step 1	Repay_status	No	37	11	77.1
		Yes	18	34	65.4
	Overall Percentage				71.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-.027	0.039	0.483	1	0.487	0.973
	Sex(1)	-.097	0.600	0.026	1	0.872	0.908
	Marital_Status(1)	-.514	0.884	0.338	1	0.561	0.598
	Educ_Level(1)	-.360	0.690	0.272	1	0.602	0.698
	Security(1)		1.104	7.033	1	0.008	0.054
	years	-2.927		0.000	1	0.998	1.000
	town_dumy(1)	.000	0.073				
		-.729	0.615	1.404	1	0.236	0.482
	Loan_Type(1)	1.554	0.987	2.476	1	0.116	4.729
	Loan_Type(2)	1.075	1.008	1.139	1	0.286	2.931
	Loan_Type(3)	-1.451	0.746	3.784	1	0.052	0.234
	Constant	2.975	1.981	2.254	1	0.133	19.589

a. Variable(s) entered on step 1: Age, Sex, Marital_Status, Educ_Level, Security, years, town_dumy, Loan_Type.