

**KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**INSTITUTE OF DISTANCE LEARNING**

**OPTIMIZING A BANK'S CREDIT RISK POLICY  
(A CASE STUDY OF PRUDENTIAL BANK LIMITED)**

**MICHAEL KOFI ASARE**



**A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS IN PARTIAL  
FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF  
MASTER OF SCIENCE IN INDUSTRIAL MATHEMATICS**

**JUNE, 2013**

## DECLARATION

I hereby declare that this submission is my own work towards the award of Master of Science degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for award of any other degree of the university, except due acknowledgement has been made in the text.

Asare, Michael Kofi, (PG 6317911)



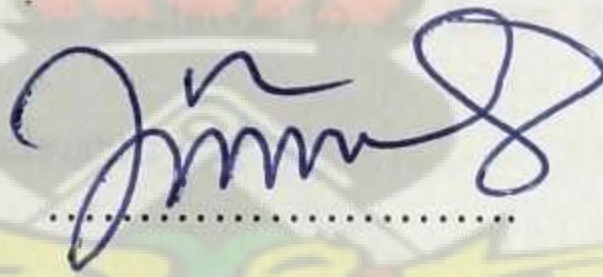
03 Oct-13

Student's Name & ID

Signature

Date

Certified by



26/5/12

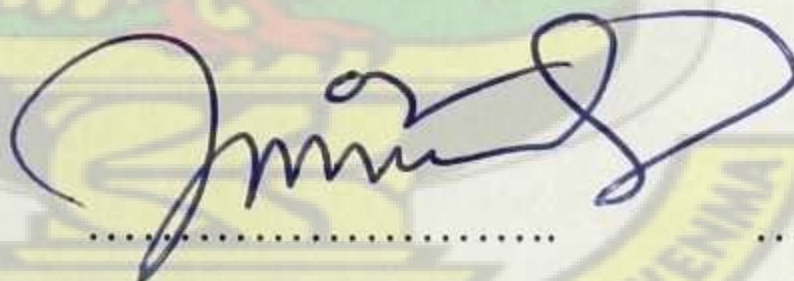
Prof. S. K. Amponsah

Signature

Date

(Supervisor)

Certified by



26/5/12

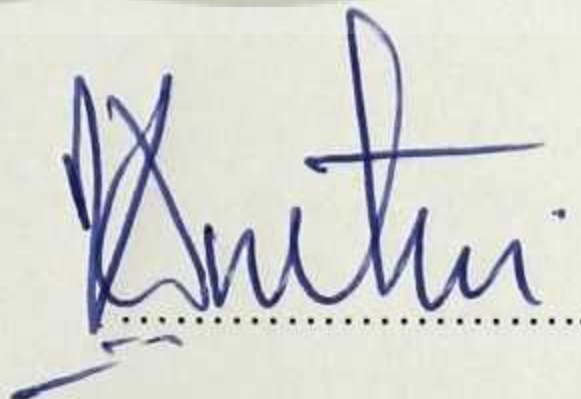
Mr. K. F. Darkwa

Signature

Date

(Head of Department)

Certified by



Prof. I. K. Dontwi

Signature

Date

(Dean, IDL)

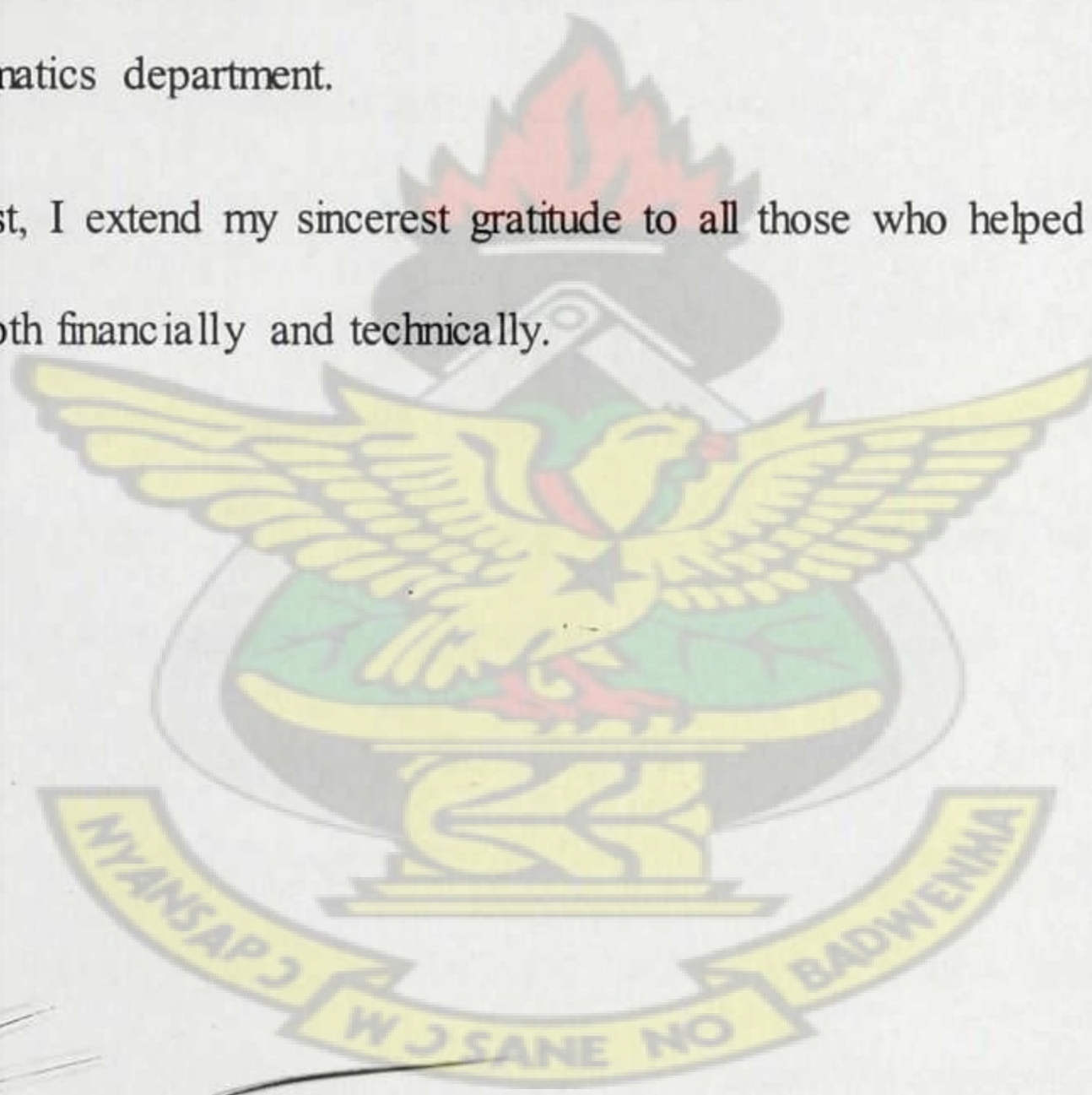
## ACKNOWLEDGEMENT

First and foremost I am grateful to the Almighty God for His grace and mercies which has seen me through my life.

I acknowledge the Management, the credit officers' core and the entire staff of the Adenta Branch, all of Prudential Bank Limited, for their support.

The study would not have been possible without the efforts of my supervisor, Prof. S. K. Amponsah. His dedication and valuable guidance made this work possible. I also wish to express my gratitude to all the lecturers in the Mathematics department.

Lastly but not the least, I extend my sincerest gratitude to all those who helped in diverse way to make this thesis a success both financially and technically.



## DEDICATION

To the Glory of God, I dedicate this work to my family.



## ABSTRACT

Financial institutions in Ghana are concerned with the “optimal allocation of limited resources to meet desired objectives.” Two mathematical models are proposed to help the financial institutions make the optimum allocation of the available scarce resources, in achieving their desired goals of maximizing the profit and minimizing the costs. The purpose of this paper is to show the practical application of Linear Programming and Logistic Regression Models in the formulation of an optimal bank credit policy. Firstly, we formulate a Linear Programming Model and develop a solution (using the Simplex Algorithm) that optimally allocates funds, where a financial institution is facing the problem of allocation of limited funds among different types of loans / advances at different markup / interest rates with varying degree of risk (bad debts). We go further, after optimal allocation of funds, to propose a Binary Logistic Regression Model (BLRM) to discriminate loan defaulters from non-defaulters. The study revealed that the available funds of GH¢166 million for credit facilities will yield a return of GH¢35.25 million (ignoring time value of money) if the funds were allocated to the economic sectors in following manner; GH¢55.44 million for Commerce, GH¢5.49 million for Agriculture, GH¢32.21 million for Education, GH¢27.43 million for Consumer, GH¢31.37 million for Export and GH¢14.10 for other financial institutions. The study also revealed that the Bank (Prudential Bank Limited) should not allocate any funds to the construction/manufacturing sector, this will save them some GH¢10,907.45. Finally, the BLRM proposed predicts that about 82% of prospective customers are likely not to default on loans granted them.

## TABLE OF CONTENT

| CONTENT                                   | PAGE |
|---|------|
| DECLARATION.....                          | i    |
| DEDICATION.....                           | ii   |
| ACKNOWLEDGEMENT .....                     | iii  |
| ABSTARCT.....                             | iv   |
| CONTENTS.....                             | v    |
| LIST OF TABLES.....                       | viii |
| LIST OF FIGURES.....                      | ix   |
| <br><b>CHAPTER 1</b>                      |      |
| <b>INTRODUCTION</b>                       |      |
| 1.0 Overview.....                         | 1    |
| 1.1 Background to the Study.....          | 1    |
| 1.2 Statement of the Problem.....         | 12   |
| 1.3 Objectives .....                      | 13   |
| 1.4 Justification .....                   | 13   |
| 1.5 Methodology .....                     | 14   |
| 1.6 About the Financial Institution ..... | 14   |
| 1.7 Limitation: .....                     | 16   |

|     |                                 |    |
|-----|---------------------------------|----|
| 1.8 | Organization of the Study ..... | 16 |
|-----|---------------------------------|----|

|     |                      |    |
|-----|----------------------|----|
| 1.9 | Chapter Summary..... | 16 |
|-----|----------------------|----|

## CHAPTER 2

### LITERATURE REVIEW

|     |                   |    |
|-----|-------------------|----|
| 2.0 | Introduction..... | 17 |
|-----|-------------------|----|

|     |                                   |    |
|-----|-----------------------------------|----|
| 2.1 | Studies on Linear Programing..... | 18 |
|-----|-----------------------------------|----|

|     |                                     |    |
|-----|-------------------------------------|----|
| 2.2 | Studies on Logistic Regression..... | 27 |
|-----|-------------------------------------|----|

|     |              |    |
|-----|--------------|----|
| 2.3 | Summary..... | 35 |
|-----|--------------|----|

## CHAPTER 3

### METHODOLOGY

|     |                   |    |
|-----|-------------------|----|
| 3.0 | Introduction..... | 36 |
|-----|-------------------|----|

|     |                        |    |
|-----|------------------------|----|
| 3.1 | Linear Programing..... | 36 |
|-----|------------------------|----|

|     |                          |    |
|-----|--------------------------|----|
| 3.2 | Logistic Regression..... | 46 |
|-----|--------------------------|----|

|     |              |    |
|-----|--------------|----|
| 3.3 | Summary..... | 56 |
|-----|--------------|----|

## CHAPTER 4

### DATA COLLECTION AND MODELING

|     |                   |    |
|-----|-------------------|----|
| 4.0 | Introduction..... | 57 |
|-----|-------------------|----|

|     |                                    |    |
|-----|------------------------------------|----|
| 4.1 | The Linear Programing Problem..... | 57 |
|-----|------------------------------------|----|

|     |                               |    |
|-----|-------------------------------|----|
| 4.2 | The Discriminating Model..... | 61 |
| 4.3 | Summary.....                  | 69 |

**CHAPTER 5**

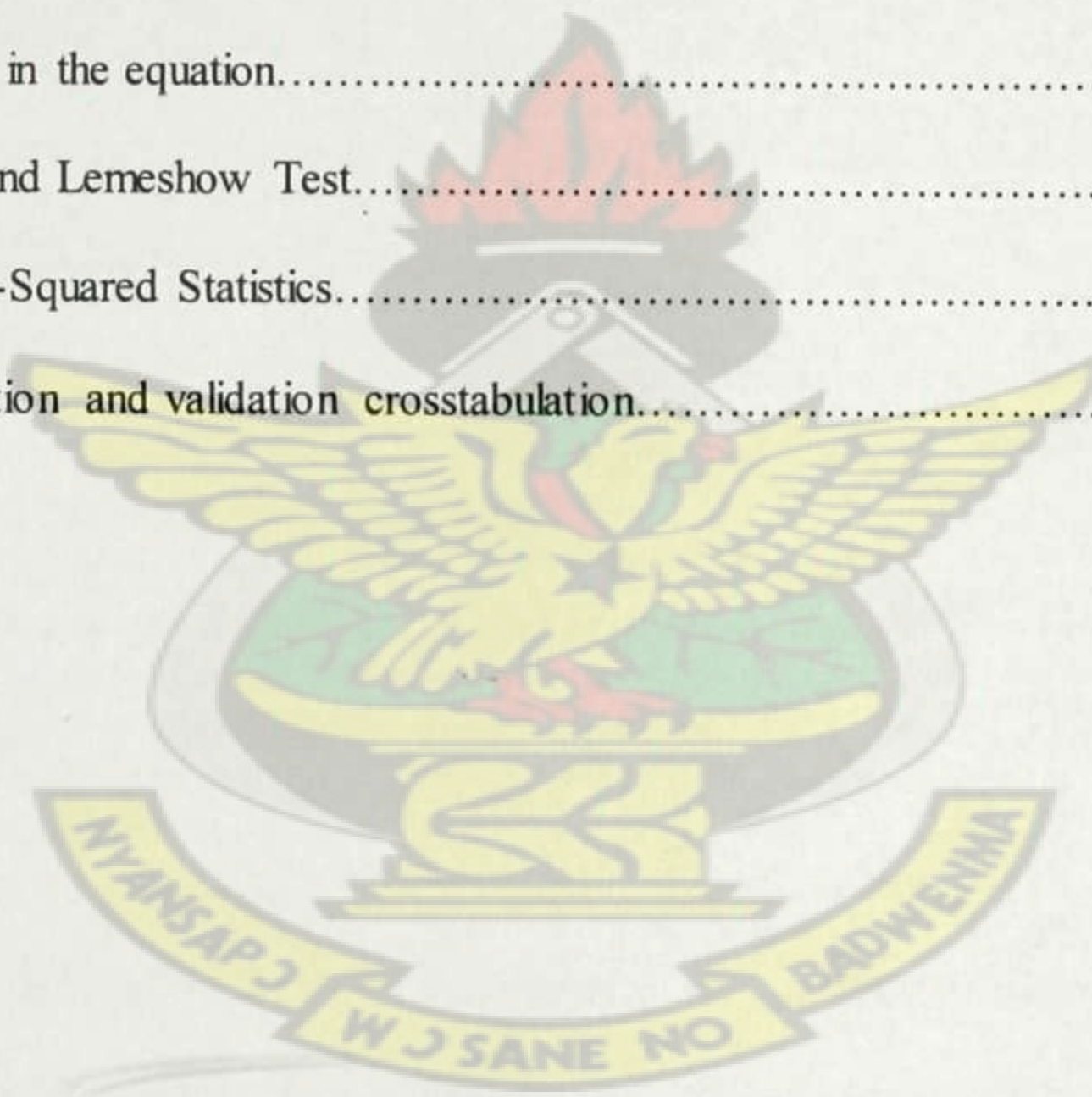
**CONCLUSION AND RECOMMENDATIONS**

|     |                         |    |
|-----|-------------------------|----|
| 5.0 | Introduction.....       | 71 |
| 5.1 | Summary of Results..... | 71 |
| 5.2 | Conclusion.....         | 73 |
| 5.3 | Recommendation.....     | 73 |
|     | References.....         | 75 |
|     | Appendix I.....         | 80 |
|     | Appendix II.....        | 83 |



## LIST OF TABLES

|            |   |    |
|------------|---|----|
| Table 3.1: | General form for Initial Simplex Tableau.....                       | 45 |
| Table 3.2: | ML parameter estimates.....   | 48 |
| Table 3.3: | Computer output for Horseshoe Crabs with Width and Color predictors | 54 |
| Table 4.1: | Economic sectors where PBL supports with credit facilities.....     | 57 |
| Table 4.2: | Proposed allocation to maximize return on GH¢166 million.....       | 60 |
| Table 4.3: | Case Processing Summary of Validate versus Previously defaulted...  | 61 |
| Table 4.4: | Crosstabulation of Validate and Previously defaulted counts.....    | 62 |
| Table 4.5: | Variables in the equation.....                                      | 63 |
| Table 4.6: | Hosmer and Lemeshow Test.....                                       | 65 |
| Table 4.7: | Pseudo R-Squared Statistics.....                                    | 66 |
| Table 4.8: | Classification and validation crosstabulation.....                  | 68 |



## LIST OF FIGURES

|             |   |    |
|-------------|---|----|
| Figure 3.1: | Logistic Regression Functions.....  | 47 |
| Figure 3.2: | Logistic Regression Model with width and colors as predictors.....  | 55 |
| Figure 4.1: | Change in deviance versus Predicted probabilities.....  | 65 |
| Figure 4.2: | Frequency distribution of Predicted Probability of defaulters by Actual<br>observed defaulters status (No/Yes)..... | 67 |
| Figure 4.3: | Distribution of Predicted Probabilities of Loan Default Among<br>Prospects.....                                     | 69 |



## **CHAPTER 1**

### **1.0 INTRODUCTION**

In this chapter we will put forward the general problem facing financial institutions in Ghana and discuss two mathematical models that can be used by these financial institutions to reduce their credit risk.

The fundamental issues facing senior bank management revolve around the structuring of a bank's credit risk policy. The desired outcome of such a policy is to maximize profit. In order to formulate such a policy to obtain the desired outcome, the management of the bank has at their disposal several optimization techniques. Linear Programming and Regression models are two of the modeling techniques that can be used to obtain such a desired outcome.

### **1.1 BACKGROUND OF THE STUDY**

#### **1.1.1 ORIGINS OF LINEAR PROGRAMMING**

Linear Programming is viewed as a revolutionary development giving us the ability for the first time to state general objectives and to find optimal policy decisions for a broad class of practical decision problems of great complexity.

Industrial production, the flow of resources in the economy, the exertion of military effort in a war, the management of finances all require the coordination of interrelated activities. What these complex undertakings share in common is the task of constructing a statement of actions to be performed, their timing and quantity that if implemented, would move the system from a given initial status as much as possible towards some defined goal.

While differences may exist in the goals to be achieved, the particular processes, and the magnitudes of effort involved, when modeled in mathematical terms, these seemingly disparate systems often have a remarkably similar mathematical structure. The computational task is then to devise for these systems an algorithm for choosing the best schedule of actions from among the possible alternatives.

The observation, in particular, that a number of economic, industrial, financial, and military systems can be modeled by mathematical systems of linear inequalities and equations has given rise to the development of linear programming field Dantzig (1949).

In 1947 G. B. Dantzig proposed the Simplex algorithm as an efficient method to solve a linear programming problem. He was then working in the SCOOP group (Scientific Computation of Optimum Programs), an American research program that resulted from the intensive scientific activity during the Second World War, in the USA, aimed at rationalising the logistics in the war effort. In the Soviet Union, Kantorovitch had already proposed a similar method for the analysis of economic plans, but his contribution remained unknown to the general scientific community. It seems also that 19th century mathematicians, in particular Fourier (1823), had also thought about similar methods. What made the contribution of Dantzig so important is its concomitance with two other phenomena.

- The considerable development of the digital computer that permitted the implementation of the algorithm to solve full size real life problems;
- The parallel development of the paradigm of inter industry exchange table, also known as the input/output matrix, proposed by Leontieff (1941) which showed that the whole economy could be represented in a sort of linear programming structure. Therefore the method of linear programming was providing both an efficient instrument to compute solutions of large scale linear optimization

problems and a general paradigm of economic equilibrium between different sectors exchanging resources and services.

### 1.1.2 ORIGINS OF LOGISTIC REGRESSION

The complete name of the correlation coefficient deceives many students into a belief that Karl Pearson developed this statistical measure himself. Although Pearson did develop a rigorous treatment of the mathematics of the Pearson Product Moment Correlation (PPMC), it was the imagination of Sir Francis Galton that originally conceived modern notions of correlation and regression. Galton, a cousin of Charles Darwin and an accomplished 19th century scientist in his own right, has often been criticized in this century for his promotion of "eugenics" (planned breeding of humans; see, for example, Paul (1995). Historians have also suggested that his cousin's lasting fame unfairly overshadowed the substantial scientific contributions Galton made to biology, psychology and applied statistics (see, for example, FitzPatrick, 1960). Galton's fascination with genetics and heredity provided the initial inspiration that led to regression and the PPMC.

Logistic regression is used extensively in numerous disciplines, including the medical and social science fields. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al., 1987, using logistic regression. It is also employed in marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription, etc.

These interesting subjects have been applied in many fields not because of their mathematical elegance but also because of their practical utilization in modeling. Both models will be the basis of this thesis.

One of the purposes of Banks is to facilitate, accumulate and allocate capital by channeling individual savings into loans to governments and businesses. Their transaction consists of making loans to customers and purchase of investment securities in the capital and or money market place.

The Ghanaian financial sector has moved from a repressed era to a more liberal environment as a response to the international calls of the World Bank and the International Monetary Fund. Several benefits, which are reminiscent of a liberal financial market, have been enjoyed as evidenced by the contribution of the financial liberalization policy to the improved economic growth and development of the Ghanaian economy. In view of this the banking industry has seen a massive growth in structure i.e. increase in branch network, provision of wide range of banking services and technological advancement.

Credits facilities are usually granted in a form of contract between the bank and the borrower. The bank will like to know the purpose of the credit facility and whether the use of the facility will yield results.

Due to poor allocation of their credit facility disbursement they are not able to optimize their profits when they give out these facilities hence this study therefore proposes two models, each at a different decision making level to enable credit facilities to be disbursed optimally for both short term and long term basis in order to maximize their profit.

### **1.1.3 OVERVIEW OF THE BANKING SECTOR IN GHANA**

Over time, technology has increased in importance in Ghanaian banks. Traditionally, banks have always sought media through which they would serve their clients more cost-effectively as well as increase the utility to their clientele. Their main concern has been to serve clients more conveniently, and in the process increase profits and competitiveness.

Bankers play very important role in the economic life of the nation. The health of the economy is closely related to the soundness of its banking system. Although banks create no new wealth but their borrowing, lending and related activities facilitate the process of production, distribution, exchange and consumption of wealth. In this way they become very effective partners in the process of development. Today modern banks are very useful for the utilization of the resources of the country. The banks are mobilizing the savings of the people for the investment purposes. If there would be no banks then a great portion of a capital of the country would remain idle. Credit facilities provided by banks works as an incentive to the producer to increase the production.

#### **1.1.4 TYPES OF BANKS**

##### **1.1.4.1 Central Bank**

A central bank, reserve bank, or monetary authority is the entity responsible for the monetary policy of a country or of a group of member states. Its primary responsibility is to maintain the stability of the national currency and money supply, but more active duties include controlling subsidized-loan interest rates, and acting as a lender of last resort to the banking sector during times of financial crisis (private banks often being integral to the national financial system). It may also have supervisory powers, to ensure that banks and other financial institutions do not behave recklessly or fraudulently. In Ghana the Central Bank is known as Bank of Ghana, which is, fully owned by the Government. It is governed by a central board (headed by a Governor) appointed by the Central Government. It issues guidelines for the functioning of all banks operating within the country.

##### **1.1.4.2 Commercial Banks**

A commercial Bank performs all kinds of banking functions such as accepting deposits, advancing loans, credit creation and agency functions. They generally advance short term loans to their customers;

in some cases they may give medium term loans. This broader definition includes many other financial institutions that are not usually thought of as banks but which nevertheless provide one or more of these broadly defined banking services. These institutions include finance companies, investment companies, investment banks, insurance companies, pension funds, security brokers and dealers, mortgage companies.

#### **1.1.4.3 Industrial Banks**

Ordinarily, the industrial banks perform three main functions: Firstly, acceptance of long term deposits: Since the industrial banks give long term loans, they cannot accept short term deposits from the public. Secondly, meeting the credit requirements of companies: in this case, the industries require the purchasing of land to erect buildings and purchasing of heavy machinery. The industries also require short term loans to buy raw materials and to make payment of wages to workers. Thirdly, it does some Other Functions - The industrial banks tender advice to big industrial firms regarding the sale and purchase of shares and debentures

#### **1.1.4.4 Agricultural Banks**

As the commercial and the industrial Banks are not in a position to meet the credit requirements of agriculture, there arises the need for setting up special types of banks to finance agriculture. Firstly, the farmers require short term loans to buy seeds, fertilizers, ploughs and other inputs. Secondly, the farmers require long term loans to purchase land, to effect permanent improvements on the land to buy equipment and to provide for irrigation works.

#### **1.1.4.5 Foreign Exchange Banks**

Their main function is to make international payments through the purchase and sale of exchange bills. As is well known, the exporters of a country prefer to receive the payment for their exports in their own

currency. Hence there arises the problem of converting the currency of one country into the currency of another. The foreign exchange bank tries to solve this problem. These banks specialize in financing foreign trade.

#### **1.1.4.6 Rural Banks**

The Rural and Community banks are unit banks owned by members of the rural community through purchase of shares and are licensed to provide financial intermediation in the rural areas.

### **1.1.5 CREDIT FACILITIES**

Lending is the principal business activity for every financial institution, and credit facilities are its major source of revenue. Conversely, credit facilities also represent the greatest source of risk to the institution's safety and soundness, and they have been the major cause of losses and institutional failures. Because of their predominate importance to the existence and success of the institution, the assets within the credit portfolio continually warrant the highest and best management skills and the most effective tools and processes to manage and control the opportunities and inherent risks.

Lending can expose a bank's earnings and capital to all of the risks. For most banks, credit facilities are the largest and most obvious source of credit risk.

Credit facilities can be classified under three main categories; Overdraft facility, Loan facility and Bank Bonds and Guarantees.

#### **1.1.5.1 Overdraft Facility**

A credit agreement made with a financial institution that permits an account holder to use or withdraw more than they have in their account, without exceeding a specified maximum negative balance. Establishing an overdraft facility with a bank can help an individual or small business with short term

cash flow problems. Although the negative balance typically needs to be repaid within a month, the competition between the Banks in Ghana has allowed the period to be extended to six months and or one year.

### **1.1.5.2 Loan Facility**

#### **1.1.5.2.1 Working Capital Loan**

Working capital loan covers only expenses incurred by existing capital and human resources (e.g. utilities, rents, payroll, etc.). Working capital loans are generally granted only to companies with a high credit rating, and are only meant to be used until a company can generate enough revenue to cover its own expenses.

#### **1.1.5.2.2 Term Loan**

A bank loan, typically with a floating interest rate, for a specified amount that matures in between one and ten years, and requires a specified repayment schedule. This is for building projects, purchase of equipment, purchase of vehicles and other fixed assets. Salaried Workers' Loans, Car Loans, Commercial Loans, Agricultural Loans and Commercial and Industrial mortgages all fall under this category.

### **1.1.5.3 Bonds and Guarantees**

#### **1.1.5.3.1 Bank Guarantee**

It is an irrevocable commitment by a bank to pay a specified sum of money in the event that the party requesting the guarantee fails to perform the promise or discharge the liability to a third person in case of the requestors default.

#### **1.1.5.3.2 Advance Payment or Mobilization Guarantee**

A guarantee issued by a bank, on behalf of a seller to a buyer, in relation to any advance payment that is made by the buyer to the seller to allow the contract to commence. If the contract is not completed the buyer can claim reimbursement of the advance payment under the guarantee.

#### **1.1.5.3.3 Performance Security/Guarantee**

A written guaranty from a third party guarantor (usually a bank or an insurance company) submitted to a principal (client or customer) by a contractor on winning the bid. A performance bond ensures payment of a sum (not exceeding a stated maximum) of money in case the contractor fails in the full performance of the contract. Performance bonds usually cover 100 percent of the contract price and replace the bid bonds on award of the contract. Unlike a fidelity bond, a performance bond is not an insurance policy and (if cashed by the principal) the payment amount is recovered by the guarantor from the contractor, also called standby letter of credit or Contract performance bond.

#### **1.1.5.3.4 Bid Security/Bond**

A written guaranty from a third party guarantor (usually a bank or an insurance company) submitted to a principal (client or customer) by a contractor (bidder) with a bid.

A bid bond ensures that on acceptance of a bid by the customer the contractor will proceed with the contract and will replace the bid bond with a performance bond. Otherwise, the guarantor will pay the customer the difference between the contractor's bid and the next highest bidder. This difference is called liquidated damages, which cannot exceed the amount of the bid bond. Unlike a fidelity bond, a bid bond is not an insurance policy, and (if cashed by the principal) the payment amount is recovered by the guarantor from the contractor. Also called bid guaranty or bid surety.

#### **1.1.5.3.5 Letters of Credit (LC or L/C)**

An LC is a document issued mostly by a financial institution which usually provides an irrevocable payment undertaking (it can also be revocable, confirmed, unconfirmed, transferable or revolving but is most commonly irrevocable/confirmed) to a beneficiary against complying documents as stated in the LC. It is often referred to as a documentary credit (DC or D/C) or simply as credit. Once the beneficiary or a presenting bank acting on its behalf, makes a presentation to the issuing bank or confirming bank, if any, within the expiry date of the LC, comprising documents complying with the terms and conditions of the LC, the applicable international standard banking practice, the issuing bank or confirming bank, if any, is obliged to honor irrespective of any instructions from the applicant to the contrary. In other words, the obligation to honor (usually payment) is shifted from the applicant to the issuing bank or confirming bank, if any. Non-banks can also issue letters of credit; however parties must balance risks. There are three types of L/C; Sight letters of Credit, Deferred Payment Letters of Credit and Revolving Deferred Payment Letters of Credit.

#### **1.1.6 TERMS OF BANK CREDIT**

##### **1.1.6.1 Secured Loan**

A secured loan is a loan in which the borrower pledges some asset (e.g. a car or property) as collateral for the loan.

##### **1.1.6.2 Unsecured Loan**

Unsecured loans are monetary loans that are not secured against the borrower's assets. These may be available from financial institutions under many different guises or marketing packages.

#### **1.1.6.3 Line of credit**

This is a pre-approved loan for up to 24 months, with the bank setting a maximum Credit line and giving the company a time limit to draw against it. The company may be required to pay an upfront commitment fee of up to one percent of the credit line.

#### **1.1.6.4 Factoring**

This type of loan is used when the business cannot qualify for accounts receivable financing. The business sells its receivables to a factoring institution called a factor and receives discounted funds immediately. The factor assumes the credit risks and collection responsibilities, with end customers making payment to a postal address held by the factor. The factor will refuse any invoices it considers to be high-risk, and the cost of this type of loan is very high.

### **1.1.7 SOURCES OF CREDIT**

Banking is the business of providing financial services to consumers and businesses. The basic services a bank provides are checking accounts, which can be used like money to make payments and purchase goods and services; savings accounts and time deposits that can be used to save money for future use; credit facilities that consumers and businesses can use to purchase goods and services and basic cash management services such as check cashing and foreign currency exchange. Four types of banks specialize in offering these basic banking services; these are commercial banks, Savings and loan associations, savings banks, and credit unions.

#### **1.1.8 ABUSES IN CREDIT FACILITY**

One form of abuse in the granting of credit facility involves granting credit in order to put the borrower in a position that one can gain advantage over. Another form of abuse is where the lender charges excessive interest. In different time periods and cultures the acceptable interest rate has varied, from no

interest at all to unlimited interest rates. Credit card companies in some countries have been accused by consumer organizations of lending at usurious interest rates and making money out of frivolous extra charges. Abuses can also take place in the form of the customer abusing the lender by not repaying the loan or with intent to defraud the lender.

## 1.2 PROBLEM STATEMENT

There has been an increase in the number of banks in Ghana over the past decade. This has made the banking industry in Ghana go through many structural changes. This change includes increases in branch network, provision of wide range of banking services and acceleration of credit activities in different ways. The performance level of a bank is not measured based only structural or physical growth but its assets. With the recent advent of the global financial crisis initiated by the collapse of the US mortgage loan markets, Young-Han and Kim (2009), there has been growing numbers of portfolio managers who spend the majority of their day in front of computers managing their portfolio investments, creating software, mitigating risks, and communicating with their research team. Only a few banks are keen on the proper allocation of funds for the specific type of credit to right individuals or organization in the right proportions in order to maximize returns on the disbursement of credit facilities. Credit facilities are usually given based on the type of credit the customer demands and the amount of money available. There is a second problem that needs attention; most Banks don't have discriminating models that differentiate between defaulters from non-defaulters. Since Credit portfolio which is considered as the Banks major asset needs a constant acceleration for the growth of the Bank. It is therefore imperative that a study be conducted to help banks optimally allocate their funds for credit facilities and be able to discriminate between defaulters and non-defaulters with a high degree of accuracy, these will optimally lead to maximization of profits.

### 1.3 OBJECTIVES OF THE STUDY

The general objective of the study aims at obtaining optimal loan allocation policy to enable appropriate disbursement for maximization of profits and to discriminate between defaulters and non-defaulters.

This study seeks to specifically:

- (i) Propose a Linear Programming Model that optimally allocates funds available for credit facilities.
- (ii) Propose a Binary Logistic Regression Model that discriminates defaulters from non-defaulters of credit facilities.

### 1.4 JUSTIFICATION OF THE STUDY

In today's world money is a scarce commodity for a lot of people. No matter how hard you work there is always a need for more money. Hence banks and financial institutions have taken an advantage of this current need for money to satisfy various needs. The money given out also serves a useful purpose because the loan portfolio of bank which comprises of the principal and the interest gained on the loan serves as the major asset of bank.

Strategic portfolio planning is a major segment of the institution's overall business and capital planning process and a primary component of effective portfolio management.

Bank failures are widely perceived to have greater adverse effects on the economy and thus are considered more important than the failure of other types of business firms. A bank fails economically when the market value of its assets declines below the market value of its liabilities. It is also imperative that loan defaults be kept at the minimum, failure to do so will also lead to the same predicament.

Hence optimal allocation of funds alone will not suffice. Since the credit portfolio benefits both the lender and the borrower, it is of essence that optimal allocation of funds to the different types of credit

facilities coupled with systemic discriminating between defaulters from non-defaulters will yield maximized returns.

## 1.5 METHODOLOGY

The data will be a secondary data and will be collected from the chosen bank, Prudential Bank Limited (PBL). The type of data to be collected from the bank will be the actual credit facilities granted to borrowers/applicants as at 31<sup>st</sup> December, 2012.

Default rate for each economic sector is presented in the data. We would calculate the recovery rate for each of the seven economic sectors presented by PBL. Allocation of funds among different sectors of the economy (in Ghana) at different interest rates with varying degree of risk (bad debts) shall be modeled using Linear Programming. The Simplex Method developed by G. B. Dantzig (1947) which has been coded in solver in Microsoft Excel software will be used to solve the Linear Programming problem, hence forecast of credit disbursement for future allocation of each type will be determined.

The data also contains the demographical data and certain indicative variables of individuals who have either applied for or enjoyed a credit facility. A discriminating model (Logistic Regression Model) using SPSS software shall be executed on this data to differentiate between defaulters and non-defaulters.

## 1.6 ABOUT THE FINANCIAL INSTITUTION

Prudential Bank Limited (PBL) operates in the Banking and Finance industry within the legal framework of the Banking Act, 2007. All the commercial banks in Ghana are being supervised by the Central bank (Bank of Ghana).

Prudential Bank Ltd (PBL) was promoted by Messrs. J.S. Addo Consultants and incorporated as private limited liability company in 1993 under the companies code Act 179. The bank opened its doors for business in August 1996 and focused on the development and financing of industry and exports .Over

the years Prudential Bank has broadened its scope of operations to actively support small and medium scale businesses with 30 branches located in Accra, Tema, Kumasi, Takoradi, Cape coast and Tamale.

Corporate mission of the bank is to provide domestic and international banking services with a strategic focus on project financing and export development. PBL is committed to playing a positive and innovative role in the financial intermediation process and, most importantly, to offer the best and the most remunerative banking services to the business community. **Quality, Creativity and Innovation** are the hall marks of Prudential Bank Limited.

PBL offer the following products and services; Current Account, Savings Account, Call Deposit Account, Fixed/Time Deposit Account, Certificate of Deposit, Funds management, Transfer of funds, Project financing, Export financing, Mortgage financing, Consumer credit facility, Cheque collection, Issue of Bonds and Guarantees, Short/Medium/Long Term Credit facilities, Import/Export Letters of Credits, Inward and Outward Bills for collection, Negotiation of Export Bills, Foreign Currency/Foreign Exchange Accounts, Transfer/Remittances, Export Advisory Services, ATM/Internet/e-Banking Products, e-Zwich services.

Prudential Bank has two subsidiary companies- Prudential Properties Limited (PPL) and Prudential Securities Limited (PSL), an investment banking company engaged in stockbrokerage, funds management, corporate finance and business advisory services as well as equity and economic research.

The bank's shareholders are:

- Messrs J.S Addo consultants Ltd
- The Social Security and National Insurance Trust (SSNIT)
- National Trust Holding Company Ltd (NTHC)
- Ghana Union Assurance Company Ltd
- PBL staff provident fund and

- Four Individual Ghanaians

## 1.7 LIMITATIONS ON THE STUDY

There is difference in the behavior of corporate (including enterprises, partnerships, club and societies) and individual accounts. The predictive variables for building the logistic regression model for each group (corporate, individual) are quite different, hence this study will focus on individual accounts and propose a binary logistic regression model that will discriminate between defaulters and non-defaulters using demographic and behavioral characteristics of individuals.

## 1.8 ORGANISATION OF THE STUDY

This thesis is organized as follows: The first chapter is a general introduction of the study including the problem statement, research objectives, justification and a brief methodology of the study. Chapter two focus mainly on literature review on previous study or work done using the Linear Programming and Logistic Regression Models. Chapter three provided the details of the research method while chapter four entails analysis and data presentation. In chapter five, the final chapter, the summary of the study is put together with the recommendations.

## 1.9 SUMMARY

In this chapter we considered the background of the proposed models to be used for the analysis, the problem statement and the objectives of the study. The justification and methodology was also put forward.

In the next chapter, we shall present the relevant and adequate literature on the proposed models.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.0 INTRODUCTION

In this chapter, we present some studies done by other researchers (authors) on the proposed mathematical models and the practical use of these models in solving real life problems.

Mathematical and Statistical Modeling might broadly be defined as the process of building effective descriptions of complex statistical data. It makes use of the tools of statistical methodology but has a strong focus on applications to real data. In that sense it acts as a bridge between the fundamental methods of the subject and important applications in a wide variety of areas.

Most business resource-allocation problems require the decision maker to take into accounts various types of constraints, such as capital, labor, legal, and behavioral restrictions. Linear programming techniques can be used to provide relatively simple and realistic solutions to problems involving constrained resource allocation decisions. A wide variety of production, finance, marketing, and distribution problems have been formulated in the linear programming framework.

Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure.

Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables are a categorical, or a mix of continuous and categorical, logistic regression is preferred.

## 2.1 STUDIES ON LINEAR PROGRAMING

The obligation to meet infinite needs with restricted resources is one of the biggest challenges encountered in the market today (Ozsan et al., 2010). Linear programming (LP) is a powerful analytical tool that can be used to determine an optimal solution that satisfies the constraints and requirements of the current situation (Better, 1988). This method consists of three quantitative components: (i) objective function (maximization of profit or minimization of costs); (ii) constraints (limitation of production sources); and (iii) decision variables. In formulating the linear programming problem, the assumption is that a series of linear (or approximately linear) relationships involving the decision variables exist over the range of alternatives being considered in the problem (Chinneck, 2004). LP output not only provides an optimal solution, it also provides sensitivity analysis. Sensitivity analysis evaluates how changes in the objective function coefficients affect the optimal solution of a linear programming model. It could examine how well the changes of objective function coefficients and the right hand side value could affect the optimal solution (Anderson et al., 2000).

LP has been used in the evaluation and optimization of raw material resources, capital, machinery, equipment, time and manpower under certain restricting circumstances to get the most benefit (Han et al., 2011).

Hassan (2005) used a linear programming model to determine the optimum cropping pattern as a prerequisite to efficient utilization of available resources of land, water, and capital for Pakistan's agriculture.

Bretas (1991) reported a general linear programming model which was developed to determine an income-maximizing set of management activities for a cash-crop farm subject to groundwater quality

standards for pesticide contamination. Ozsan et al., (2010) reported that a linear programming model was used to determine the maximum profit in marble processing plants.

Garver (1970) studied one aspect of long-range planning of electric power systems which involves the exploration of various designs for the bulk power transmission network. He used linear programming for network analysis to determine where capacity shortages exist and, most importantly, where to add new circuits to relieve the shortages is presented. The new method of network estimation produces a feasible transmission network with near-minimum circuit miles using as input any existing network plus a load and generation schedule. An example is used to present the two steps of the method: 1) linear flow estimation and 2) new circuit selection. The method has become a fundamental part of computer programs for transmission network synthesis.

Abara (1989) formulated and solved the fleet assignment problem as an integer linear programming model, permitting assignment of two or more fleets to a flight schedule simultaneously. The objective function can take a variety of forms including profit maximization, cost minimization, and the optimal utilization of a particular fleet type. Several departments at American Airlines use the model to assist in fleet planning and schedule development. It will become one of 10 key decision modules for the next generation scheduling system currently being developed by American Airlines Decision Technologies.

Wang et al., (2010) put forward that most of the commercial P2P video streaming deployments support hundreds of channels and are referred to as multichannel systems. Measurement studies show that bandwidth resources of different channels are highly unbalanced and thus recent research studies have proposed various protocols to improve the streaming qualities for all channels by enabling cross-channel cooperation among multiple channels. However, there is no general framework for comparing existing and potential designs for multi-channel P2P systems. The goal of this paper is to establish tractable

models for answering the fundamental question in multi-channel system designs: Under what circumstances, should a particular design be used to achieve the desired streaming quality with the lowest implementation complexity? To achieve this goal, the authors first classify existing and potential designs into three categories, namely Naive Bandwidth allocation Approach (NBA), Passive Channel-aware bandwidth allocation Approach (PCA) and Active Channel-aware bandwidth allocation Approach (ACA). Then, they define the bandwidth satisfaction ratio as a performance metric to develop linear programming models for the three designs. The proposed models are independent of implementations and can be efficiently solved due to the linear property, which provides a way of numerically exploring the design space of multi-channel systems and developing closed-form solutions for special systems.

Lai et al., (2006) put forward that credit risk evaluation has been the major focus of financial and banking industry due to recent financial crises and regulatory concern of Basel II. Recent studies have revealed that emerging artificial intelligent techniques are advantageous to statistical models for credit risk evaluation. In this study, the authors discuss the use of Least Square Support Vector Machine (LSSVM) technique to design a credit risk evaluation system to discriminate good creditors from bad ones. Relative to the Vapnik's Support Vector Machine, the LSSVM can transform a quadratic programming problem into a linear programming problem thus reducing the computational complexity. For illustration, a published credit dataset for consumer credit is used to validate the effectiveness of the LSSVM.

Ferrier and Lovell, (1990) compared two techniques for estimating production economies and efficiencies. One approach involves the econometric estimation of a cost frontier; the second is a series of linear programs which calculate a production frontier. The two techniques are very different in principle, both possessing certain advantages and disadvantages. The authors' empirical implementation

of these techniques shows that they yield very similar results regarding cost economies, and dissimilar results regarding cost efficiencies. These are important findings to the extent that policy decisions and evaluation often rely on only one of the two types of approaches available.

Markowitz (1957) showed that it is common for matrices in industrial applications of linear programming to have a large proportion of zero coefficients. While every item (raw material, intermediate material, end item, equipment item) in, say, a petroleum refinery may be indirectly related to every other, any particular process uses few of these. Thus the matrix describing petroleum technology has a small percentage of non-zeros. If spacial or temporal distinctions are introduced into the model the percentage of non-zeros generally falls further.

The author also discussed a form of inverse which is especially convenient to obtain and use for matrices with a high percentage of zeros. The application of this form of inverse in linear programming is also discussed.

Manne (1960) demonstrated how a typical sequential probabilistic model may be formulated in terms of (a) an initial decision rule and (b) a Markov process, and then optimized by means of linear programming. This linear programming technique may turn out to be an efficient alternative to the functional equation approach in the numerical analysis of such problems. Regardless of computational significance, however, it is of interest that there should be such a close relationship between the two traditionally distinct areas of dynamic programming and linear programming.

Shayeghi and Bagheri (2012) demonstrated that since the emerging of distributed generation (DG) technologies, their penetration into power systems has provided new options in the design and operation of electric networks. In their paper, DG units are considered as a novel alternative for supplying the load of sub-transmission system. Thus, the mathematical model of considering DG on the expansion planning

of sub-transmission system is developed. Fix and variable costs of the plan and the related constraints are formulated in the proposed model. The proposed objective function and its constraint are converted to an optimization problem where the hybrid decimal codification genetic algorithm (DCGA) and linear programming (LP) technique are employed to solve it. Solution of the proposed method gives the optimal capacity of substations; optimal location and capacity of DGs as well as optimal configuration of the sub-transmission lines. To demonstrate the effectiveness of the proposed approach, it is applied on a realistic sub-transmission system of Zanjan Regional Electrical Company, Iran, and the results are evaluated.

Yue (2012) compared strategies and costs of protecting impatiens in greenhouse culture from western flower thrips that would provide a plant of acceptable quality to the market and would address the issue of development of resistance to commonly used pesticides by evaluating biopesticides. Partial budgets based on alternative strategies were identified. Six control strategies were identified from a combination of commercial growers, research experts and biopesticide recommendations from product distributors. The research-recommended strategy 6 had the highest total production cost (\$197.44), while one of the grower strategies based on conventional pesticides had the lowest total cost (\$153.28). The second growers' strategy had the second lowest total cost by relying on scouting and pesticide application as needed. This strategy used the smallest quantity of pesticides, and was expected to reduce or prevent resistance and minimize environmental impacts. Biopesticides had higher prices than conventional pesticides. Three biopesticide recommendation strategies (3, 4 and 5) were in the midrange of production cost. The treatments containing biopesticides usually had higher product and production cost than treatments that included only nonbiopesticides. An integer linear programming model was developed to determine the optimal WFT control program for impatiens. Constraints included pesticide mortality and label limits on consecutive or total applications per crop cycle. All pesticides in the linear

programming solution were conventional. Biopesticides were not included in the solution because mortalities of biopesticides were far below the threshold, according to research reported through the IR4 program. The costs of using pesticides include economic product costs and environmental costs. Using biopesticides to replace conventional pesticides in a rotation scheme of conventional ones with different modes of action could reduce water and soil pollution while maintaining crop quality.

Wood et al., (2013) presented a novel computer vision algorithm to analyze 3D stacks of confocal images of fluorescently stained single cells. The goal of the algorithm is to create representative *in silico* model structures that can be imported into finite element analysis software for mechanical characterization. Segmentation of cell and nucleus boundaries is accomplished via standard thresholding methods. Using novel linear programming methods, a representative actin stress fiber network is generated by computing a linear superposition of fibers having minimum discrepancy compared with an experimental 3D confocal image. Qualitative validation is performed through analysis of seven 3D confocal image stacks of adherent Vascular Smooth Muscle Cells (VSMCs) grown in 2D culture. The presented method is able to automatically generate 3D geometries of the cell's boundary, nucleus, and representative F-actin network based on standard cell microscopy data. These geometries can be used for direct importation and implementation in structural finite element models for analysis of the mechanics of a single cell to potentially speed discoveries in the fields of regenerative medicine, mechanobiology, and drug discovery.

Atias and Sharan (2013) demonstrated that large scale screening experiments have become the workhorse of molecular biology, producing data at an ever increasing scale. The interpretation of such data, particularly in the context of a protein interaction network, has the potential to shed light on the molecular pathways underlying the phenotype or process in question. A host of approaches have been

developed in recent years to tackle this reconstruction challenge. These approaches aim to infer a compact sub-network that connects the genes revealed by the screen while optimizing local (individual path lengths) or global (likelihood) aspects of the subnetwork. Yosef et al., [Yosef et al., Mol Syst Biol, 2009, 5:248] were the first to provide a joint optimization of both criteria, albeit approximate in nature. Here we devise an integer linear programming formulation of the joint optimization problem, allowing us to solve it to optimality in minutes on current networks. We apply our algorithm, iPoint, to various data sets in yeast and human and evaluate its performance against state-of-the-art algorithms. We show that iPoint attains very compact and accurate solutions that outperform previous network inference algorithms with respect to their local and global attributes, their consistency across multiple experiments targeting the same pathway, and their agreement with current biological knowledge.

Gokbayrak and Yildirim (2013) described that Wireless Mesh Networks (WMNs) provide cost effective solutions for setting up a communications network over a certain geographic area. In this paper, the authors studied strategic problems of WMNs such as selecting the gateway nodes along with several operational problems such as routing, power control, and transmission slot assignment. Under the assumptions of the physical interference model and the tree-based routing restriction for traffic flow, a Mixed Integer Linear Programming (MILP) formulation is presented, in which the objective is to maximize the minimum service level provided at the nodes. A set of valid inequalities is derived and added to the model in an attempt to improve the solution quality. Since the MILP formulation becomes computationally infeasible for larger instances, we propose a heuristic method that is aimed at solving the problem in two stages. In the first stage, we devise a simple MILP problem that is concerned only with the selection of gateway nodes. In the second stage, the MILP problem in the original formulation is solved by fixing the gateway nodes from the first stage. Computational experiments are provided to evaluate the proposed models and the heuristic method.

Wheaton (1974) reviewed the early work of Herbert-Stevens in which linear programming was used to find a competitive equilibrium to an urban land market. First, it is demonstrated that a solution to the Herbert-Stevens model does not meet well-established criteria for equilibrium. Secondly, a new linear programming model is suggested which is proven to achieve equilibrium if certain conditions are met. An iterative procedure for meeting these conditions is suggested, and an operational version of the model exhibits no problem in obtaining convergence. The revised model represents a feasible way of simulating urban land markets.

Bennett and Mangasarian (1992) proposed a single linear programming formulation which generates a plane that minimizes an average sum of misclassified points belonging to two disjoint points sets in  $n$ -dimensional real space. When the convex hulls of the two sets are also disjoint, the plane completely separates the two sets. When the convex hulls intersect, our linear program, unlike all previously proposed linear programs, is guaranteed to generate some error-minimizing plane, without the imposition of extraneous normalization constraints that inevitably fail to handle certain cases. The effectiveness of the proposed linear program has been demonstrated by successfully testing it on a number of databases. In addition, it has been used in conjunction with the multisurface method of piecewise-linear separation to train a feed-forward neural network with a single hidden layer.

Jansen and Porkolab (2003) studied the problem of scheduling a set of  $n$  independent parallel tasks on  $m$  processors, where in addition to the processing time there is a size associated with each task indicating that the task can be processed on any subset of processors of the given size. Based on a linear programming formulation, they propose an algorithm for computing a preemptive schedule with minimum makespan, and show that the running time of the algorithm depends polynomially on  $m$  and only linearly on  $n$ . Thus for any fixed  $m$ , an optimal preemptive schedule can be computed in  $O(n)$  time.

They also present extensions of this approach to other (more general) scheduling problems with malleable tasks, due dates and maximum lateness minimization.

Villasana et al., (1985) established that in long range transmission planning, where new load growth, new generation sites and perhaps a new voltage level are to be considered, a computer aided method of visualizing new circuits in a network context is needed. The new method presented meets this need by the combined use of a linear (dc) power flow transmission model and a transportation model (also known as a trans-shipment model). The dc transmission model is solved for the facilities network by obeying both of Kirchhoff's laws, flow conservation at each bus and voltage conservation around each loop. The transportation model is solved for the overloads by obeying only the bus flow conservation law while minimizing a cost objective function.

Tyteca (1997) used linear programming models to define standardised, aggregate environmental performance indicators for firms. The best practice frontier obtained corresponds to decision making units showing the best environmental behaviour. Results are obtained with data from U.S. fossil fuel-fired electric utilities, starting from four alternative models, among which are three linear programming models that differ in the way they account for undesirable outputs (pollutants) and resources used as inputs. The results indicate important discrepancies in the rankings obtained by the four models. Rather than contradictory, these results are interpreted as giving different, complementary kinds of information that should all be taken into account by public decision-makers.

Yang and Chuang (1994) demonstrated that the topology design problem is formulated as a general optimization problem and is solved by sequential linear programming. Two objectives are considered: one is to maximize the stiffness of the structure and the other is to maximize the lowest eigenvalue. A total material usage constraint is imposed for both cases. The density of each finite element is chosen as

the design variable and its relationship with Young's modulus is expressed by an empirical formula. Typically, the number of design variables is large, as the finite element mesh must be fine enough to represent the shape of the structure. To handle the large number of design variables, an efficient strategy for sensitivity analysis and optimization must be established. In this research, the adjoint variable is used for sensitivity analysis and the linear programming method is used to obtain the optimal topology. The advantage of this approach is its generality as opposed to the optimality criteria method; it can handle various problems, for example, multiple objective functions and multiple design criteria.

## 2.2 STUDIES ON LOGISTIC REGRESSION

Zhang and Yu (1998) stated in their study on logistic regression that when the incidence of an outcome of interest is common in the study population ( $>10\%$ ), the adjusted odds ratio derived from the logistic regression can no longer approximate the risk ratio. The more frequent the outcome, the more the odds ratio overestimates the risk ratio when it is more than 1 or underestimates it when it is less than. They propose a simple method to approximate a risk ratio from the adjusted odds ratio and derive an estimate of an association or treatment effect that better represents the true relative risk.

Classifying an observation into one of several populations is discriminant analysis, or classification. Relating qualitative variables to other variables through a logistic cdf functional form is logistic regression. Estimators generated for one of these problems are often used in the other. If the populations are normal with identical covariance matrices, discriminant analysis estimators are preferred to logistic regression estimators for the discriminant analysis problem. In most discriminant analysis applications, however, at least one variable is qualitative (ruling out multivariate normality). Under nonnormality, we prefer the logistic regression model with maximum likelihood estimators for solving

both problems. In this article we summarize the related arguments, and report on our own supportive empirical studies, Press and Wilson (1978).

As a first step forward in regional hazard management, multivariate statistical analysis in the form of logistic regression was used to produce a landslide susceptibility map in the Kakuda-Yahiko Mountains of Central Japan. There are different methods to prepare landslide susceptibility maps. The use of logistic regression in this study stemmed not only from the fact that this approach relaxes the strict assumptions required by other multivariate statistical methods, but also to demonstrate that it can be combined with Bivariate Statistical Analyses (BSA) to simplify the interpretation of the model obtained at the end. In susceptibility mapping, the use of logistic regression is to find the best fitting function to describe the relationship between the presence or absence of landslides (dependent variable) and a set of independent parameters such as slope angle and lithology. Here, an inventory map of 87 landslides was used to produce a dependent variable, which takes a value of 0 for the absence and 1 for the presence of slope failures. Lithology, bed rock-slope relationship, lineaments, slope gradient, aspect, elevation and road network were taken as independent parameters. The effect of each parameter on landslide occurrence was assessed from the corresponding coefficient that appears in the logistic regression function. The interpretations of the coefficients showed that road network plays a major role in determining landslide occurrence and distribution. Among the geomorphological parameters, aspect and slope gradient have a more significant contribution than elevation, although field observations showed that the latter is a good estimator of the approximate location of slope cuts. Using a predicted map of probability, the study area was classified into five categories of landslide susceptibility: extremely low, very low, low, medium and high. The medium and high susceptibility zones make up 8.87% of the total study area and involve mid-altitude slopes in the eastern part of Kakuda Mountain and the central and southern parts of Yahiko Mountain, Ayalew and Yamagishi (2005).

Swaminathan and Rogers (1990) outlined a logistic regression model for characterizing differential item functioning (DIF) between two groups is presented. A distinction is drawn between uniform and non-uniform DIF in terms of the parameters of the model. A statistic for testing the hypothesis of no DIF is developed. Through simulation studies, it is shown that the logistic regression procedure is more powerful than the Mantel-Haenszel procedure for detecting non-uniform DIF and as powerful in detecting uniform DIF.

Logistic regression analysis of two retrospective series ( $n=205$  and  $n=1667$ , respectively) of out-of-hospital cardiac arrests was performed on data sets from a Southwestern city (population, 415 000; area, 406 km<sup>2</sup>) and a Northwestern county (population, 1 038 000; area, 1399 km<sup>2</sup>). Both are served by similar two-tiered emergency response systems. All arrests were witnessed and occurred before the arrival of emergency responders, and the initial cardiac rhythm observed was ventricular fibrillation. The main outcome measure was survival to hospital discharge. Patient age, initiation of CPR by bystanders, interval from collapse to CPR, interval from collapse to defibrillation, bystander CPR/collapse-to-CPR interval interaction, and collapse-to-CPR/collapse-to-defibrillation interval interaction were significantly associated with survival. There was not a significant difference between observed survival rates at the two sites after control for significant predictors. A simplified predictive model retaining only collapse to CPR and collapse to defibrillation intervals performed comparably to the more complicated explanatory model. The effectiveness of prehospital interventions for out-of-hospital cardiac arrest may be estimated from their influence on collapse to CPR and collapse to defibrillation intervals. A model derived from combined data from two geographically distinct populations did not identify site as a predictor of survival if clinically relevant predictor variables were controlled for. This model can be generalized to other US populations and used to project the local effectiveness of interventions to improve cardiac arrest survival (Valenzuela et al., 1997).

Logistic regression analysis may well be used to develop a prognostic model for a dichotomous outcome. Especially when limited data are available, it is difficult to determine an appropriate selection of co-variables for inclusion in such models. Also, predictions may be improved by applying some sort of shrinkage in the estimation of regression coefficients. In this study we compare the performance of several selection and shrinkage methods in small data sets of patients with acute myocardial infarction, where we aim to predict 30-day mortality. Selection methods included backward stepwise selection with significance levels  $\alpha$  of 0.01, 0.05, 0.157 (the AIC criterion) or 0.50, and the use of qualitative external information on the sign of regression coefficients in the model. Estimation methods included standard maximum likelihood, the use of a linear shrinkage factor, penalized maximum likelihood, the Lasso, or quantitative external information on univariable regression coefficients. We found that stepwise selection with a low  $\alpha$  (for example, 0.05) led to a relatively poor model performance, when evaluated on independent data. Substantially better performance was obtained with full models with a limited number of important predictors, where regression coefficients were reduced with any of the shrinkage methods. Incorporation of external information for selection and estimation improved the stability and quality of the prognostic models. We therefore recommend shrinkage methods in full models including prespecified predictors and incorporation of external information, when prognostic models are constructed in small data sets (Steyerberg et al., 2000).

Collins et al., (2002) gave a unified account of boosting and logistic regression in which each learning problem is cast in terms of optimization of Bregman distances. The striking similarity of the two problems in this framework allows us to design and analyze algorithms for both simultaneously, and to easily adapt algorithms designed for one problem to the other. For both problems, they give new algorithms and explain their potential advantages over existing methods. These algorithms are iterative and can be divided into two types based on whether the parameters are updated sequentially (one at a

time) or in parallel (all at once). The authors also describe a parameterized family of algorithms that includes both a sequential- and a parallel-update algorithm as special cases, thus showing how the sequential and parallel approaches can themselves be unified. For all of the algorithms, they give convergence proofs using a general formalization of the auxiliary-function proof technique. As one of our sequential-update algorithms is equivalent to AdaBoost, this provides the first general proof of convergence for AdaBoost. The authors show that all of our algorithms generalize easily to the multiclass case, and they contrast the new algorithms with the iterative scaling algorithm. The authors conclude with a few experimental results with synthetic data that highlight the behavior of the old and newly proposed algorithms in different settings.

Cepeda et al., (2002), used Monte Carlo simulations to compare logistic regression with propensity scores in terms of bias, precision, empirical coverage probability, empirical power, and robustness when the number of events is low relative to the number of confounders. The authors simulated a cohort study and performed 252,480 trials. In the logistic regression, the bias decreased as the number of events per confounder increased. In the propensity score, the bias decreased as the strength of the association of the exposure with the outcome increased. Propensity scores produced estimates that were less biased, more robust, and more precise than the logistic regression estimates when there were seven or fewer events per confounder. The logistic regression empirical coverage probability increased as the number of events per confounder increased. The propensity score empirical coverage probability decreased after eight or more events per confounder. Overall, the propensity score exhibited more empirical power than logistic regression. Propensity scores are a good alternative to control for imbalances when there are seven or fewer events per confounder; however, empirical power could range from 35% to 60%. Logistic regression is the technique of choice when there are at least eight events per confounder.

A random vector  $\mathbf{x}$  arises from one of two multivariate normal distributions differing in mean but not covariance. A training set  $x_1, x_2, \dots, x_n$  of previous cases, along with their correct assignments, is known. These can be used to estimate Fisher's discriminant by maximum likelihood and then to assign  $\mathbf{x}$  on the basis of the estimated discriminant, a method known as the normal discrimination procedure. Logistic regression does the same thing but with the estimation of Fisher's discriminant done conditionally on the observed values of  $x_1, x_2, \dots, x_n$ . This article computes the asymptotic relative efficiency of the two procedures. Typically, logistic regression is shown to be between one half and two thirds as effective as normal discrimination for statistically interesting values of the parameters (Efron, 1975).

Logistic Regression (LR) is a widely used multivariable method for modeling dichotomous outcomes. This article examines use and reporting of LR in the medical literature by comprehensively assessing its use in a selected area of medical study. Medline, followed by bibliography searches, identified 15 peer-reviewed English-language articles with original data, employing LR, published between 1985 and 1999, pertaining to patient interest in genetic testing for cancer susceptibility. Articles were examined for each of 10 criteria for proper use and reporting of LR models. Substantial shortcomings were found in both use of LR and reporting of results. For many studies, the ratio of the number of outcome events to predictor variables (events per variable) was sufficiently small to call into question the accuracy of the regression model. Additionally, no studies reported validation analysis, regression diagnostics, or goodness-of-fit measures. It is recommended that authors, reviewers, and editors pay greater attention to guidelines concerning the use and reporting of LR models (Bagley et al., 2001).

Logistic Regression was applied to accident-related data collected from traffic police records in order to examine the contribution of several variables to accident severity. A total of 560 subjects involved in

serious accidents were sampled. Accident severity (the dependent variable) in this study is a dichotomous variable with two categories, fatal and non-fatal. Therefore, each of the subjects sampled was classified as being in either a fatal or non-fatal accident. Because of the binary nature of this dependent variable, a logistic Regression approach was found suitable. Of nine independent variables obtained from police accident reports, two were found most significantly associated with accident severity, namely, location and cause of accident. A statistical interpretation is given of the model-developed estimates in terms of the odds ratio concept. The findings show that logistic regression as used in this research is a promising tool in providing meaningful interpretations that can be used for future safety improvements in Riyadh (Al-Ghamdi, 2002).

Qin and Zhang (1996) tested the logistic regression assumption under a case-control sampling plan. After reparameterisation, the assumed logistic regression model is equivalent to a two-sample semiparametric model in which the log ratio of two density functions is linear in data. By identifying this model with a biased sampling model, they proposed a Kolmogorov-Smirnov-type statistic to test the validity of the logistic link function. Moreover, they pointed out that this test statistic can also be used in mixture sampling. They presented a bootstrap procedure along with some results on simulation and on analysis of two real datasets.

Magder and Hughes (1996) stated that in epidemiologic research, logistic regression is often used to estimate the odds of some outcome of interest as a function of predictors. However, in some datasets, the outcome of interest is measured with imperfect sensitivity and specificity. It is well known that the misclassification induced by such an imperfect diagnostic test will lead to biased estimates of the odds ratios and their variances. In this paper, the authors show that when the sensitivity and specificity of a diagnostic test are known, it is straightforward to incorporate this information into the fitting of logistic

regression models. An EM algorithm that produces unbiased estimates of the odds ratios and their variances is described. The resulting odds ratio estimates tend to be farther from the null but have greater variance than estimates found by ignoring the imperfections of the test. The method can be extended to the situation where the sensitivity and specificity differ for different study subjects, i.e., non-differential misclassification. The method is useful even when the sensitivity and specificity are not known, as a way to see the degree to which various assumptions about sensitivity and specificity affect one's estimates. The method can also be used to estimate sensitivity and specificity under certain assumptions or when a validation subsample is available. Several examples are provided to compare the results of this method with those obtained by standard logistic regression.

Logistic regression is a commonly used technique for the analysis of retrospective and prospective epidemiological and clinical studies with binary response variables. Usually this analysis is performed using large sample approximations. When the sample size is small or the data structure sparse, the accuracy of the asymptotic approximations is in question. On other occasions, singularity of the covariance matrix of parameter estimates precludes asymptotic analysis.

Under these circumstances, use of exact inferential procedures would seem to be a prudent alternative. Cox (1970) showed that exact inference on the parameters of a logistic model with binary response requires consideration of the distribution of sufficient statistics for these parameters. To date, however, resorting to the exact method has not been computationally feasible except in a few special situations. This article presents an efficient recursive algorithm that generates the joint and conditional distributions of the sufficient statistics and thus makes it feasible to perform exact inference for a much wider range of situations. Various methods of improving the efficiency of the basic algorithm, such as the application of appropriate criteria to delete infeasible vectors, recoding covariates, sorting observations by covariate values, and use of a two-step recursive procedure, are also described.

The algorithm given in this article enables the data analyst to perform exact inference for models with or without interaction terms and for matched as well as unmatched designs. Exact analysis proposed by Cox (1970) was restricted to a single parameter. Since our algorithm can be used to generate any combination of joint and conditional distributions of the sufficient statistics, it paves the way for multi-parametric exact inference. Further, this algorithm also provides a tool for comparing exact and asymptotic inferential procedures. Such comparisons would, it is hoped, provide statisticians with guidelines stating when each of the procedures should be preferred (Hirji et al., 1987).

### 2.3 SUMMARY

In this chapter we presented relevant works, studies and literature reviews on the proposed models (Linear programming and Logistic Regression Models).

In the next chapter, we shall treat theories, axioms and equations of the proposed models.



## CHAPTER 3

### METHODOLOGY

#### 3.0 INTRODUCTION

This chapter of the study reviews relevant fundamentals that will help us deduce the appropriate models and the best way to solve them.

#### 3.1 LINEAR PROGRAMING

A Linear Program (LP) is an optimization problem in which the objective function is linear in the unknowns and the constraints consist of linear equalities and linear inequalities. The exact form of these constraints may differ from one problem to another, but as shown below, any linear program can be transformed into the following standard form:

$$\begin{array}{ll} \text{minimize} & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{subject to} & \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array} \\ \text{and} & x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0, \end{array} \quad (3.1)$$

where the  $b_i$ 's,  $c_i$ 's and  $a_{ij}$ 's are fixed real constants, and the  $x_i$ 's are real numbers to be determined. We always assume that each equation has been multiplied by minus unity, if necessary, so that each  $b_i \geq 0$ .

In more compact vector notation, this standard problem becomes

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^T \mathbf{x} \\ \text{subject to} & \mathbf{Ax} = \mathbf{b} \quad \text{and} \quad \mathbf{x} \geq 0 \end{array} \quad (3.2)$$

Here  $\mathbf{x}$  is an  $n$ -dimensional column vector,  $\mathbf{c}^T$  is an  $n$ -dimensional row vector,  $\mathbf{A}$  is an  $m \times n$  matrix, and  $\mathbf{b}$  is an  $m$ -dimensional column vector. The vector inequality  $\mathbf{x} \geq 0$  means that each component of  $\mathbf{x}$  is nonnegative.

Before giving some examples of areas in which linear programming problems arise naturally, we indicate how various other forms of linear programs can be converted to the standard form.

**Example 1 (Slack variables).** Consider the problem

$$\begin{aligned} &\text{minimize} && c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ &\text{subject to} && \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &\leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &\leq b_2 \\ \vdots &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &\leq b_m \end{aligned} \\ &\text{and} && x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0, \end{aligned}$$

In this case the constraint set is determined entirely by linear inequalities. The problem may be alternatively expressed as

$$\begin{aligned} &\text{minimize} && c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ &\text{subject to} && \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + y_1 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + y_2 &= b_2 \\ \vdots &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + y_m &= b_m \end{aligned} \\ &\text{and} && x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0, \\ &\text{and} && y_1 \geq 0, y_2 \geq 0, \dots, y_m \geq 0. \end{aligned}$$

The new positive variables  $y_i$  introduced to convert the inequalities to equalities are called **slack variables** (or more loosely, slacks). By considering the problem as one having  $n + m$  unknowns  $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$ , the problem takes the standard form. The matrix  $m \times (n + m)$  that now

describes the linear equality constraints is of the special form  $[A, I]$  (that is, its columns can be partitioned into two sets; the first  $n$  columns make up the original  $A$  matrix and the last  $m$  columns make up an  $m \times m$  identity matrix).

**Example 2 (Surplus variables).** If the linear inequalities of **Example 1** are reversed so that a typical inequality is

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \geq b_i,$$

it is clear that this is equivalent to

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - y_i = b_i$$

with  $y_i \geq 0$ . Variables, such as  $y_i$ , adjoined in this fashion to convert a "greater than or equal to" inequality to equality are called **surplus variables**.

It should be clear that by suitably multiplying by minus unity, and adjoining slack and surplus variables, any set of linear inequalities can be converted to standard form if the unknown variables are restricted to be nonnegative.

**Example 3 (Free variables—first method).** If a linear program is given in standard form except that one or more of the unknown variables is not required to be nonnegative, the problem can be transformed to standard form by either of two simple techniques. To describe the first technique, suppose in (1), for example, that the restriction  $x_1 \geq 0$  is not present and hence  $x_1$  is free to take on either positive or negative values. We then write

$$x_1 = u_1 - v_1, \tag{3.3}$$

where we require  $u_1 \geq 0$  and  $v_1 \geq 0$ . If we substitute  $u_1 - v_1$  for  $x_1$  everywhere in (3.1), the linearity of the constraints is preserved and all variables are now required to be non-negative. The problem is then

expressed in terms of the  $n + 1$  variables  $u_1, v_1, x_2, x_3, \dots, x_n$ . There is obviously a certain degree of redundancy introduced by this technique, however, since a constant added to  $u_1$  and  $v_1$  does not change  $x_1$  (that is, the representation of a given value  $x_1$  is not unique). Nevertheless, this does not hinder the simplex method of solution.

**Example 4 (Free variables—second method).** A second approach for converting to standard form when  $x_1$  is unconstrained in sign is to eliminate,  $x_1$  together with one of the constraint equations. Take any one of the  $m$  equations in (3.1) which has a nonzero coefficient for  $x_1$ . Say, for example,

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i, \quad (3.4)$$

where  $a_{i1} \neq 0$ . Then  $x_1$  can be expressed as a linear combination of the other variables plus a constant. If this expression is substituted for  $x_1$  everywhere in (3.1), we are led to a new problem of exactly the same form but expressed in terms of the variables  $x_2, x_3, \dots, x_n$  only. Furthermore, the  $i^{\text{th}}$  equation, used to determine  $x_1$ , is now identically zero and it too can be eliminated. This substitution scheme is valid since any combination of non-negative variables  $x_2, x_3, \dots, x_n$  leads to a feasible  $x_1$  from (3.4), since the sign of  $x_1$  is unrestricted. As a result of this simplification, we obtain a standard linear program having  $n - 1$  variables and  $m - 1$  constraint equations. The value of the variable  $x_1$  can be determined after solution through (3.4).

**Example 5 (Specific case).** As a specific instance of the above technique consider the problem

$$\begin{aligned} &\text{minimize} && x_1 + 3x_2 + 4x_3 \\ &\text{subject to} && x_1 + 2x_2 + x_3 = 5 \\ & && 2x_1 + 3x_2 + x_3 = 6 \\ & && x_2 \geq 0, \quad x_3 \geq 0. \end{aligned}$$

Since  $x_1$  is free, we solve for it from the first constraint, obtaining

$$x_1 = 5 - 2x_2 + x_3 \quad (3.5)$$

Substituting this into the objective and the second constraint, we obtain the equivalent problem (subtracting five from the objective)

$$\begin{aligned} &\text{minimize} && x_2 + 3x_3 \\ &\text{subject to} && x_2 + x_3 = 4 \\ &&& x_2 \geq 0, \quad x_3 \geq 0. \end{aligned}$$

which is a problem in standard form. After the smaller problem is solved (the answer is  $x_1 = 4, x_3 = 0$ ) the value for  $x_1$  ( $x_1 = -3$ ) can be found from (3.5).

### 3.1.1 BASIC SOLUTIONS

Consider the system of equalities

$$\mathbf{Ax} = \mathbf{b} \tag{3.6}$$

where  $\mathbf{x}$  is an  $n$ -vector,  $\mathbf{b}$  an  $m$ -vector, and  $\mathbf{A}$  is an  $m \times n$  matrix. Suppose that from the  $n$  columns of  $\mathbf{A}$  we select a set of  $m$  linearly independent columns (such a set exists if the rank of  $\mathbf{A}$  is  $m$ ). For notational simplicity assume that we select the first  $m$  columns of  $\mathbf{A}$  and denote the  $m \times m$  matrix determined by these columns by  $\mathbf{B}$ . The matrix  $\mathbf{B}$  is then nonsingular and we may uniquely solve the equation.

$$\mathbf{Bx}_B = \mathbf{b} \tag{3.7}$$

for the  $m$ -vector  $\mathbf{x}_B$ . By putting  $\mathbf{x} = (\mathbf{x}_B, 0)$  (that is, setting the first  $m$  components of  $\mathbf{x}$  equal to those of  $\mathbf{x}_B$  and the remaining components equal to zero), we obtain a solution to  $\mathbf{Ax} = \mathbf{b}$ . This leads to the following definition.

**Definition:** Given the set of  $m$  simultaneous linear equations in  $n$  unknowns (3.6), let  $\mathbf{B}$  be any nonsingular  $m \times m$  submatrix made up of columns of  $\mathbf{A}$ . Then, if all  $n - m$  components of  $\mathbf{x}$  not associated with columns of  $\mathbf{B}$  are set equal to zero, the solution to the resulting set of equations is said to

be a **basic solution** to (3.6) with respect to the basis **B**. The components of **x** associated with columns of **B** are called **basic variables**.

In the above definition we refer to **B** as a basis, since **B** consists of  $m$  linearly independent columns that can be regarded as a basis for the space  $E^m$ . The basic solution corresponds to an expression for the vector **b** as a linear combination of these basis vectors. In general, of course, Eq. (3.6) may have no basic solutions. However, we may avoid trivialities and difficulties of a nonessential nature by making certain elementary assumptions regarding the structure of the matrix **A**. First, we usually assume that  $n > m$ , that is, the number of variables  $x_i$  exceeds the number of equality constraints. Second, we usually assume that the rows of **A** are linearly independent, corresponding to linear independence of the  $m$  equations. A linear dependency among the rows of **A** would lead either to contradictory constraints and hence no solutions to (3.6), or to a redundancy that could be eliminated. Formally, we explicitly make the following assumption in our development, unless noted otherwise.

Full rank assumption: The  $m \times n$  matrix **A** has  $m < n$ , and the  $m$  rows of **A** are linearly independent.

Under the above assumption, the system (3.6) will always have a solution and, in fact, it will always have at least one basic solution.

The basic variables in a basic solution are not necessarily all nonzero. This is noted by the following definition.

Definition: If one or more of the basic variables in a basic solution has value zero, that solution is said to be a **degenerate basic solution**.

We note that in a non-degenerate basic solution the basic variables, and hence the basis **B**, can be immediately identified from the positive components of the solution. There is ambiguity associated with

a degenerate basic solution, however, since the zero-valued basic and non-basic variables can be interchanged. So far in the discussion of basic solutions we have treated only the equality constraint (3.6) and have made no reference to positivity constraints on the variables. Similar definitions apply when these constraints are also considered. Thus, consider now the system of constraints

$$Ax = b$$

$$x \geq 0, \quad (3.8)$$

which represent the constraints of a linear program in standard form.

Definition: A vector  $x$  satisfying (3.8) is said to be feasible for these constraints. A feasible solution to the constraints (3.8) that is also basic is said to be a **basic feasible solution**; if this solution is also a degenerate basic solution, it is called a **degenerate basic feasible solution**.

### 3.1.2 THE SIMPLEX METHOD

The simplex algorithm, developed by George Dantzig in 1947, solves LP problems by constructing a feasible solution at a vertex of the polytope and then walking along a path on the edges of the polytope to vertices with non-decreasing values of the objective function until an optimum is reached. Many pivots are made with no increase in the objective function.

The simplex algorithm is quite efficient and has been proved to solve "random" problems efficiently.

To solve a standard maximization problem using the simplex method, take the following steps:

**STEP 1:** Convert to a system of equations by introducing slack variables to turn the constraints into equations, and rewriting the objective function in standard form.

**STEP 2:** Write down the initial tableau.

**STEP 3:** Select the pivot column: Choose the negative number with the largest magnitude in the bottom row (excluding the rightmost entry). Its column is the pivot column. (If there are two candidates, choose

either one.) If all the numbers in the bottom row are zero or positive (excluding the rightmost entry), then you are done: the basic solution maximizes the objective function (see below for the basic solution).

**STEP 4:** Select the pivot in the pivot column: The pivot must always be a positive number. For each positive entry  $b$  in the pivot column, compute the ratio  $a/b$ , where  $a$  is the number in the Answer column in that row. Of these test ratios, choose the smallest one. The corresponding number  $b$  is the pivot.

**STEP 5:** Use the pivot to clear the column by using Gauss Elimination, and then re-label the pivot row with the label from the pivot column. The variable originally labeling the pivot row is the departing or exiting variable and the variable labeling the column is the entering variable.

**STEP 6:** Go to Step 3.

To get the basic solution corresponding to any tableau in the simplex method, set to zero all variables that do not appear as row labels (these are the inactive variables).

The value of a variable that does appear as a row label (an active variable) is the number in the rightmost column in that row divided by the number in that row in the column labeled by the same variable.

To solve a linear programming problem with constraints of the form  $Ax + By + \dots \geq N$  with  $N$  positive, subtract a surplus variable from the left-hand side. The basic solution corresponding to the initial tableau will not be feasible since some of the active variables will be negative.

To solve a minimization problem using the simplex method, convert it into a maximization problem. If you need to minimize  $c$ , instead maximize  $p = -c$ .

The terms used in the tableau are defined as follows:

$c_j$  = Objective function coefficients for variable  $j$

$b_i$  = Right-hand side coefficients (value) for constraint  $i$

$a_{ij}$ = coefficients of variable  $j$  in constraint  $i$

$c_B$ = Objective function coefficients of the basic variables.

From Table 3.1, the top row of the table presents the  $c_j$  , the objective function coefficients.

The next row gives the headings for the various columns which are then followed by constraints coefficients. The  $Z_j$  row and the  $(c_j - Z_j)$  row which provides the current value of the objective function and the net contribution per unit of the  $j^{th}$  variable respectively are presented. The leftmost column in the tableau indicates the values of the objective function coefficients associated with the basic variable, with a set of constraints.



Table 3.1 General Form for Initial Simplex Tableau

| INITIAL SIMPLEX TABLEAU |                 |                    |             |     |                   |                 |                     |                     |          |                         |                                      |
|-------------------------|-----------------|--------------------|-------------|-----|-------------------|-----------------|---------------------|---------------------|----------|-------------------------|--------------------------------------|
|                         |                 | Decision variables |             |     |                   | Slack Variables |                     |                     |          | Solution                | Obj fn coeff                         |
| $c_j$                   |                 | $c_1$              | $c_2$       | ... | $c_n$             |                 | 0                   | 0                   | ...      | 0                       |                                      |
| $c_B$                   | Basic Variables | $x_1$              | $x_2$       | ... | $x_n$             |                 | $s_1$               | $s_2$               | ...      | $s_m$                   | Headings                             |
| 0                       | $s_1$           | $a_{11}$           | $a_{12}$    | ... | $a_{1n}$          |                 | 1                   | 0                   | ...      | 0                       |                                      |
| $\vdots$                | $s_2$           | $a_{21}$           | $a_{22}$    | ... | $a_{2n}$          |                 | 0                   | 1                   | ...      | 0                       |                                      |
| 0                       | $\vdots$        |                    | $\vdots$    |     | $\vdots$          |                 | $\vdots$            | $\vdots$            | $\vdots$ | $\vdots$                |                                      |
|                         | $s_m$           | $a_{m1}$           | $a_{m2}$    | ... | $a_{mn}$          |                 | 0                   | 0                   | ...      | 1                       |                                      |
|                         |                 | $Z_1$              | $Z_2$       | ... | $Z_{mn}$          |                 | $Z_{11}$            | $Z_{12}$            | ...      | $Z_{1m}$                | Current value of objective fn        |
|                         | $c_j - Z_j$     | $c_1 - Z_1$        | $c_2 - Z_2$ | ... | $c_{mn} - Z_{mn}$ | ...             | $c_{1_1} - Z_{1_1}$ | $c_{1_2} - Z_{1_2}$ | ...      | $c_{1_\pi} - Z_{1_\pi}$ | Reduced Cost (Net Contribution/Unit) |

## 3.2 LOGISTIC REGRESSION

### 3.2.1 SIMPLE LOGISTIC REGRESSION

Logistic regression analysis examines the influence of various factors on a dichotomous outcome by estimating the probability of the event's occurrence. It does this by examining the relationship between one or more independent variables and the log odds of the dichotomous outcome by calculating changes in the log odds of the dependent as opposed to the dependent variable itself. The log odds ratio is the ratio of two odds and it is a summary measure of the relationship between two variables. The use of the log odds ratio in logistic regression provides a more simplistic description of the probabilistic relationship of the variables and the outcome in comparison to a linear regression by which linear relationships and more rich information can be drawn.

There are two models of logistic regression to include binomial/binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical variables. Logistic regression is best used in this condition. When the dependent variable is not dichotomous and is comprised of more than two cases, a multinomial logistic regression can be employed. Also referred to as logit regression, multinomial logistic regression has very similar results to binary logistic regression.

The logistic regression or logit model takes the following form:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log \text{it}(\pi(x)) = \alpha + \beta x \quad (3.9)$$

and thus

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{(\alpha + \beta x)}}{1 + e^{(\alpha + \beta x)}} \quad (3.10)$$

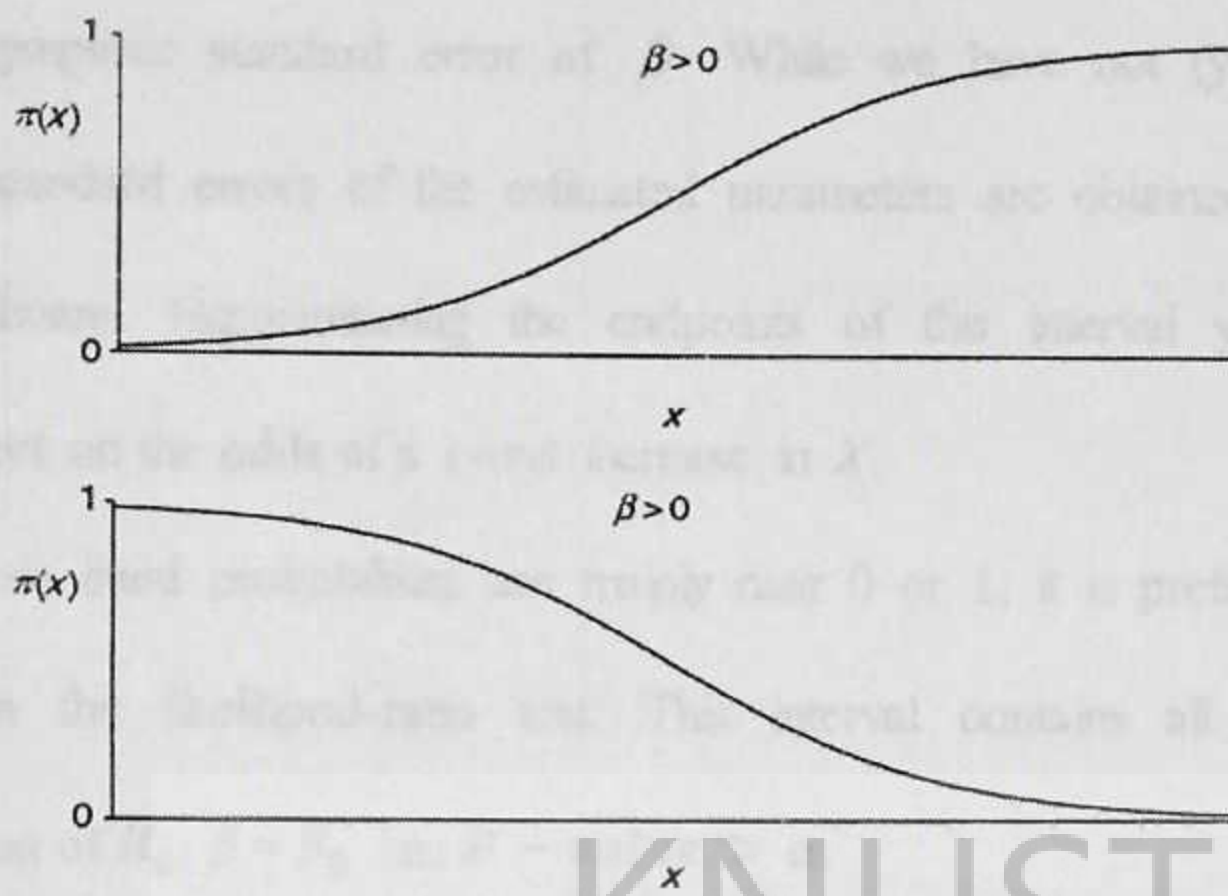


Figure 3.1: Logistic regression functions

This function represents S-shaped curves (often realistic shapes for the relationship). The random component for the (success, failure) determination is binomial and the link function the logit transformation.

The parameter  $\beta$  in the logistic regression model determines the rate of increase or decrease of the curve. When  $\beta > 0$ ,  $\pi(x)$  increases as  $x$  increases, as in Figure 3.0 (a). When  $\beta < 0$ ,  $\pi(x)$  decreases as  $x$  increases, as in Figure 3.0 (b). The magnitude of  $\beta$  determines how fast the curve increases or decreases. As  $|\beta|$  increases, the curve has a steeper rate of change. When  $\beta = 0$ , the curve flattens to a horizontal straight line.

## 3.2.2 INFERENCE FOR LOGISTIC REGRESSION

### 3.2.2.1 Confidence Intervals for Effects

A large-sample Wald confidence interval for the parameter in the logistic regression model  $\text{logit}(\pi(x)) = \alpha + \beta x$ , is

$$\hat{\beta} \pm z_{\alpha/2} (ASE) \quad (3.11)$$

with ASE the asymptotic standard error of  $\beta$ . 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. Exponentiating the endpoints of this interval yield one of the  $\exp \beta$ , the multiplicative effect on the odds of a 1-unit increase in  $X$ .

When  $n$  is small or fitted probabilities are mainly near 0 or 1, it is preferable to construct a confidence interval based on the likelihood-ratio test. This interval contains all the  $\beta_0$  values for which the likelihood-ratio test of  $H_0: \beta = \beta_0$  has  $P - value > \alpha$ .

Example: Data are available on a study with 100 participants that investigates age as a possible risk factor for heart disease. Results of fitting the Logistic regression model to the data. Software SAS was used to obtain the ML parameter estimates for the logistic regression model.

Table 3.2: ML parameter estimates

| Variable | Estimated Coefficient | Standard Error |
|----------|-----------------------|----------------|
| Constant | -5.310                | 1.134          |
| Age      | 0.111                 | 0.024          |

Problem

- (i) Write the predicted model as a function of age.
- (ii) Calculate the predicted probability at the minimum age level (20years).
- (iii) Calculate the predicted probability at the maximum age level (69years).
- (iv) Calculate the median effective level.

Solution

- (i) The predicted value for coronary heart disease as a function of Age is:

$$\hat{\pi}(x) = \frac{\exp(-5.31 + 0.111 \times AGE)}{1 + \exp(-5.31 + 0.111 \times AGE)}$$

Since  $\beta > 0$ , the predicted value is higher at higher values of AGE.

$$(ii) \hat{\pi}(x) = \frac{\exp(-5.31 + 0.111 \times 20)}{1 + \exp(-5.31 + 0.111 \times 20)} = 4.4\%$$

$$(iii) \hat{\pi}(x) = \frac{\exp(-5.31 + 0.111 \times 69)}{1 + \exp(-5.31 + 0.111 \times 69)} = 91.2\%$$

(iv) The median effective level is the age level at which the predicted probability equals 50%, which is –  
 $(-5.310)/0.111 = 47.8$

A 95% confidence interval for the effect of age is  $0.111 \pm 1.96 \times 0.024$ , or  $(0.064, 0.158)$

The confidence interval for  $\exp \beta$ , the effect on the odds per year equals  
 $(\exp 0.064, \exp 0.158) = (1.066, 1.171)$

We can thus conclude that a 1 year increase in age results in an increase of at least 6.6% and at most 17% in the odds of have a coronary heart disease.

### 3.2.2.2 Significance Testing

#### 3.2.2.2.1 Wald Test

The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter  $\beta$  to an estimate of its standard error. For the logistic regression model,  $H_0: \beta = 0$  states that the probability of success is independent of  $X$ . Wald test statistics are simple. For large samples,

$$z = \frac{\hat{\beta}}{ASE_{\hat{\beta}}} \quad (3.12)$$

has a standard normal distribution when  $\beta = 0$ . Refer  $z$  to the standard normal table to get a one-sided or two-sided  $p$ -value. Equivalently, for the two-sided  $H_a: \beta \neq 0$ ,

$$z^2 = \left( \frac{\hat{\beta}}{ASE_{\hat{\beta}}} \right)^2 \quad (3.13)$$

has a large-sample chi-squared null distribution with  $df = 1$ .

### 3.2.2.2.2 Likelihood Ratio Statistic

Although the Wald test is adequate for large samples, the likelihood-ratio test is more powerful and more reliable for sample sizes often used in practice. The test statistic compares the maximum  $L_0$  of the log-likelihood function when  $\beta = 0$  (i.e. when  $\pi(x)$  is forced to be identical at all  $x$  values) to the maximum  $L_1$  of the log-likelihood function for unrestricted  $\beta$ . The test statistic,

$$G^2 = -2(L_0 - L_1) \quad (3.14)$$

also has a large-sample chi-squared null distribution with  $df = 1$ . Most software for logistic regression reports the maximized log-likelihoods  $L_0$  and  $L_1$  and the likelihood-ratio statistic derived from those maxima.

Example: Using the heart coronary disease data

(i) Wald statistics

$z = \hat{\beta} / ASE = 0.111 / 0.024 = 4.610$ , the two-tailed  $p$ -value (based on a standard normal distribution) is 0.0001

$z^2 = \hat{\beta}^2 / ASE^2 = 0.111^2 / 0.024^2 = 21.252$ , the two-tailed  $p$ -value (based on a chi-squared distribution with  $df = 1$ ) is 0.0001

Both statistics show how a strong evidence of a positive effect of AGE on coronary heart disease.

(ii) Likelihood ratio statistic

The log-likelihood for the model containing only a constant term is  $L_0 = -68.332$ . The log-likelihood for the model containing the independent variable AGE, along with the constant term is  $L_1 = -53.667$ .

The value for the likelihood ratio test statistic is thus

$-2[-68.332 - (-53.667)] = 29.31$  the  $p$ -value (based on a chi-squared distribution with  $df = 1$ ) is 0.0001. This provides even stronger evidence than the Wald statistic of an AGE effect.

### 3.2.2.2.3 Distribution of Probability Estimates

The estimated probability that  $Y=1$  at a fixed setting  $x$  of  $X$  equals

$$\hat{\pi}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}$$

How can we construct a confidence interval for the true predicted probability? We outline it as follows:

- Start from the linear predictor  $\hat{\alpha} + \hat{\beta}x$
- The estimated linear predictor has large-sample ASE given by the estimated square root of  $Var(\hat{\alpha} + \hat{\beta}x) = Var(\hat{\alpha}) + 2xCov(\hat{\alpha}, \hat{\beta}) + x^2Var(\hat{\beta})$ .
- This can be calculated using the covariance matrix of the model parameters (reported by standard software packages).
- A 95% confidence interval for the true logit is thus

$$(\hat{\alpha} + \hat{\beta}x) \pm 1.96ASE$$

- Substituting the endpoints of this interval in the exponents of

$$\hat{\pi}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}$$

gives a corresponding interval for the predicted probability.

Example: For the coronary hearts disease example, calculate the predicted probability of having a coronary heart disease at age 67, together with the 95% confidence interval for the true probability.

Solution:

The logistic regression fit yields the following predicted probability:

$$\hat{\pi}(67) = \frac{\exp(-5.3095 + 0.1109 * 67)}{1 + \exp(-5.3095 + 0.1109 * 67)} = 0.893$$

The predicted logit is  $-5.3095 + 0.1109 * 67 = 2.121$ . The software reports the following:

$$\text{Var}(\hat{\alpha}) = 1.285, \quad \text{Var}(\hat{\beta}) = 0.000579, \quad \text{Cov}(\hat{\alpha}, \hat{\beta}) = -0.0267$$

We have used only a single explanatory variable so far, but the rest of the unit allows additional predictors. The remarks of this subsection apply regardless of the number of predictors. The standard errors of the estimates are the square roots of the variances from the main diagonal of the covariance matrix.

The estimated variance of the predicted logit equals:

$$1.285 + 2(67)(-0.0267) + 67^2(0.000579) = 0.306.$$

The 95% confidence interval for the true logit equals  $2.121 \pm 1.96 \times \sqrt{0.306}$  or  $(1.036, 3.205)$ .

This translates to the following interval for the coronary heart disease probability at age 67.

$$\left\{ \frac{\exp(1.036)}{1 + \exp(1.036)}, \frac{\exp(3.205)}{1 + \exp(3.205)} \right\} = (0.738, 0.961)$$

### 3.2.3 LOGISTIC REGRESSION FOR SEVERAL PREDICTORS

#### 3.2.3.1 Multiple Logistic Regression

Next we will consider the general logistic regression model with multiple explanatory variables. Denote the  $k$  predictors for a binary response  $Y$  by  $x_1, x_2, \dots, x_k$ . The model for the log odds is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (3.15)$$

The parameter  $\beta_i$  refers to the effect of  $x_i$  on the log odds that  $Y = 1$ , controlling the other  $x$ 's. For example,  $\exp(\beta_i)$  is the multiplicative effect on the odds of a 1-unit increase in  $x_i$ , at fixed levels of the other  $x$ 's.

Example:

To illustrate, we use the horseshoe crab data. Here, we let  $Y$  indicate whether a female crab has any satellites (other males who could mate with her). That is,  $Y = 1$  if a female crab has at least one satellite, and  $Y = 0$  if she has no satellite. We both the female crab's shell width (in cm) and color as predictors. Color has four categories: medium light, medium, medium dark, dark. Color is a surrogate for age, older crabs tending to have darker shells. To treat color as a nominal-scale predictor, we use three indicator variables for the four categories. The model is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x, \quad (3.16)$$

where  $x$  denotes width and

$c_1 = 1$  for color = medium light, 0 otherwise

$c_2 = 1$  for color = medium, 0 otherwise

$c_3 = 1$  for color = medium dark, 0 otherwise

The crab color is dark (category 4) when  $c_1 = c_2 = c_3 = 0$ .

Table 3.3 Computer Output for Model for Horseshoe Crabs with Width and Color Predictors

| Parameter | Estimate | Std. Error | Like. Ratio Confidence | 95% Limits | Chi Square | Pr > ChiSq |
|-----------|----------|------------|------------------------|------------|------------|------------|
| intercept | -12.7151 | 2.7618     | -18.4564               | -7.5788    | 21.20      | <.0001     |
| c1        | 1.3299   | 0.8525     | -0.2738                | 3.1354     | 2.43       | 0.1188     |
| c2        | 1.4023   | 0.5484     | 0.3527                 | 2.5260     | 6.54       | 0.0106     |
| c3        | 1.1061   | 0.5921     | -0.0279                | 2.3138     | 3.49       | 0.0617     |
| width     | 0.4680   | 0.1055     | 0.2713                 | 0.6870     | 19.66      | <.0001     |

For instance, for dark crabs,  $c_1 = c_2 = c_3 = 0$ , and the prediction equation is

$$\text{logit}[P^{\wedge}(Y = 1)] = -12.715 + 0.468x.$$

By contrast, for medium-light crabs,  $c_1 = 1$ , and

$$\text{logit}[P^{\wedge}(Y = 1)] = (-12.715 + 1.330) + 0.468x = -11.385 + 0.468x$$

The model assumes a lack of interaction between color and width. Width has the same effect (coefficient 0.468) for all colors. This implies that the shapes of the four curves relating width to  $P(Y = 1)$  (for the four colors) are identical. For each color, a 1 cm increase in width has a multiplicative effect of  $\exp(0.468) = 1.60$  on the odds that  $Y = 1$ .

The parallelism of curves in the horizontal dimension implies that two curves never cross. At all width values, for example, color 4 (dark) has a lower estimated probability of a satellite than the other colors.

To illustrate, a dark crab of average width (26.3 cm) has estimated probability

$$\exp[-12.715 + 0.468(26.3)] / \{1 + \exp[-12.715 + 0.468(26.3)]\} = 0.399.$$

By contrast, a medium-light crab of average width has estimated probability

$$\exp[-11.385 + 0.468(26.3)] / \{1 + \exp[-11.385 + 0.468(26.3)]\} = 0.715.$$

The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors. For example, the difference in color parameter estimates between medium-light crabs and dark crabs equals 1.330. So, at any given width, the estimated odds that a medium-light crab has a satellite are  $\exp(1.330) = 3.8$  times the estimated odds for a dark crab. Using the probabilities just calculated at width 26.3, the odds equal  $0.399/0.601 = 0.66$  for a dark crab and  $0.715/0.285 = 2.51$  for a medium-light crab, for which  $2.51/0.66 = 3.8$

Figure 3.1 below displays the fitted model. Any one curve is any other curve shifted to the right or to the left.

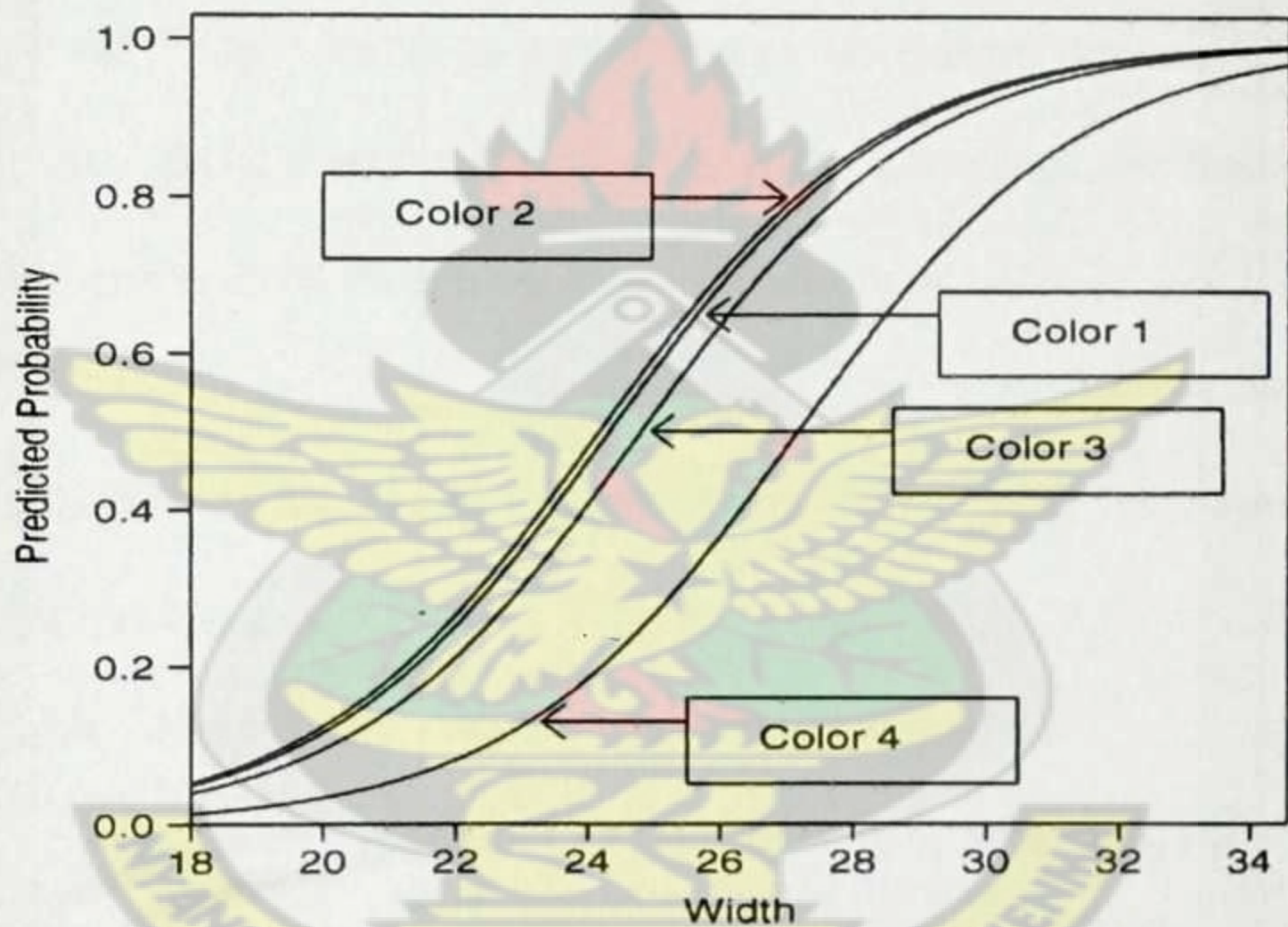


Figure 3.2: Logistic regression model with width and colors as predictors

The models we have considered so far assume a lack of interaction between width and color. Let us check now whether this is sensible. We can allow interaction by adding cross products of terms for width and color. Each color then has a different-shaped curve relating width to the probability of a satellite, so a comparison of two colors varies according to the value of width.

For example, consider the model just discussed that has a dummy variable  $c = 0$  for dark-colored crabs and  $c = 1$  otherwise. The model with an interaction term has the prediction equation;

$$\text{logit}[P^{\wedge}(Y = 1)] = -5.854 - 6.958c + 0.200x + 0.322(c \times x)$$

### 3.3 SUMMARY

In this chapter we treated linear programming and logistic regression models, we considered their theories and axioms and applied them to some examples.

In the next chapter, we shall present the data collection and models.



## CHAPTER 4

### DATA COLLECTION AND MODELING

#### 4.0 INTRODUCTION

In this chapter we shall analyze the data taken from Prudential Bank Limited (PBL) using their banking application software named Flexcube, the models (linear Programing and Logistic Regression) are formulated and solved to help the Bank maximize its profits and minimize the number of defaulters.

#### 4.1 THE LINEAR PROGRAMING PROBLEM

PBL has decided that it will increase its credit portfolio from the existing value of GH¢476 million by GH¢166 million for the year 2013. Therefore the Bank is in the process of formulating a credit policy involving GH¢166 million. Being a full-service facility, the bank is obligated to grant credit facilities to different clientele.

Table 4.0 provides the type of loans, the interest rate charged by the bank and the probability of bad debt as estimated from past experience.

Table 4.1: Economic sectors where PBL supports with credit facilities

| Industry Sector              | Interest Rate | Probability of bad Debt | Recovery Rate |
|------------------------------|---------------|-------------------------|---------------|
| Commerce                     | 0.30          | 0.08                    | 0.92          |
| Agriculture                  | 0.28          | 0.10                    | 0.90          |
| Education                    | 0.30          | 0.06                    | 0.94          |
| Construction (Manufacturing) | 0.32          | 0.09                    | 0.91          |
| Consumer                     | 0.36          | 0.08                    | 0.92          |
| Export                       | 0.28          | 0.06                    | 0.94          |
| Finance                      | 0.28          | 0.04                    | 0.96          |

Bad debts are assumed unrecoverable and hence produce no interest revenue. For policy reasons, there are limits on how the bank allocates its funds. Competition with other financial institutions in the city requires that the bank.

- (i) Allocate at least 60% of total funds available to Commerce, Export, Education and Construction sectors of the economy.
- (ii) The sum of funds allocated for Consumer and Financial sectors must not exceed 25% of total funds available.
- (iii) To support the exportation and as part of PBL's corporate mandate, funds allocated to export and Agric sectors must at least be 50% more than funds allocated to Consumer, Finance and Education sectors.
- (iv) The sum of funds allocated to Commerce and Construction must at least be greater than 70% that allocated to Finance, Consumer, Education and Agric.
- (v) To assist the Agricultural sector, funds allocated to Agric sector must at least be 20% of funds allocated to Construction and Consumer sectors.
- (vi) The Bank also states that the total ratio for bad debt on all credit facilities must not exceed 0.07

#### 4.1.1 THE PROPOSED LP MODEL

The variables of the model are defined as follows;

$$x_1 = \text{Commerce}, \quad x_2 = \text{Agriculture}, \quad x_3 = \text{Education},$$

$$x_4 = \text{Construction(Manufacturing)}, \quad x_5 = \text{Consumer}, \quad x_6 = \text{Export}, \quad x_7 = \text{Finance}$$

The objective of PBL is to maximize its net returns comprised of the difference between the revenue from interest and lost funds due to bad debts.

Therefore the Objective function is:

$$Z_{max} = 0.30(0.92)x_1 + 0.28(0.90)x_2 + 0.30(0.94)x_3 + 0.32(0.91)x_4 + 0.36(0.92)x_5 + 0.28(0.94)x_6 + 0.28(0.96)x_7$$

The function simplifies to

$$Z_{max} = 0.1960x_1 + 0.1520x_2 + 0.2194x_3 + 0.2012x_4 + 0.2512x_5 + 0.2032x_6 + 0.2288x_7$$

This function is subject to the following seven constraints:

(i) Limit on total funds available

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \leq 166$$

(ii) Limit on Commerce, Export, Education and Construction(Manufacturing) sectors

$$x_1 + x_3 + x_4 + x_6 \geq 0.6(166) \text{ which simplifies to}$$

$$x_1 + x_3 + x_4 + x_6 \geq 99.60$$

(iii) Limit on Consumer and Finance sectors

$$x_5 + x_7 \leq 0.25(166) \text{ which simplifies to}$$

$$x_5 + x_7 \leq 41.50$$

(iv) Limit on Agriculture and Export sectors

$$x_2 + x_6 \geq 0.5(x_3 + x_5 + x_7) \text{ which simplifies to}$$

$$x_2 - 0.5x_3 - 0.5x_5 + x_6 - 0.5x_7 \geq 0$$

(v) Limit on Commerce and Construction(Manufacturing)

$$x_1 + x_4 \geq 0.7(x_2 + x_3 + x_5 + x_7) \text{ which simplifies to}$$

$$x_1 - 0.7x_2 - 0.7x_3 + x_4 - 0.7x_5 - 0.7x_7 \geq 0$$

(vi) Limit on Agriculture

$$x_2 \geq 0.2(x_4 + x_5) \text{ which simplifies to}$$

$$x_2 - 0.2x_4 - 0.2x_5 \geq 0$$

(vii) Limit on bad debt

$$\frac{0.08x_1+0.10x_2+0.06x_3+0.09x_4+0.08x_5+0.06x_6+0.04x_7}{x_1+x_2+x_3+x_4+x_5+x_6+x_7} \leq 0.07$$

which simplifies to

$$0.01x_1 + 0.03x_2 - 0.01x_3 + 0.02x_4 + 0.01x_5 - 0.01x_6 - 0.03x_7 \leq 0$$

and

(viii) Non-negativity

$$x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, x_5 \geq 0, x_6 \geq 0, x_7 \geq 0$$

The allocation of funds to the various sectors as formulated above was transferred onto Microsoft excel.

The solver option as an add-in in excel was used to solve the LP problem.

The solution to the LP problem is summarized in Table 4.1 below;

Table 4.2: Proposed allocation to maximize return on GH¢166 million

| Allocation                   | Final Value | Reduced Cost | Objective Coefficient |
|------------------------------|-------------|--------------|-----------------------|
| Amt to allocate Commerce     | 55.4369065  | 0            | 0.1960                |
| Amt to allocate Agriculture  | 5.486851717 | 0            | 0.1520                |
| Amt to allocate Education    | 32.208729   | 0            | 0.2194                |
| Amt to allocate Construction | 0           | -0.010907451 | 0.2012                |
| Amt to allocate Consumer     | 27.43425858 | 0            | 0.2512                |
| Amt to allocate Export       | 31.36751278 | 0            | 0.2032                |
| Amt to allocate Finance      | 14.06574142 | 0            | 0.2288                |

This optimal allocation yields a revenue (return) of approximately GH¢35.25 million.

The entire table and copy of the Microsoft excel sheets are attached as appendix I.

4.2 THE DISCRIMINATING MODEL (Binary Logistic Regression Model)

4.2.1 LOGISTIC REGRESSION ANALYSIS

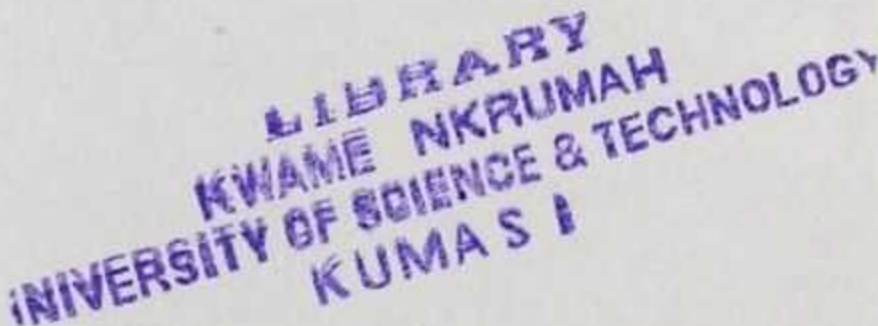
This study uses information on 850 past and prospective customers to execute a Logistic Regression Analysis. Of these, 700 cases are customers who were previously given loans. We use a random sample of 489 of these 700 customers to create a risk model. We then set aside the remaining 211 customers as a holdout or validation sample on which to test the credit-risk model then use the model to classify the 150 prospective customers as good or bad credit risks. Binary logistic regression is an appropriate technique to use on these data because the “dependent” or criterion variable is dichotomous (loan default vs. no default).

First we display the crosstabulations below, which confirm our sample characteristics. The first table shows that we will be using 700 cases for building and validating our model, holding the 150 prospects aside for later scoring using the model’s coefficients.

Table 4.3: Case Processing Summary of Validate versus Previously defaulted

|                                 | Cases |         |         |         |       |         |
|---------------------------------|-------|---------|---------|---------|-------|---------|
|                                 | Valid |         | Missing |         | Total |         |
|                                 | N     | Percent | N       | Percent | N     | Percent |
| Validate * Previously defaulted | 700   | 82.4%   | 150     | 17.6%   | 850   | 100.0%  |

The second table shows that we have created a variable called “validate.” Customers have been randomly assigned one of two values of this variable. The 489 customers who will be used to build the model are assigned a value of 1. The remaining 211 customers will be assigned a value of zero, and will constitute the validation sample on which the model will be tested.



The crosstabulations also show that the modeling sample contains 360 customers who did not default on a previous loan, and 129 who did default. The validation or holdout sample contains 157 customers who did not default, and 54 who did.

Table 4.4: Crosstabulation of Validate and Previously defaulted counts

|                      |     |                               | validate |        | Total  |
|----------------------|-----|-------------------------------|----------|--------|--------|
|                      |     |                               | 0        | 1      |        |
| Previously defaulted | No  | Count                         | 157      | 360    | 517    |
|                      |     | % within Previously defaulted | 30.4%    | 69.6%  | 100.0% |
|                      |     | % within validate             | 74.4%    | 73.6%  | 73.9%  |
|                      | Yes | Count                         | 54       | 129    | 183    |
|                      |     | % within Previously defaulted | 29.5%    | 70.5%  | 100.0% |
|                      |     | % within validate             | 25.6%    | 26.4%  | 26.1%  |
| Total                |     | Count                         | 211      | 489    | 700    |
|                      |     | % within Previously defaulted | 30.1%    | 69.9%  | 100.0% |
|                      |     | % within validate             | 100.0%   | 100.0% | 100.0% |

### 4.2.2 THE BINARY LOGISTIC REGRESSION MODEL

Now we run a logistic regression modeling analysis and examine the results. The model will be testing several candidate predictors, including:

- (i) Age
- (ii) Level of education
- (iii) Number of years with current employer
- (iv) Number of years at current address
- (iv) Household income (in thousands of GH¢)
- (v) Debt-to-income ratio
- (vi) Number of dependents
- (vii) Other debt (in thousands of GH¢)

Our logistic regression modeling analysis will use an automatic stepwise procedure using Wald test as a decision statistic. This begins by selecting the strongest candidate predictor, then testing additional candidate predictors, one at a time, for inclusion in the model. At each step, we check to see whether a new candidate predictor will improve the model significantly. We also check to see whether, if the new predictor is included in the model, any other predictors already in the model should stay or be removed. If a newly entered predictor does a better job of explaining loan default behavior, then it is possible for a predictor already in the model to be removed from the model because it no longer uniquely explains enough. This stepwise procedure continues until all the candidate predictors have been thoroughly tested for inclusion and removal. When the analysis is finished, we have the following table that contains various statistics.

Table 4.5: Variables in the equation

|                     | Variable | B      | S.E.  | Wald    | df | Sig.  | Exp(B) |
|---------------------|----------|--------|-------|---------|----|-------|--------|
| Step 1 <sup>a</sup> | debtinc  | 0.129  | 0.016 | 61.777  | 1  | 0.000 | 1.138  |
|                     | Constant | -2.500 | 0.228 | 119.948 | 1  | 0.000 | 0.082  |
| Step 2 <sup>b</sup> | employ   | -0.131 | 0.022 | 34.850  | 1  | 0.000 | 0.877  |
|                     | debtinc  | 0.140  | 0.018 | 61.974  | 1  | 0.000 | 1.150  |
|                     | Constant | -1.695 | 0.258 | 43.051  | 1  | 0.000 | 0.184  |
| Step 3 <sup>c</sup> | employ   | -0.194 | 0.029 | 44.691  | 1  | 0.000 | 0.824  |
|                     | income   | 0.017  | 0.005 | 12.880  | 1  | 0.000 | 1.017  |
|                     | debtinc  | 0.147  | 0.018 | 63.826  | 1  | 0.000 | 1.159  |
|                     | Constant | -2.064 | 0.285 | 52.349  | 1  | 0.000 | 0.127  |
| Step 4 <sup>d</sup> | employ   | -0.194 | 0.030 | 43.110  | 1  | 0.000 | 0.823  |
|                     | address  | -0.060 | 0.022 | 7.770   | 1  | 0.005 | 0.942  |
|                     | income   | 0.021  | 0.005 | 16.810  | 1  | 0.000 | 1.021  |
|                     | debtinc  | 0.152  | 0.019 | 64.793  | 1  | 0.000 | 1.164  |
|                     | Constant | -1.822 | 0.296 | 37.781  | 1  | 0.000 | 0.162  |

- a. Variable(s) entered on step 1: debtinc (Debt to income ratio).
- b. Variable(s) entered on step 2: employ (Years with current employer).

- c. Variable(s) entered on step 3: income (Household income).
- d. Variable(s) entered on step 4: address (Years at current address)

Table 4.4 leftmost column shows that the stepwise model-building process included four steps. In the first step, a constant as well as the debt to income ratio predictor variable “debtinc” are entered into the model. At the second step the years with current employer “employ” is added to the model. The third step adds household income “income”. And the final step adds number of years at current address “address”.

The “B” column shows the Beta Coefficients, abbreviated with a “B” associated with each predictor. We see that number of years at current employer and number of years at current address have negative coefficients, indicating that customers who have spent less time at their current employer or current address are somewhat more likely to default on a loan. The predictors measuring the household income and debt-to-income ratio both have positive coefficients, indicating that higher household income or higher debt-to-income ratios are associated with a greater likelihood of defaulting on a loan.

Exp(B) represents the ratio-change in the odds of the event of interest for a one-unit change in the predictor. For example, Exp(B) for number of years with current employer is equal to 0.823, which means that the odds of default for a person who has been employed at their current job for two years are just 0.823 times the odds of default for a person who has been employed at their current job for 1 year, all other things being equal. The odds of default for a person with 1 more year on the job are  $1 \times 0.823 = 0.823$ , so the corresponding probability of default is 0.4515 using equation (3.10).

### 4.2.3 MODEL DIAGNOSTICS

The Binary Logistic Regression procedure reports the Hosmer-Lemeshow goodness-of-fit statistic.

Table 4.6: Hosmer and Lemeshow Test

| Step | Chi-square | df | Sig. |
|------|------------|----|------|
| 1    | 7.567      | 8  | .477 |
| 2    | 5.341      | 8  | .721 |
| 3    | 6.188      | 8  | .626 |
| 4    | 8.193      | 8  | .415 |

The Hosmer-Lemeshow statistic indicates a poor fit if the significance value is less than 0.05. Here, it reports a significance level of 0.415; the model adequately fits the data.

KNUST

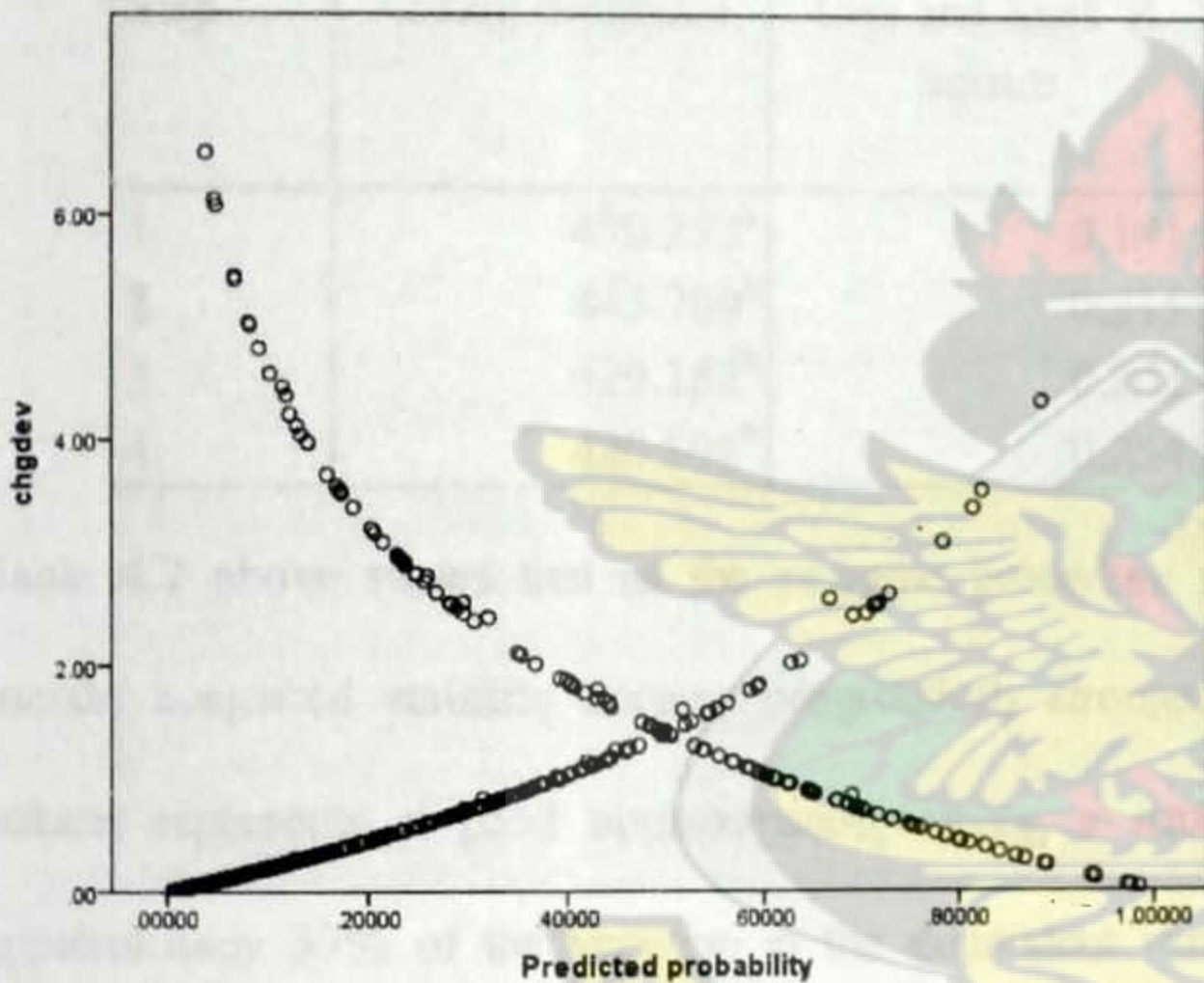


Figure 4.1: Change in Deviance versus Predicted Probabilities

In Figure 4.1 above “chgdev” is squared studentized residuals calculated from the data. The change in deviance plot helps us to identify cases that are poorly fit by the model. Larger changes in deviance indicate poorer fits.

There are two distinct patterns in the plot: a curve that extends from the lower left to the upper right, and a curve that extends from the upper left to the lower right. The curve that extends from the lower left to the upper right corresponds to cases in which the dependent variable has a value of 0. Thus, non-

defaulters who have large model-predicted probabilities of default are adequately fit by the model. The curve that extends from the upper left to the lower right corresponds to cases in which the dependent variable has a value of 1. Thus, defaulters who have small model-predicted probabilities of default are also fairly fit by the model.

By identifying the cases that are poorly fit by the model, we can focus on how those customers are different, and hopefully discover another predictor that will improve the model.

Table 4.7: Pseudo R-Squared Statistics.

| Step | -2 Log likelihood    | Cox and Snell R Square | Nagelkerke R Square |
|------|----------------------|------------------------|---------------------|
| 1    | 490.252 <sup>a</sup> | 0.141                  | 0.205               |
| 2    | 445.709 <sup>b</sup> | 0.215                  | 0.315               |
| 3    | 429.182 <sup>b</sup> | 0.241                  | 0.353               |
| 4    | 420.898 <sup>b</sup> | 0.254                  | 0.371               |

Table 4.7 above shows that as the stepwise procedure moved forward from step one to step four, the pseudo r-squared statistics became progressively stronger. The Nagelkerke statistic in the far righthand column represents a good approximation, having a maximum possible value of 1.00. It shows that approximately 37% of the variation in the dependent variable is explained by the four predictors in our final model. This suggests that the model adequately fits the data.

Figure 4.2 below provides additional information about the model's strength. It shows probability distributions for the probability of defaulting, separately for actual non-defaulters and actual defaulters. The binary logistic regression model assigns probabilities of defaulting to each customer, ranging from zero to 1.00 (zero to 100%). In this case, it uses a cut point of exactly 0.50 (50% probability) as the dividing line between predicted non-defaulters and predicted defaulters.

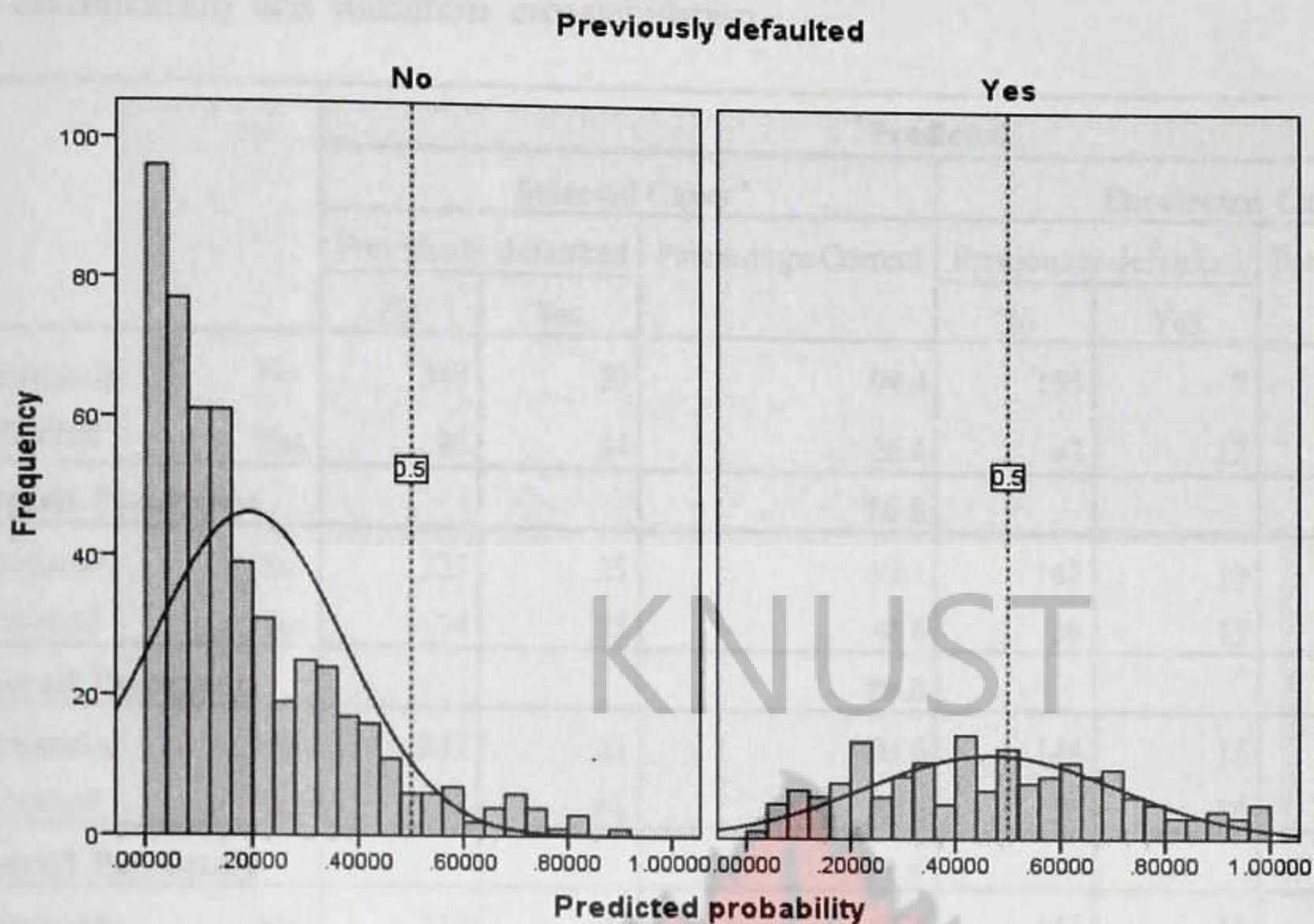


Figure 4.2: Frequency distribution of Predicted Probability of defaulters by Actual observed defaulters status (No/Yes).

The leftmost graph shows that the modeling process assigned the bulk of the actual non-defaulters very low probabilities of defaulting, far below the 50% probability cut point. And the right-hand graph shows that the model assigned the bulk of the defaulters low probabilities of defaulting, a little above the 50% cut point. So this adds more confirmation that we have a fairly good model.

The expectation for the right-hand graph is the model assigns the bulk of the defaulters high probabilities of defaulting, far above the 50% cut point.

#### 4.2.4 CLASSIFICATION AND VALIDATION OF THE MODEL

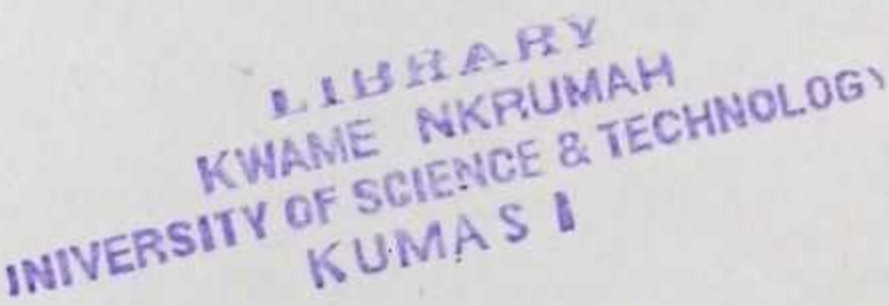
Crosstabulating observed response categories with predicted categories helps us to determine how well the model identifies defaulters.

Table 4.8: classification and validation crosstabulation

| Observed           |                      |     | Predicted                   |     |                    |                                 |     |                    |
|--------------------|----------------------|-----|-----------------------------|-----|--------------------|---------------------------------|-----|--------------------|
|                    |                      |     | Selected Cases <sup>b</sup> |     |                    | Unselected Cases <sup>c,d</sup> |     |                    |
|                    |                      |     | Previously defaulted        |     | Percentage Correct | Previously defaulted            |     | Percentage Correct |
|                    |                      |     | No                          | Yes |                    | No                              | Yes |                    |
| Step 1             | Previously defaulted | No  | 340                         | 20  | 94.4               | 150                             | 7   | 95.5               |
|                    |                      | Yes | 95                          | 34  | 26.4               | 42                              | 12  | 22.2               |
| Overall Percentage |                      |     |                             |     | 76.5               |                                 |     | 76.8               |
| Step 2             | Previously defaulted | No  | 335                         | 25  | 93.1               | 147                             | 10  | 93.6               |
|                    |                      | Yes | 74                          | 55  | 42.6               | 39                              | 15  | 27.8               |
| Overall Percentage |                      |     |                             |     | 79.8               |                                 |     | 76.8               |
| Step 3             | Previously defaulted | No  | 337                         | 23  | 93.6               | 144                             | 13  | 91.7               |
|                    |                      | Yes | 68                          | 61  | 47.3               | 36                              | 18  | 33.3               |
| Overall Percentage |                      |     |                             |     | 81.4               |                                 |     | 76.8               |
| Step 4             | Previously defaulted | No  | 337                         | 23  | 93.6               | 143                             | 14  | 91.1               |
|                    |                      | Yes | 69                          | 60  | 46.5               | 35                              | 19  | 35.2               |
| Overall Percentage |                      |     |                             |     | 81.2               |                                 |     | 76.8               |

Table 4.8 shows that the model correctly classified 93.6% of the modeling sample’s non-defaulters and about 46.5% of the modeling sample’s defaulters, for an overall correct classification percentage of about 81.2%. Similarly, when applied to the holdout or validation sample, the model correctly identified about 91.1% of the non-defaulters and about 35.2% of the defaulters, for an overall correct classification percentage of about 76.8%. The classifications above also confirm that the model is an adequate model; though it classifies non-defaulter with a high degree of accuracy same cannot be said for classification of defaulters.

Now that we have a predictive model, we use it to score the prospect file. The graph below shows the result after we have scored our 150 prospects.



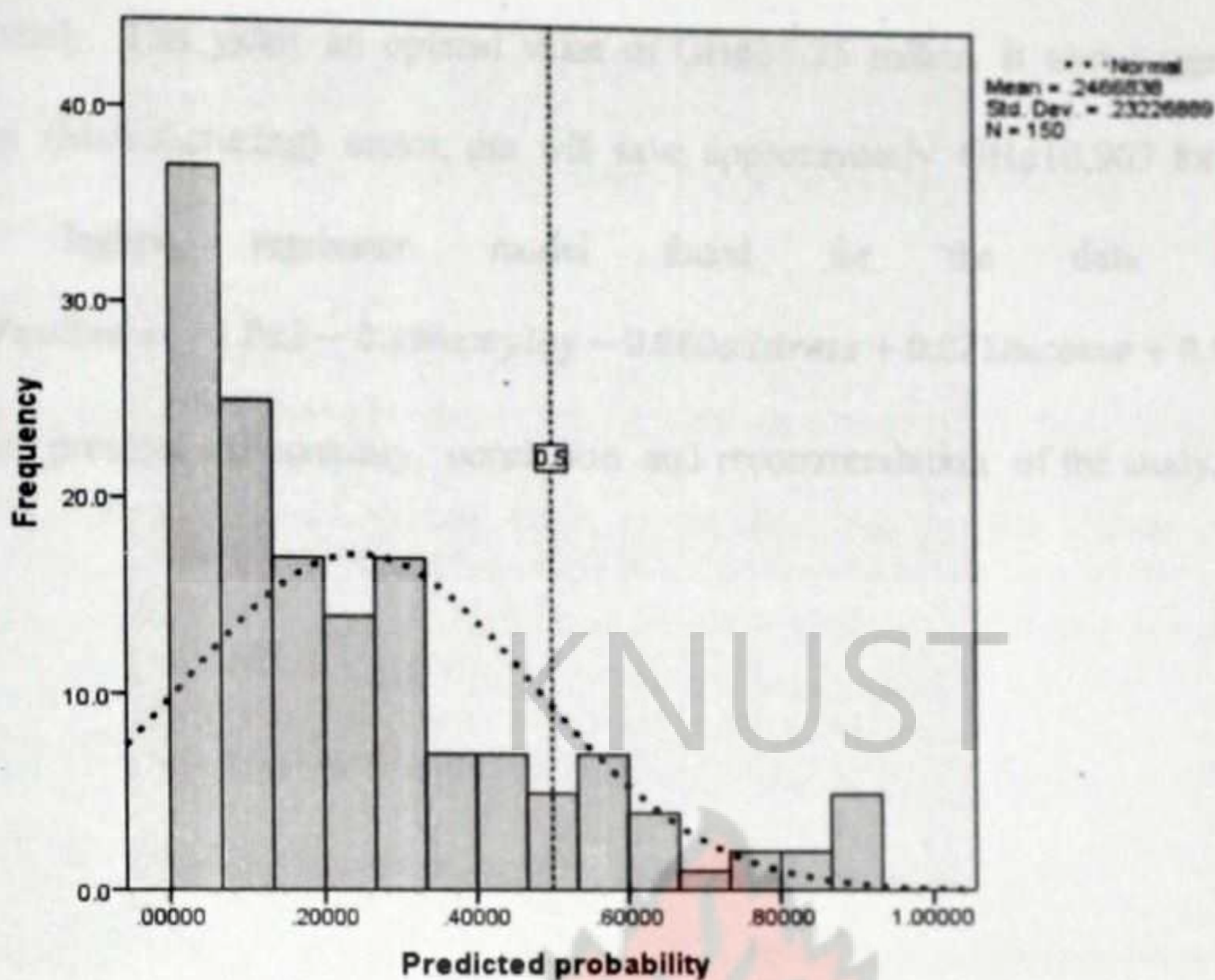


Figure 4.3: Distribution of Predicted Probabilities of Loan Default among Prospects

Figure 4.3 shows that approximately 82% of the prospects would not be expected to default on a loan. (If we had used much larger customer and prospect samples, as would typically be the case if PBL had a larger customer base, then the prospect sample's results would more closely resemble the modeling sample's results.) Note that the separation of prospects into predicted defaulter and non-defaulter subgroups is not quite as clean as for the modeling sample. Although larger samples would mitigate this difference, it is typical for a model to deteriorate slightly when applied to a sample that is different from the one on which the model was built, due to natural sampling error.

### 4.3 SUMMARY

In this chapter data from PBL was analyzed and the following findings were made;

The LP equations solved advices that the allocation of GH¢166 million should be approximately 55, 6, 32, 27, 31, 14 (all in millions of GH¢) to Commerce, Agriculture, Education, Consumer, Export and

Finance respectively. This yields an optimal value of GH¢35.25 million. It also suggest that PBL ignores the Construction (Manufacturing) sector, this will save approximately GH¢10,907 for the Bank.

The binary logistic regression model found for the data from PBL is;  
 $previously\ defaulted = -1.822 - 0.194employ - 0.060address + 0.021income + 0.152debtinc$

The next chapter presents the summary, conclusion and recommendation of the study.



## CHAPTER 5

### CONCLUSIONS AND RECOMMENDATIONS

#### 5.0 INTRODUCTION

This chapter presents the summary, conclusions drawn from the study and recommendations to help Prudential Bank Limited optimally allocate funds to maximize returns and reduce default rate by using the binary logistic regression model to discriminate between defaulters and non-defaulters.

#### 5.1 SUMMARY OF RESULTS

##### 5.1.1 OPTIMAL ALLOCATION

The allocation problem went through nine iterations before an optimal solution was found. The optimal solution gives the optimal Value (Return/Profit) of GH¢35.25million which occurs at

$$x_1(\text{Commerce}) = \text{GH¢}55.44 \text{ million}$$

$$x_2(\text{Agric}) = \text{GH¢}5.49 \text{ million}$$

$$x_3(\text{Education}) = \text{GH¢}32.21 \text{ million}$$

$$x_5(\text{Consumer}) = \text{GH¢}27.43 \text{ million}$$

$$x_6(\text{Export}) = \text{GH¢}31.37 \text{ million}$$

$$x_7(\text{Finance}) = \text{GH¢}14.10 \text{ million}$$

From the results, the value of  $x_4(\text{Construction/Manufacturing})$  is zero at optimum because the probability of bad debt is high (refer Table 4.1). Though it's the second highest in the table, Agriculture had an optimal value because of the special policy on Agriculture. Refer to appendix III for tables showing the solution list, ranging values, linear programming results and the iterations as displayed by Solver for Microsoft Excel.

### 5.1.2 THE BINARY LOGISTIC REGRESSION MODEL

The regression model for the data from PBL was found to be

*previously defaulted*

$$= -1.822 - 0.194\text{employ} - 0.060\text{address} + 0.021\text{income} + 0.152\text{debtinc}$$

We see that number of years at current employer and address have negative coefficients, indicating that customers who have spent less time at their current employer or address are somewhat more likely to default on a loan. The predictors measuring the household income and debt-to-income ratio both have positive coefficients, indicating that higher household income or higher debt-to-income ratios are associated with a greater likelihood of defaulting on a loan.

A critical issue for the credit officers in the Bank is the cost of Type I and Type II errors. That is, what is the cost of classifying a defaulter as a non-defaulter (Type I error) and what is the cost of classifying a non-defaulter as a defaulter (Type II error).

If bad debt is the primary concern, then the Bank wants to lower the Type I error and maximize the "sensitivity". (Sensitivity is the probability that a "positive" case [a defaulter] is correctly classified.) If growing the Bank's customer base is the priority, then the Bank wants to lower the Type II error and maximize the "specificity". (Specificity is the probability that a "negative" case [a non-defaulter] is correctly classified.)

Usually both are major concerns, so the Bank has to choose a decision rule for classifying customers that gives the best mix of sensitivity and specificity. In this study we arbitrarily chose a probability cut point of 0.50 (50%). But in practice, depending on the Bank's specific objectives, we may want to experiment with various cut points to see how these affect our models' sensitivity and specificity by examining the rates of correct classification for each model.

## 5.2 CONCLUSION

We began this study stating the fact that most Banks in Ghana do not use mathematical methods to assist them in decision making typically in the allocation of limited funds for credit facility disbursement.

As a result of this most banks are unable to optimize their profit margins. The LP model proposed for Prudential Bank Limited will assist the Bank to disburse its available funds for credit facilities more effectively and profitably.

We have also demonstrated the use of risk modeling using logistic regression analysis to identify demographic and behavioral characteristics associated with likelihood to default on a credit facility. We identified four important influences, and we confirmed the validity of the model using several diagnostic analytic procedures. We also used the results of the model to score a prospect sample, and we briefly discussed the importance of examining a model's sensitivity and specificity in the context of a bank's specific, real-world objectives.

## 5.3 RECOMMENDATIONS

From the study we realize that using scientific methods to allocate and disburse credit facilities helps the banks increase returns on funds. Hence we recommend that PBL adapt the LP model of this study in their allocation of funds reserved for credit facilities. The LP model is created in excel in such a way that the data entry table is linked to the solution table on the same worksheet so any change in the data will reflect immediately in the solution. This makes it easy for PBL to allocate their funds whenever they want to. The worksheet also allows the deletion or addition of constraints and it shows the percentage returns to be made on available funds.

We also recommend that Banks and other financial institutions be educated to employ scientific methods such as the use of mathematical models to assist them in their decision making processes.

Finally, with respect to the BLRM proposed, we recommend that a research team not less than two members do further research on the “sensitivity” and “specificity” of the model to meet the objectives of the financial institution.

KNUST



## REFERENCES

1. Abara, J. (1989). Applying integer linear programming to the fleet assignment problem. *Interfaces*, 19(4), 20-28.
2. Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6), 729-741.
3. Amponsah, S. K. (2009). Optimization Techniques 1. IDL KNUST.
4. Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *The journal of wildlife management*, 912-923.
5. Atias, N., and Sharan, R. (2013). iPoint: an integer programming based algorithm for inferring protein subnetworks. *Molecular BioSystems*.
6. Ayalew, L., and Yamagishi, H. (2005). The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology*, 65(1), 15-31.
7. Bakhsh, K. H. U. D. A., Hassan, I., and Maqbool, A. (2005). Factors affecting cotton yield: a case study of Sargodha (Pakistan). *Journal of Agriculture and Social Sciences*, 1(4), 322-334.
8. Bennett, K. P., and Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1), 23-34.
9. Betters, D. R. (1988). Planning optimal economic strategies for agroforestry systems. *Agroforestry systems*, 7(1), 17-31.
10. Boersma, E., Pieper, K. S., Steyerberg, E. W., Wilcox, R. G., Chang, W. C., Lee, K. L., and Simoons, M. L. (2000). Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation: results from an international trial of 9461 patients. *Circulation*, 101(22), 2557-2567.

11. Boyd, C. R., Tolson, M. A., and Copes, W. S. (1987). Evaluating trauma care: the TRISS method. *The Journal of Trauma and Acute Care Surgery*, 27(4), 370-378.
12. Bretas, F. S., 1990. Linear programming analysis of pesticide pollution of groundwater.
13. Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2002). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158(3), 280-287.
14. Chinneck, J. W. (2004). Practical optimization: a gentle introduction. Electronic document: <http://www.sce.carleton.ca/faculty/chinneck/po.html>.
15. Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3), 253-285.
16. Cox, DR (1970). *The Analysis of Binary Data*. London: Methuen.
17. Dantzig, G. B. (1998). *Linear programming and extensions*. Princeton university press.
18. Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352), 892-898.
19. Ferrier, G. D., and Lovell, C. K. (1990). Measuring cost efficiency in banking: econometric and linear programming evidence. *Journal of Econometrics*
20. Fitzpatrick, T. B. (1960). Albinism. *Journal of Investigative Dermatology*, 35(4), 209-214.
21. Frempong N. K. and Adjei I. J. (2011). Statistical Model 2. IDL KNUST.
22. Garver, L. L. (1970). Transmission network estimation using linear programming. *Power Apparatus and Systems, IEEE Transactions on*, (7), 1688-1697.
23. Gokbayrak, K., and Alper Yildirim, E. (2013). Joint gateway selection, transmission slot assignment, routing and power control for wireless mesh networks. *Computers & Operations Research*.

24. Hirji, K. F., Mehta, C. R., and Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, 82(400), 1110-1117.
25. Jansen, K., and Porkolab, L. (2003). Computing optimal preemptive schedules for parallel tasks: linear programming approaches. *Mathematical programming*, 95(3), 617-630.
26. Kim, Y. H., and Kim, S. (2009). Political Economy of International Policy Coordination for Market Regulation.
27. Lai, K. K., Yu, L., Zhou, L., and Wang, S. (2006). Credit risk evaluation with least square support vector machine. In *Rough Sets and Knowledge Technology* (pp. 490-495). Springer Berlin Heidelberg.
28. Luenberger, D. G., and Ye, Y. (2008). *Linear and nonlinear programming*, International Series in Operations Research & Management Science, 116.
29. Magder, L. S., and Hughes, J. P. (1996). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2), 195-203.
30. Manne, A. S. (1960). Linear programming and sequential decisions. *Management Science*, 6(3), 259-267.
31. Markowitz, H. M. (1957). The elimination form of the inverse and its application to linear programming. *Management Science*, 3(3), 255-269.
32. Ozsan, O., Simsir, F., and Pamukcu, C. (2010). Application of linear programming in production planning at marble processing plants. *Journal of Mining Science*, 46(1), 57-65.
33. Paul, D. B. (1995). *Controlling human heredity: 1865 to the present* (p. 2). Atlantic Highlands, NJ: Humanities Press.
34. Press, S. J., and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), 699-705.

35. Qin, J., and Zhang, B. (1996). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3), 609-618.
36. Rogers, H. J., and Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17(2), 105-116.
37. Shayeghi, H., and Bagheri, A. (2012). Dynamic sub-transmission system expansion planning incorporating distributed generation using hybrid DCGA and LP technique. *International Journal of Electrical Power & Energy Systems*, 48, 111-122.
38. Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4), 361-370.
39. Tashkin, D. P., Kanner, R., Bailey, W., Buist, S., Anderson, P., Nides, M. A., and Jamerson, B. D. (2001). Smoking cessation in patients with chronic obstructive pulmonary disease: a double-blind, placebo-controlled, randomised trial. *Lancet*, 357(9268), 1571-1575.
40. Tobin, J. (1965). The theory of portfolio selection. The theory of interest rates, 3-51 Macmillan Press.
- Transactions of the ASAE. V33(1):167-172.
41. Tyteca, D. (1997). Linear programming models for the measurement of environmental performance of firms—concepts and empirical results. *Journal of productivity analysis*, 8(2), 183-197.
42. Uryasev, S., and Rockafellar, R. T. (2000). Optimisation of Conditional Value-at-Risk. *Journal of Risk*, 2(3), 21-41.
43. Valenzuela, N., Botero, R., and Martínez, E. (1997). Field study of sex determination in *Podocnemis expansa* from Colombian Amazonia. *Herpetologica*, 390-398.

44. Villasana R., Garver L. and Salon S., (1985) "Transmission network planning using linear programming", IEEE Trans PAS, vol. ... pp. 349-355,
45. Wang, M., Xu, L., and Ramamurthy, B. (2010, March). Linear programming models for multi-channel P2P streaming systems. In INFOCOM, 2010 Proceedings IEEE (pp. 1-5). IEEE.
46. Wheaton, W. C. (1974). A bid rent approach to housing demand.
47. Wood, S. T., Dean, B. C., and Dean, D. (2013). A linear programming approach to reconstructing subcellular structures from confocal images for automated generation of representative 3D cellular models. Medical image analysis.
48. Yang, R. J., and Chuang, C. H. (1994). Optimal topology design using linear programming. Computers & Structures, 52(2), 265-275.
49. Yue, X. (2012). A Linear Programing Model and Partial Budget Analysis to Optimise Management startegies of Western Flower thrips in Greenhouse Inpatiens Production (Doctoral dissertation, Louisiana State University).
50. Zhang, J., and Kai, F. Y. (1998). What's the relative risk?. JAMA: the journal of the American Medical Association, 280(19), 1690-1691.
51. Zwang, Y., and Yarden, Y. (2009). Systems Biology of Growth Factor-Induced Receptor Endocytosis. Traffic, 10(4), 349-363.

| Ind Sector      | Int Rate   | Prob of bad Debt | Recovery Rate | bad debt limit |           |             |          | LIMIT      |        |  |  |
|-----------------|------------|------------------|---------------|----------------|-----------|-------------|----------|------------|--------|--|--|
| 1 Commerce      | 0.30       | 0.08             | 0.92          | 0.0700         | 0.01      |             |          | 166.0000   |        |  |  |
| 2 Agric         | 0.28       | 0.10             | 0.90          | 0.0700         | 0.03      |             |          |            |        |  |  |
| 3 Education     | 0.30       | 0.06             | 0.94          | 0.0700         | -0.01     |             |          |            |        |  |  |
| 4 Construction  | 0.32       | 0.09             | 0.91          | 0.0700         | 0.02      |             |          |            |        |  |  |
| 5 Consumer      | 0.36       | 0.08             | 0.92          | 0.0700         | 0.01      |             |          |            |        |  |  |
| 6 Export        | 0.28       | 0.06             | 0.94          | 0.0700         | -0.01     |             |          |            |        |  |  |
| 7 Finance       | 0.28       | 0.04             | 0.96          | 0.0700         | -0.03     |             |          |            |        |  |  |
| Obj fxn coeff   | 0.1960     | 0.1520           |               |                |           |             |          |            |        |  |  |
| Amt to allocate | 55.4369065 | 5.4869           | 32.208729     | 0              | 27.434259 | 31.36751278 | 14.06574 | 35.2498363 |        |  |  |
| s.t.            |            |                  |               |                |           |             |          |            |        |  |  |
| A               | 1          | 1                | 1             | 1              | 1         | 1           | 1        | 166.00     | 166.00 |  |  |
| B               | 1          | 0                | 1             | 1              | 0         | 1           | 0        | 119.01     | 99.60  |  |  |
| C               | 0          | 0                | 0             | 0              | 1         | 0           | 1        | 41.50      | 41.50  |  |  |
| D               | 0          | 1                | -0.5          | -0.5           | -0.5      | 1           | -0.5     | 0.00       | 0.00   |  |  |
| E               | 1          | -0.7             | -0.7          | -0.7           | -0.7      | 0           | -0.7     | 0.00       | 0.00   |  |  |
| F               | 0          | 1                | 0             | -0.2           | -0.2      | 0           | 0        | 0.00       | 0.00   |  |  |
| G               | 0.01       | 0.03             | -0.01         | 0.02           | 0.01      | -0.01       | -0.03    | 0.00       | 0.00   |  |  |

LIBRARY  
KWAME NKRUMAH  
UNIVERSITY OF SCIENCE & TECHNOLOGY  
KUMASI

Microsoft Excel 14.0 Solver Report  
Worksheet: [Project ch4.xlsx]Sheet1  
Report Created: 3/29/2013 1:19:34 PM

Result: Solver found a solution. All Constraints and optimality conditions are satisfied.

Solver Engine

Engine: Simplex LP  
Solution Time: 9 Seconds.  
Iterations: 8 Subproblems: 0

Solver Options

Max Time Unlimited, Iterations Unlimited, Precision 0.000001, Use Automatic Scaling, Show Iteration Results  
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%, Assume NonNegative

Objective Cell (Max)

| Cell    | Name                    | Original Value | Final Value |
|---------|-------------------------|----------------|-------------|
| \$J\$14 | Amt to allocate REVENUE | 34.4952        | 35.24983627 |

Variable Cells

| Cell    | Name                         | Original Value | Final Value | Integer |
|---------|------------------------------|----------------|-------------|---------|
| \$B\$14 | Amt to allocate Commerce     | 23             | 55.4369065  | Contin  |
| \$C\$14 | Amt to allocate Agric        | 23.0000        | 5.4869      | Contin  |
| \$D\$14 | Amt to allocate Education    | 24             | 32.208729   | Contin  |
| \$E\$14 | Amt to allocate Construction | 24             | 0           | Contin  |
| \$F\$14 | Amt to allocate Consumer     | 24             | 27.43425858 | Contin  |
| \$G\$14 | Amt to allocate Export       | 24             | 31.36751278 | Contin  |
| \$H\$14 | Amt to allocate Finance      | 24             | 14.06574142 | Contin  |

Constraints

| Cell    | Name             | Cell Value | Formula          | Status      | Slack |
|---------|------------------|------------|------------------|-------------|-------|
| \$I\$18 | A Total Utilised | 166.00     | \$I\$18<=\$J\$18 | Binding     | 0     |
| \$I\$19 | B Total Utilised | 119.01     | \$I\$19>=\$J\$19 | Not Binding | 19.41 |
| \$I\$20 | C Total Utilised | 41.50      | \$I\$20<=\$J\$20 | Binding     | 0     |
| \$I\$21 | D Total Utilised | 0.00       | \$I\$21>=\$J\$21 | Binding     | 0.00  |
| \$I\$22 | E Total Utilised | 0.00       | \$I\$22>=\$J\$22 | Binding     | 0.00  |
| \$I\$23 | F Total Utilised | 0.00       | \$I\$23>=\$J\$23 | Binding     | 0.00  |
| \$I\$24 | G Total Utilised | 0.00       | \$I\$24<=\$J\$24 | Binding     | 0     |

Variable Cells

| Cell    | Name                         | Final Value | Reduced Cost | Objective Coefficient | Allowable Increase | Allowable Decrease |
|---------|------------------------------|-------------|--------------|-----------------------|--------------------|--------------------|
| \$B\$14 | Amt to allocate Commerce     | 55.4369065  | 0            | 0.196                 | 0.022728889        | 0.011573183        |
| \$C\$14 | Amt to allocate Agric        | 5.486851717 | 0            | 0.152                 | 0.08718            | 0.052209091        |
| \$D\$14 | Amt to allocate Education    | 32.208729   | 0            | 0.2194                | 0.01363119         | 0.007647559        |
| \$E\$14 | Amt to allocate Construction | 0           | -0.010907451 | 0.2012                | 0.010907451        | 1E+30              |
| \$F\$14 | Amt to allocate Consumer     | 27.43425858 | 0            | 0.2512                | 0.024281           | 0.010441818        |
| \$G\$14 | Amt to allocate Export       | 31.36751278 | 0            | 0.2032                | 0.004800989        | 0.075214118        |
| \$H\$14 | Amt to allocate Finance      | 14.06574142 | 0            | 0.2288                | 0.010441818        | 0.024281           |

Constraints

| Cell    | Name             | Final Value  | Shadow Price | Constraint R.H. Side | Allowable Increase | Allowable Decrease |
|---------|------------------|--------------|--------------|----------------------|--------------------|--------------------|
| \$I\$18 | A Total Utilised | 166          | 0.208844777  | 166                  | 256.7466667        | 19.6282127         |
| \$I\$19 | B Total Utilised | 119.0131483  | 0            | 99.6                 | 19.41314828        | 1E+30              |
| \$I\$20 | C Total Utilised | 41.5         | 0.014014536  | 41.5                 | 17.8366443         | 25.20418848        |
| \$I\$21 | D Total Utilised | -7.10543E-15 | -0.003547261 | 0                    | 69.43897638        | 42.45390509        |
| \$I\$22 | E Total Utilised | -1.06581E-14 | -0.014942294 | 0                    | 73.48958333        | 55.01714286        |
| \$I\$23 | F Total Utilised | -1.77636E-15 | -0.070049671 | 0                    | 24.16054545        | 6.828636364        |
| \$I\$24 | G Total Utilised | -2.22045E-16 | 0.209751644  | 0                    | 0.700218182        | 1.365727273        |

## APPENDIX II

### Descriptives

Descriptive Statistics

|                               | N         | Minimum   | Maximum   | Mean      | Std. Deviation | Skewness  |
|-------------------------------|-----------|-----------|-----------|-----------|----------------|-----------|
|                               | Statistic | Statistic | Statistic | Statistic | Statistic      | Statistic |
| Age in years                  | 850       | 20        | 56        | 35.03     | 8.041          | .335      |
| Level of education            | 850       | 1         | 5         | 1.71      | .928           | 1.217     |
| Years with current employer   | 850       | 0         | 33        | 8.57      | 6.778          | .863      |
| Years at current address      | 850       | 0         | 34        | 8.37      | 6.895          | .924      |
| Household income in thousands | 850       | 13.00     | 446.00    | 46.6753   | 38.54305       | 3.701     |
| Debt to income ratio (x100)   | 850       | .10       | 41.30     | 10.1716   | 6.71944        | 1.125     |
| Number of dependents          | 850       | 0         | 6         | 1.54      | 1.258          | .635      |
| Other debt in thousands       | 850       | .05       | 35.20     | 3.0788    | 3.39880        | 3.206     |
| Previously defaulted          | 700       | 0         | 1         | .26       | .440           | 1.088     |
| validate                      | 700       | 0         | 1         | .70       | .459           | -.867     |
| Valid N (listwise)            | 700       |           |           |           |                |           |

Descriptive Statistics

|                    | Skewness   |
|--------------------|------------|
|                    | Std. Error |
| Age in years       | .084       |
| Level of education | .084       |

|                               |  |      |
|-------------------------------|--|------|
| Years with current employer   |  | .084 |
| Years at current address      |  | .084 |
| Household income in thousands |  | .084 |
| Debt to income ratio (x100)   |  | .084 |
| Number of dependents          |  | .084 |
| Other debt in thousands       |  | .084 |
| Previously defaulted          |  | .092 |
| validate                      |  | .092 |
| Valid N (listwise)            |  |      |

# Frequencies

## Frequency Table

|         |        | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|--------|-----------|---------|---------------|--------------------|
| Valid   | 0      | 211       | 24.8    | 30.1          | 30.1               |
|         | 1      | 489       | 57.5    | 69.9          | 100.0              |
|         | Total  | 700       | 82.4    | 100.0         |                    |
| Missing | System | 150       | 17.6    |               |                    |
| Total   |        | 850       | 100.0   |               |                    |

## Level of education

|  | Frequency | Percent | Valid Percent | Cumulative Percent |
|--|-----------|---------|---------------|--------------------|
|  |           |         |               |                    |

|       |                              |     |       |       |       |
|-------|------------------------------|-----|-------|-------|-------|
| Valid | Did not complete high school | 460 | 54.1  | 54.1  | 54.1  |
|       | High school degree           | 235 | 27.6  | 27.6  | 81.8  |
|       | Some college                 | 101 | 11.9  | 11.9  | 93.6  |
|       | College degree               | 49  | 5.8   | 5.8   | 99.4  |
|       | Post-undergraduate degree    | 5   | .6    | .6    | 100.0 |
|       | Total                        | 850 | 100.0 | 100.0 |       |

KNUST

Years with current employer

|         | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|-----------|---------|---------------|--------------------|
| Valid 0 | 72        | 8.5     | 8.5           | 8.5                |
| 1       | 59        | 6.9     | 6.9           | 15.4               |
| 2       | 50        | 5.9     | 5.9           | 21.3               |
| 3       | 50        | 5.9     | 5.9           | 27.2               |
| 4       | 57        | 6.7     | 6.7           | 33.9               |
| 5       | 49        | 5.8     | 5.8           | 39.6               |
| 6       | 53        | 6.2     | 6.2           | 45.9               |
| 7       | 45        | 5.3     | 5.3           | 51.2               |
| 8       | 38        | 4.5     | 4.5           | 55.6               |
| 9       | 52        | 6.1     | 6.1           | 61.8               |
| 10      | 38        | 4.5     | 4.5           | 66.2               |
| 11      | 32        | 3.8     | 3.8           | 70.0               |

|    |    |     |     |      |
|----|----|-----|-----|------|
| 12 | 38 | 4.5 | 4.5 | 74.5 |
| 13 | 32 | 3.8 | 3.8 | 78.2 |
| 14 | 16 | 1.9 | 1.9 | 80.1 |
| 15 | 23 | 2.7 | 2.7 | 82.8 |
| 16 | 33 | 3.9 | 3.9 | 86.7 |
| 17 | 14 | 1.6 | 1.6 | 88.4 |
| 18 | 22 | 2.6 | 2.6 | 90.9 |
| 19 | 17 | 2.0 | 2.0 | 92.9 |
| 20 | 7  | .8  | .8  | 93.8 |
| 21 | 10 | 1.2 | 1.2 | 94.9 |
| 22 | 13 | 1.5 | 1.5 | 96.5 |
| 23 | 6  | .7  | .7  | 97.2 |
| 24 | 5  | .6  | .6  | 97.8 |
| 25 | 4  | .5  | .5  | 98.2 |
| 26 | 1  | .1  | .1  | 98.4 |
| 27 | 3  | .4  | .4  | 98.7 |
| 28 | 1  | .1  | .1  | 98.8 |
| 29 | 2  | .2  | .2  | 99.1 |
| 30 | 3  | .4  | .4  | 99.4 |
| 31 | 3  | .4  | .4  | 99.8 |

Years with current employer

|          | Frequency | Percent | Valid Percent | Cumulative Percent |
|----------|-----------|---------|---------------|--------------------|
| Valid 33 | 2         | .2      | .2            | 100.0              |
| Total    | 850       | 100.0   | 100.0         |                    |

Years at current address

|         | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|-----------|---------|---------------|--------------------|
| 0       | 60        | 7.1     | 7.1           | 7.1                |
| 1       | 71        | 8.4     | 8.4           | 15.4               |
| 2       | 71        | 8.4     | 8.4           | 23.8               |
| 3       | 55        | 6.5     | 6.5           | 30.2               |
| 4       | 58        | 6.8     | 6.8           | 37.1               |
| 5       | 43        | 5.1     | 5.1           | 42.1               |
| 6       | 50        | 5.9     | 5.9           | 48.0               |
| Valid 7 | 41        | 4.8     | 4.8           | 52.8               |
| 8       | 49        | 5.8     | 5.8           | 58.6               |
| 9       | 45        | 5.3     | 5.3           | 63.9               |
| 10      | 37        | 4.4     | 4.4           | 68.2               |
| 11      | 36        | 4.2     | 4.2           | 72.5               |
| 12      | 28        | 3.3     | 3.3           | 75.8               |
| 13      | 22        | 2.6     | 2.6           | 78.4               |
| 14      | 24        | 2.8     | 2.8           | 81.2               |

|    |    |     |     |       |
|----|----|-----|-----|-------|
| 15 | 18 | 2.1 | 2.1 | 83.3  |
| 16 | 22 | 2.6 | 2.6 | 85.9  |
| 17 | 20 | 2.4 | 2.4 | 88.2  |
| 18 | 14 | 1.6 | 1.6 | 89.9  |
| 19 | 16 | 1.9 | 1.9 | 91.8  |
| 20 | 8  | .9  | .9  | 92.7  |
| 21 | 10 | 1.2 | 1.2 | 93.9  |
| 22 | 9  | 1.1 | 1.1 | 94.9  |
| 23 | 11 | 1.3 | 1.3 | 96.2  |
| 24 | 4  | .5  | .5  | 96.7  |
| 25 | 9  | 1.1 | 1.1 | 97.8  |
| 26 | 10 | 1.2 | 1.2 | 98.9  |
| 27 | 4  | .5  | .5  | 99.4  |
| 29 | 1  | .1  | .1  | 99.5  |
| 30 | 1  | .1  | .1  | 99.6  |
| 31 | 2  | .2  | .2  | 99.9  |
| 34 | 1  | .1  | .1  | 100.0 |

Years at current address

|             | Frequency | Percent | Valid Percent | Cumulative Percent |
|-------------|-----------|---------|---------------|--------------------|
| Valid Total | 850       | 100.0   | 100.0         |                    |

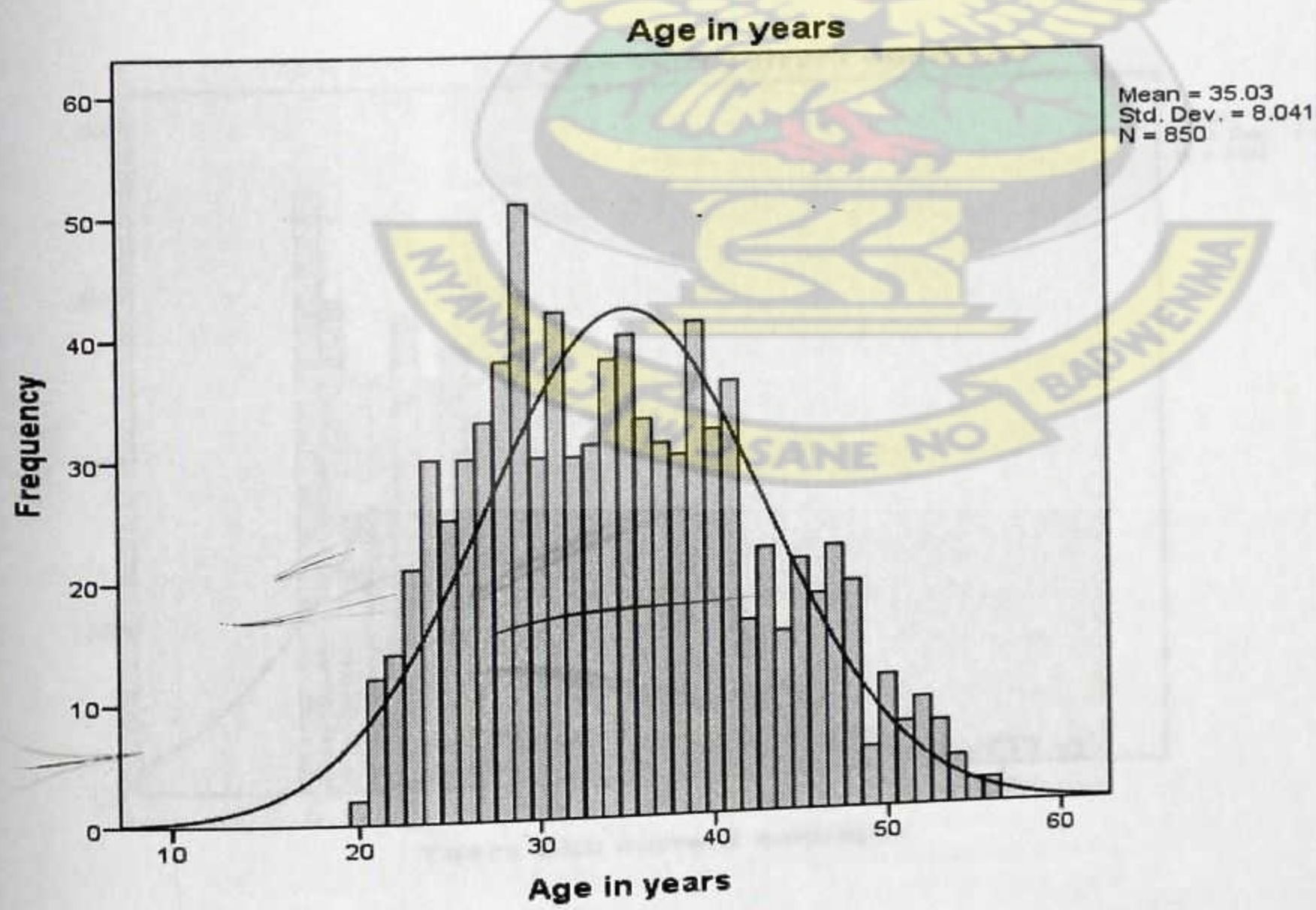
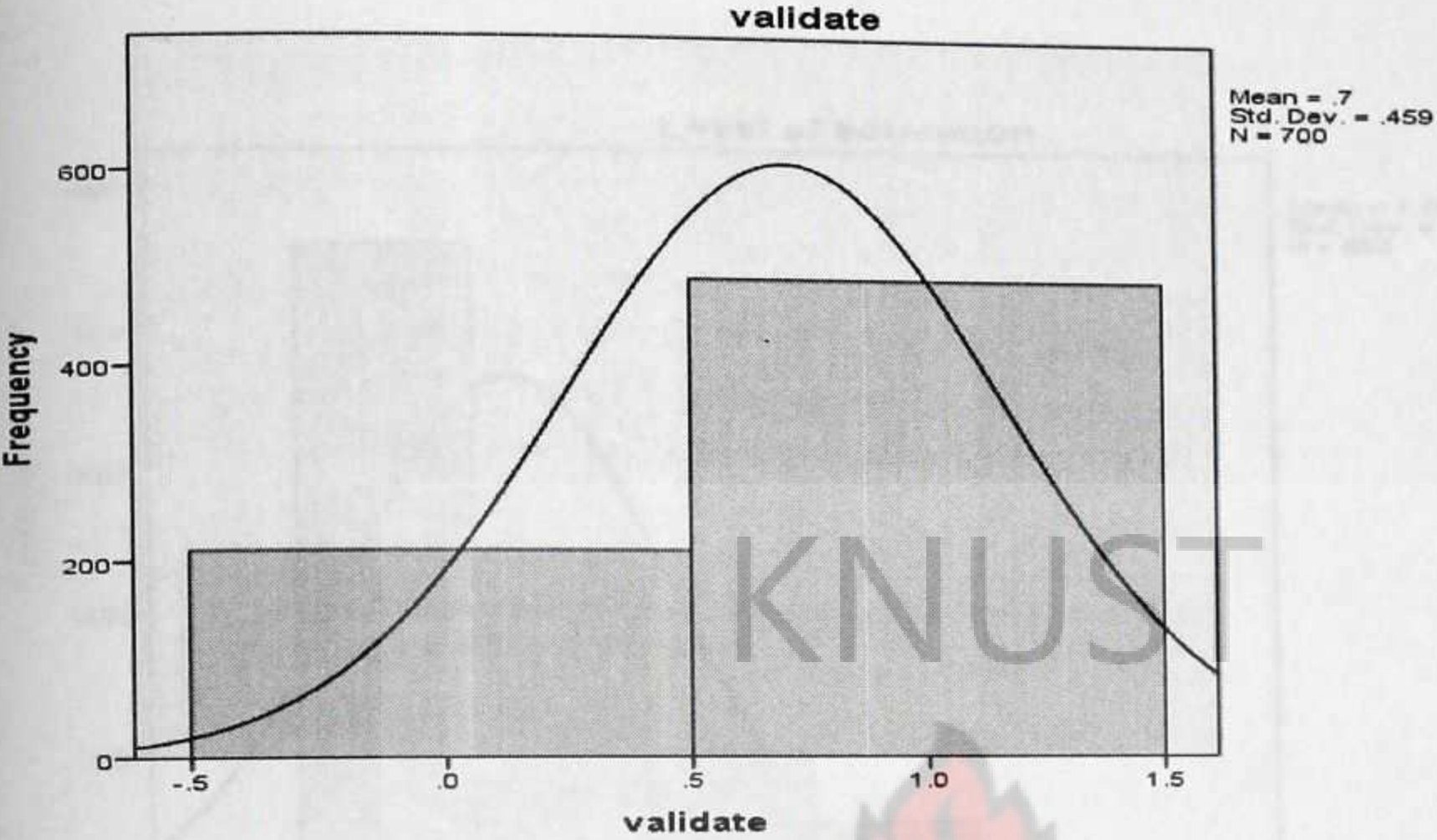
Number of dependents

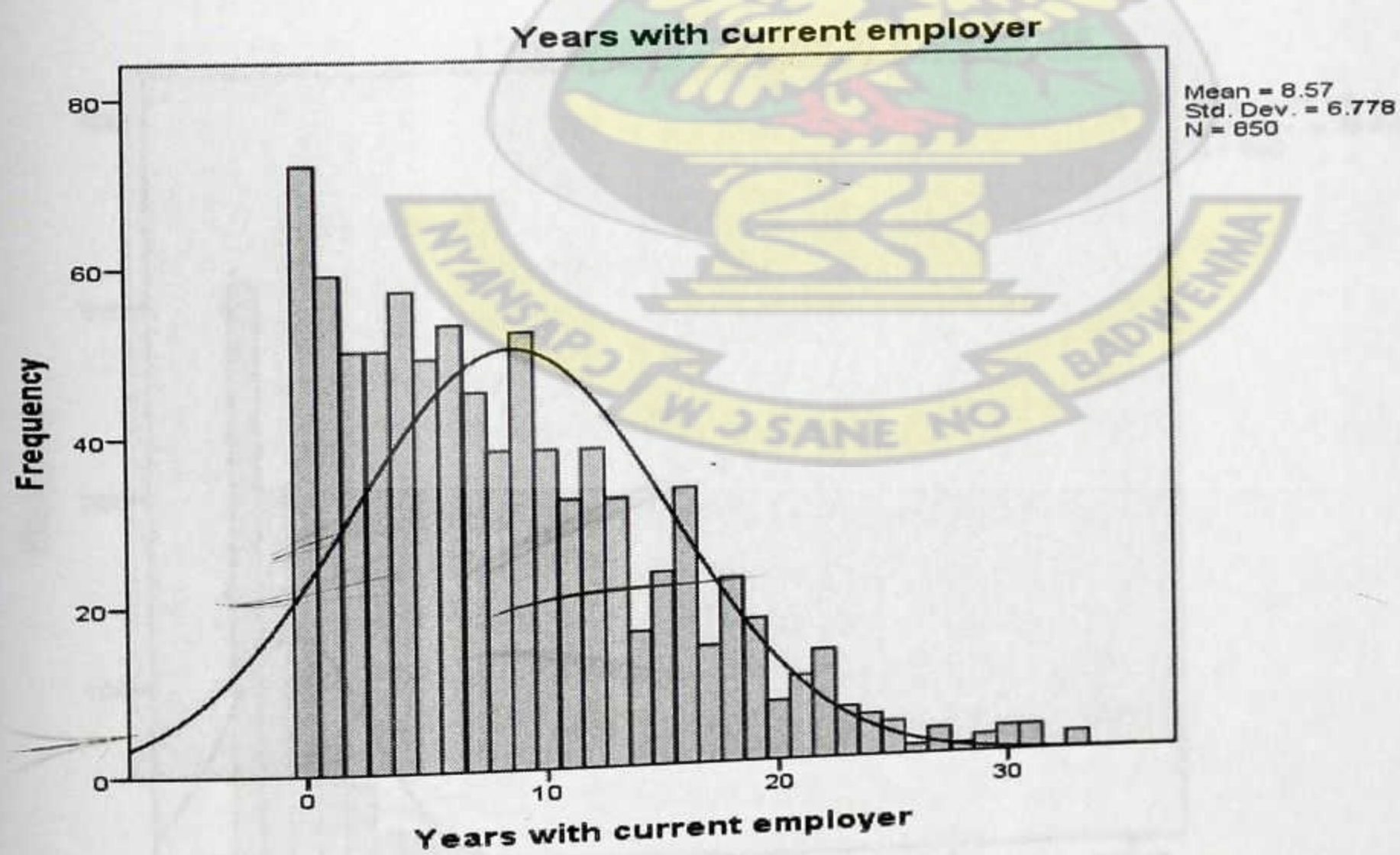
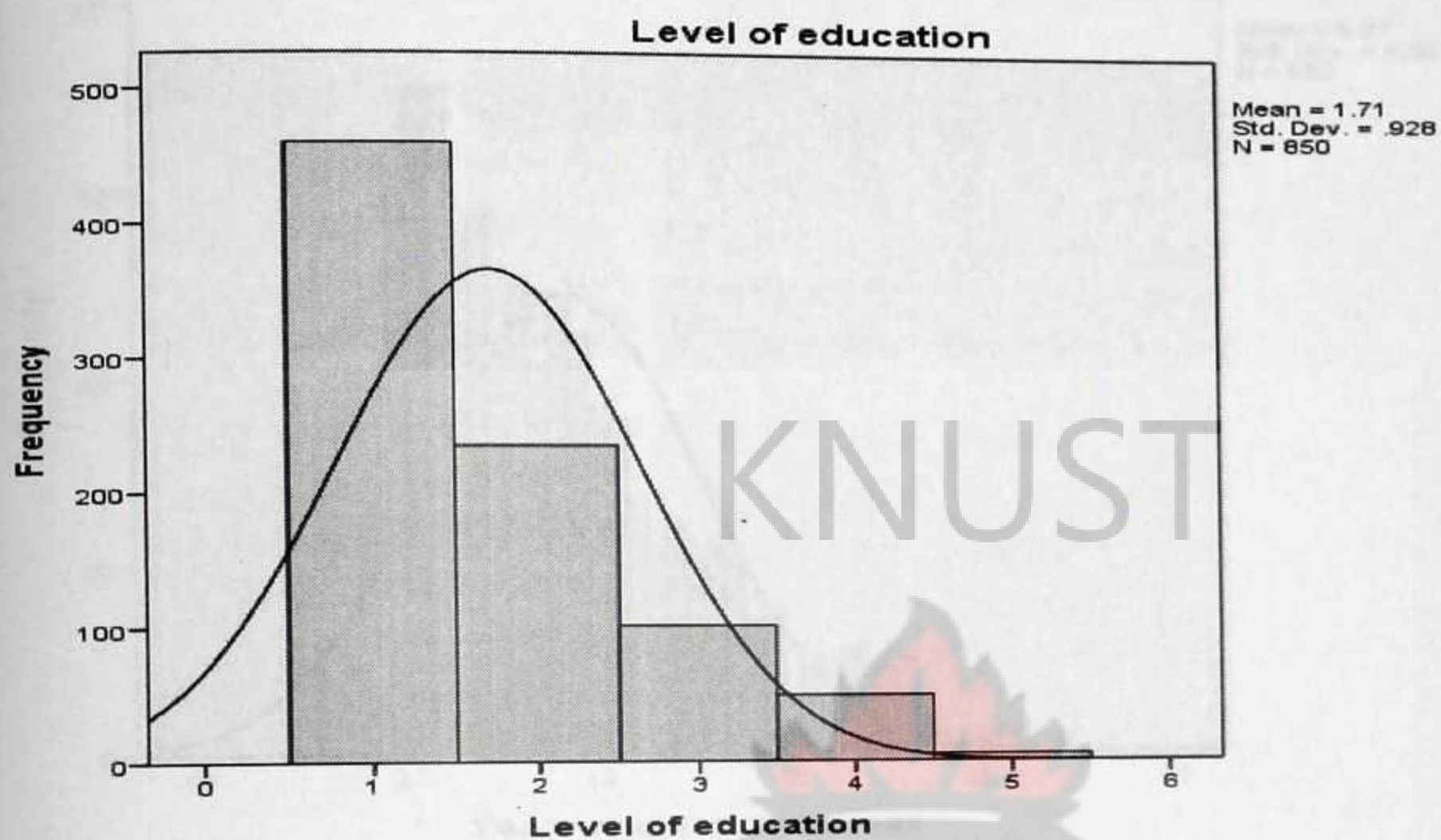
|       | Frequency | Percent | Valid.Percent | Cumulative<br>Percent |
|-------|-----------|---------|---------------|-----------------------|
| 0     | 206       | 24.2    | 24.2          | 24.2                  |
| 1     | 231       | 27.2    | 27.2          | 51.4                  |
| 2     | 241       | 28.4    | 28.4          | 79.8                  |
| 3     | 109       | 12.8    | 12.8          | 92.6                  |
| 4     | 45        | 5.3     | 5.3           | 97.9                  |
| 5     | 16        | 1.9     | 1.9           | 99.8                  |
| 6     | 2         | .2      | .2            | 100.0                 |
| Total | 850       | 100.0   | 100.0         |                       |

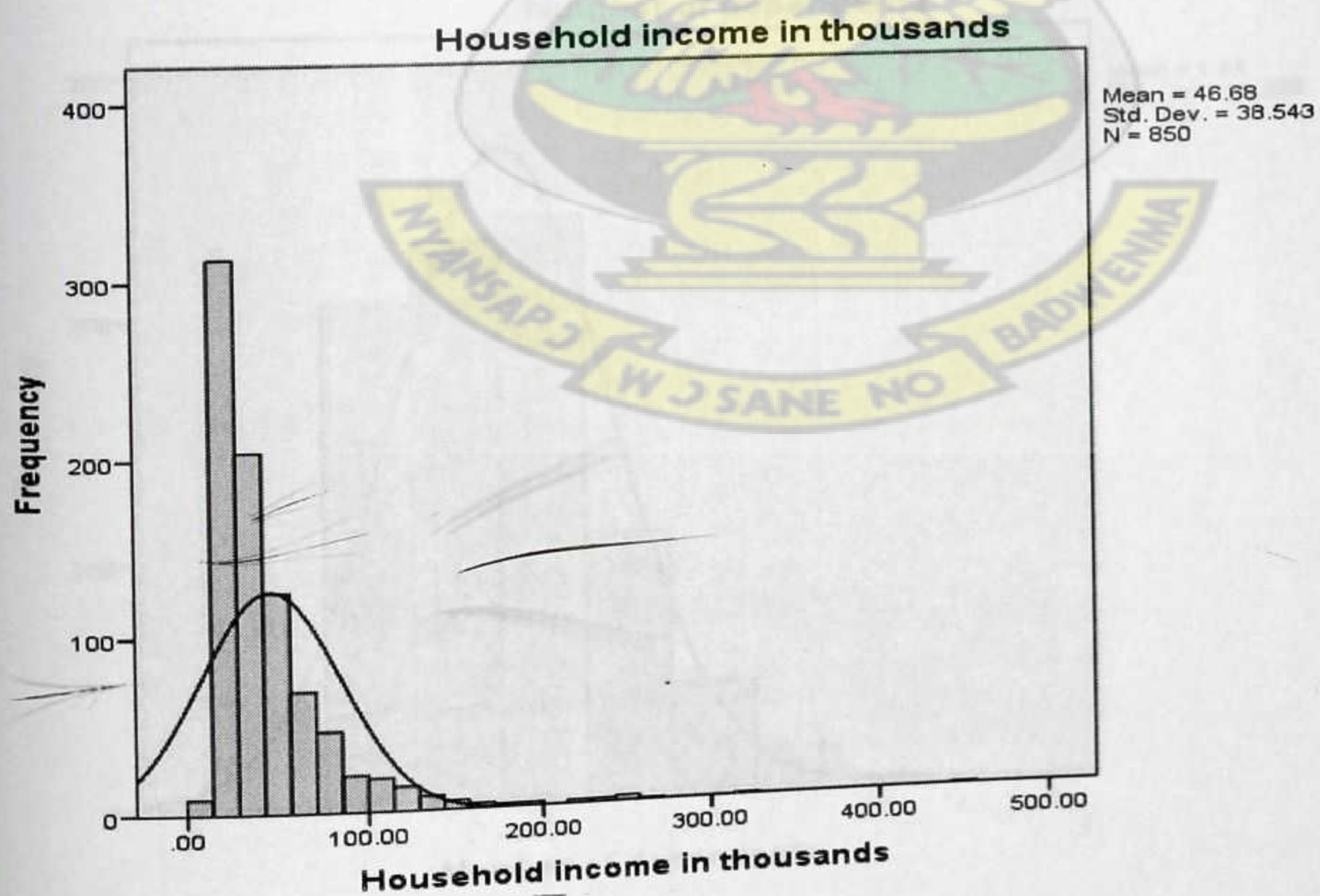
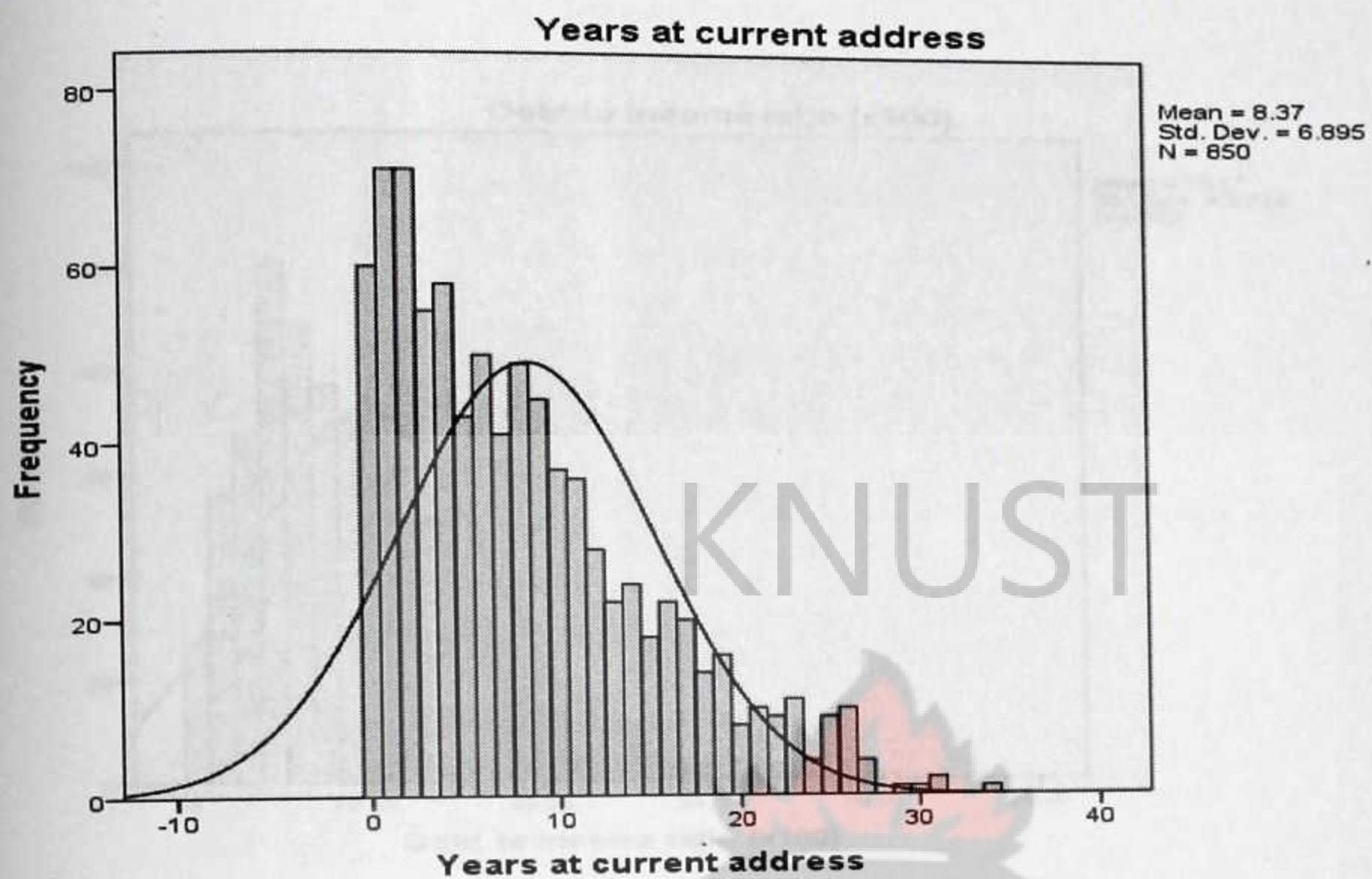
Previously defaulted

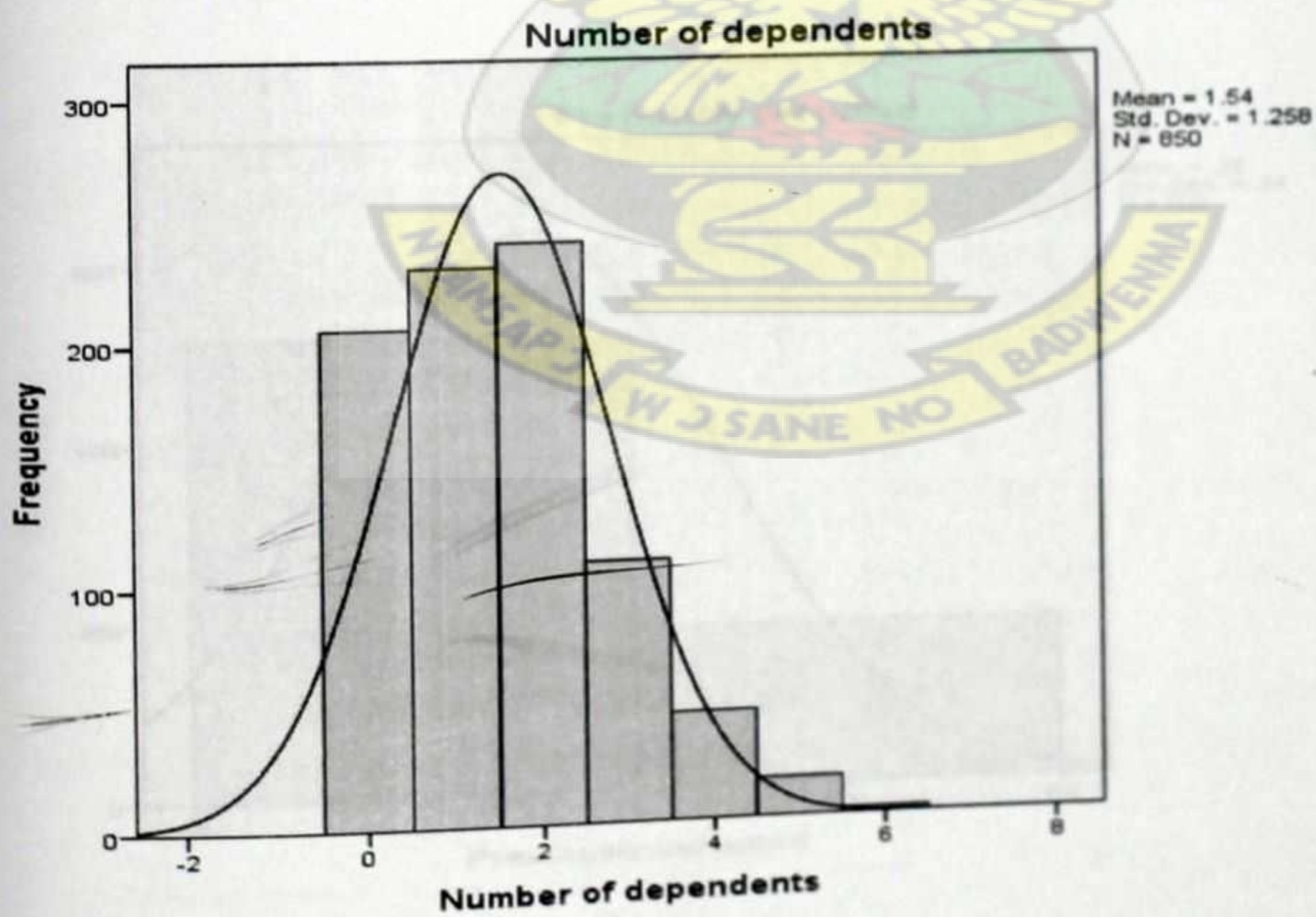
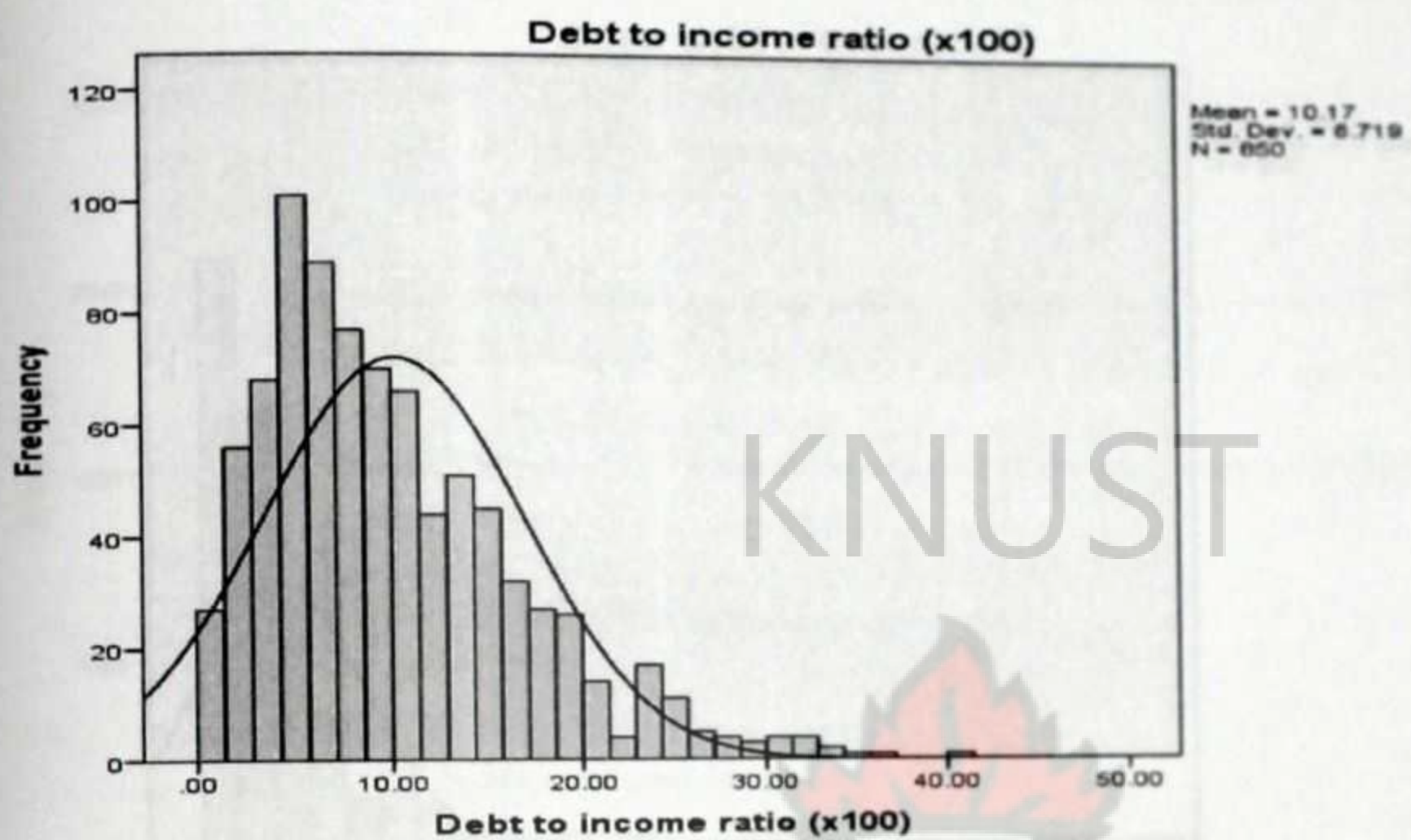
|                | Frequency | Percent | Valid Percent | Cumulative<br>Percent |
|----------------|-----------|---------|---------------|-----------------------|
| No             | 517       | 60.8    | 73.9          | 73.9                  |
| Valid Yes      | 183       | 21.5    | 26.1          | 100.0                 |
| Total          | 700       | 82.4    | 100.0         |                       |
| Missing System | 150       | 17.6    |               |                       |
| Total          | 850       | 100.0   |               |                       |

Histogram

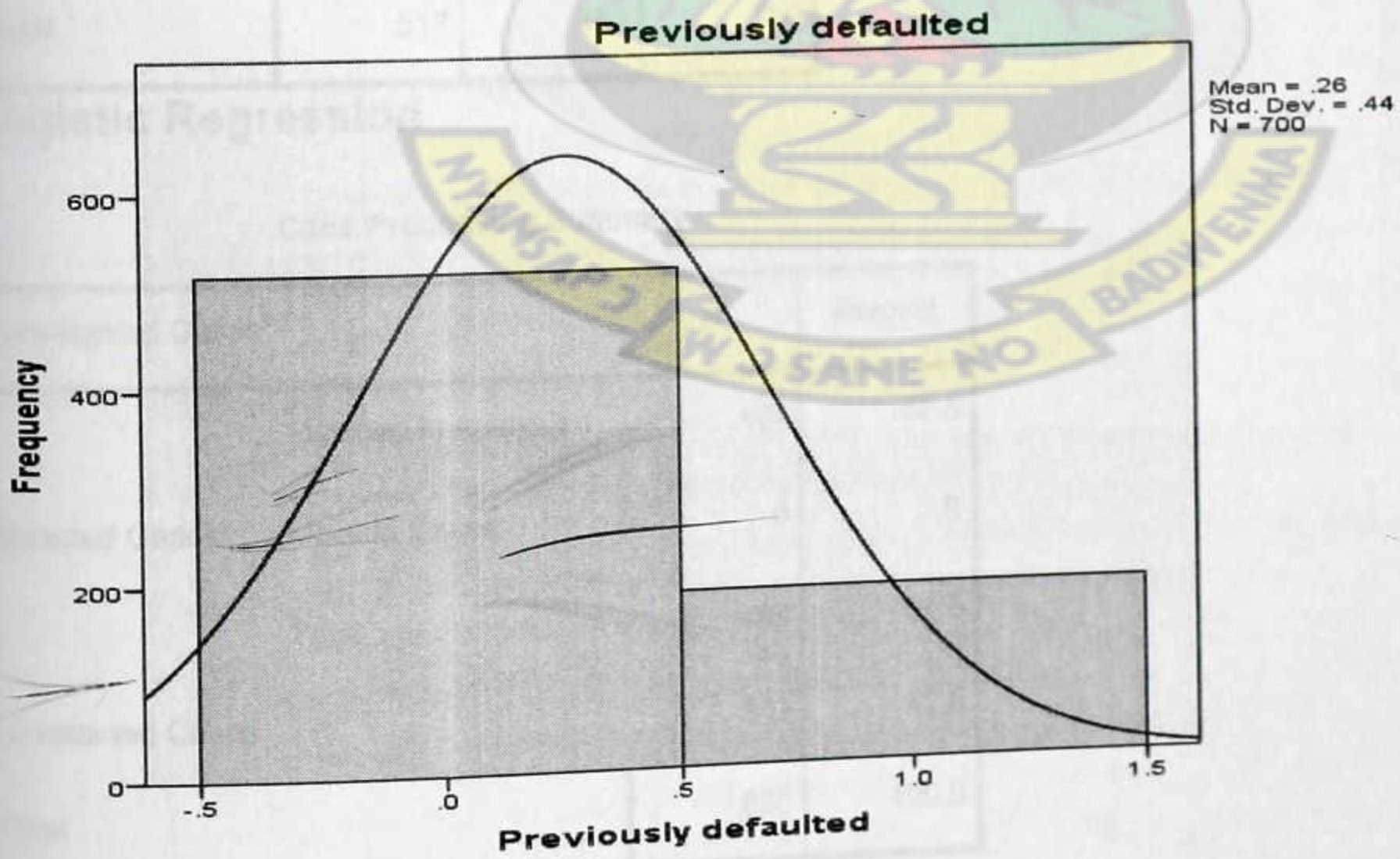
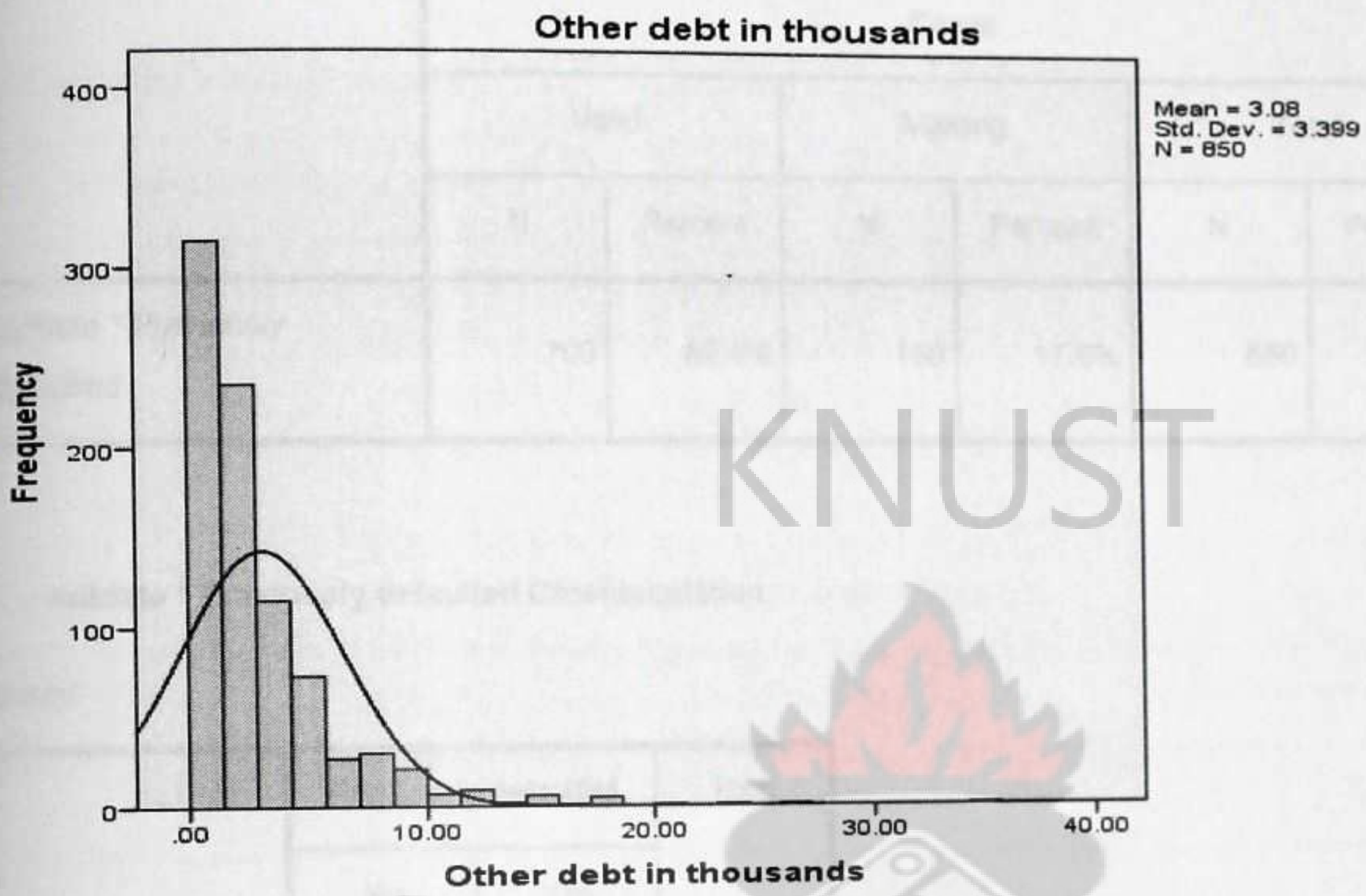








# Case Processing Summary



### Case Processing Summary

|                                 | Cases |         |         |         |       |         |
|---------------------------------|-------|---------|---------|---------|-------|---------|
|                                 | Valid |         | Missing |         | Total |         |
|                                 | N     | Percent | N       | Percent | N     | Percent |
| validate * Previously defaulted | 700   | 82.4%   | 150     | 17.6%   | 850   | 100.0%  |

### validate \* Previously defaulted Crosstabulation

Count

|       | Previously defaulted |     | Total |
|-------|----------------------|-----|-------|
|       | No                   | Yes |       |
| 0     | 157                  | 54  | 211   |
| 1     | 360                  | 129 | 489   |
| Total | 517                  | 183 | 700   |

### Logistic Regression

#### Case Processing Summary

| Unweighted Cases <sup>a</sup> |                      | N   | Percent |
|-------------------------------|----------------------|-----|---------|
| Selected Cases                | Included in Analysis | 489 | 57.5    |
|                               | Missing Cases        | 0   | .0      |
|                               | Total                | 489 | 57.5    |
| Unselected Cases              |                      | 361 | 42.5    |
| Total                         |                      | 850 | 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)   | (4)   |
| Level of education | Did not complete high school | 269       | 1.000            | .000  | .000  | .000  |
|                    | High school degree           | 134       | .000             | 1.000 | .000  | .000  |
|                    | Some college                 | 58        | .000             | .000  | 1.000 | .000  |
|                    | College degree               | 25        | .000             | .000  | .000  | 1.000 |
|                    | Post-undergraduate degree    | 3         | .000             | .000  | .000  | .000  |

Block 0: Beginning Block

Classification Table<sup>a,b</sup>

| Observed |  | Predicted                   |     |                                 |     |
|----------|--|-----------------------------|-----|---------------------------------|-----|
|          |  | Selected Cases <sup>c</sup> |     | Unselected Cases <sup>d,e</sup> |     |
|          |  | Previously defaulted        |     | Percentage Correct              |     |
|          |  | No                          | Yes | No                              | Yes |

|        |                      |     |     |   |       |     |   |
|--------|----------------------|-----|-----|---|-------|-----|---|
| Step 0 | Previously defaulted | No  | 360 | 0 | 100.0 | 157 | 0 |
|        |                      | Yes | 129 | 0 | .0    | 54  | 0 |
|        | Overall Percentage   |     |     |   | 73.6  |     |   |

Classification Table<sup>a,b</sup>

| Observed |                      | Predicted                     |       |
|----------|----------------------|-------------------------------|-------|
|          |                      | Unselected Cases <sup>c</sup> |       |
|          |                      | Percentage Correct            |       |
| Step 0   | Previously defaulted | No                            | 100.0 |
|          |                      | Yes                           | .0    |
|          | Overall Percentage   |                               | 74.4  |

- a. Constant is included in the model.
- b. The cut value is .500
- c. Selected cases validate EQ 1
- d. Unselected cases validate NE 1
- e. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Variables in the Equation

|        |          | B      | S.E. | Wald    | df | Sig. | Exp(B) |
|--------|----------|--------|------|---------|----|------|--------|
| Step 0 | Constant | -1.026 | .103 | 100.029 | 1  | .000 | .358   |

Variables not in the Equation

|                    |           |           | Score  | df   | Sig. |
|--------------------|-----------|-----------|--------|------|------|
| Step 0             | Variables | age       | 7.460  | 1    | .006 |
|                    |           | ed        | 8.994  | 4    | .061 |
|                    |           | ed(1)     | 6.089  | 1    | .014 |
|                    |           | ed(2)     | 1.145  | 1    | .285 |
|                    |           | ed(3)     | 2.224  | 1    | .136 |
|                    |           | ed(4)     | 2.516  | 1    | .113 |
|                    |           | employ    | 36.746 | 1    | .000 |
|                    |           | address   | 9.483  | 1    | .002 |
|                    |           | income    | 1.107  | 1    | .293 |
|                    |           | debtinc   | 76.418 | 1    | .000 |
|                    |           | numdepend | 1.169  | 1    | .280 |
|                    | othdebt   | 12.531    | 1      | .000 |      |
| Overall Statistics |           | 131.618   | 11     | .000 |      |

Block 1: Method = Forward Stepwise (Wald)

Omnibus Tests of Model Coefficients

|        |       | Chi-square | df | Sig. |
|--------|-------|------------|----|------|
| Step 1 | Step  | 74.052     | 1  | .000 |
|        | Block | 74.052     | 1  | .000 |

|        |       |         |   |      |
|--------|-------|---------|---|------|
|        | Model | 74.052  | 1 | .000 |
|        | Step  | 44.543  | 1 | .000 |
| Step 2 | Block | 118.595 | 2 | .000 |
|        | Model | 118.595 | 2 | .000 |
|        | Step  | 16.527  | 1 | .000 |
| Step 3 | Block | 135.122 | 3 | .000 |
|        | Model | 135.122 | 3 | .000 |
|        | Step  | 8.284   | 1 | .004 |
| Step 4 | Block | 143.406 | 4 | .000 |
|        | Model | 143.406 | 4 | .000 |

Model Summary

| Step | -2 Log likelihood    | Cox & Snell R Square | Nagelkerke R Square |
|------|----------------------|----------------------|---------------------|
| 1    | 490.252 <sup>a</sup> | .141                 | .205                |
| 2    | 445.709 <sup>b</sup> | .215                 | .315                |
| 3    | 429.182 <sup>b</sup> | .241                 | .353                |
| 4    | 420.898 <sup>b</sup> | .254                 | .371                |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

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

Hosmer and Lemeshow Test

| Step | Chi-square | df | Sig. |
|------|------------|----|------|
| 1    | 7.567      | 8  | .477 |
| 2    | 5.341      | 8  | .721 |
| 3    | 6.188      | 8  | .626 |
| 4    | 8.193      | 8  | .415 |

Contingency Table for Hosmer and Lemeshow Test

|    | Previously defaulted = No |          | Previously defaulted = Yes |          | Total |
|----|---------------------------|----------|----------------------------|----------|-------|
|    | Observed                  | Expected | Observed                   | Expected |       |
| 1  | 44                        | 44.388   | 5                          | 4.612    | 49    |
| 2  | 45                        | 43.344   | 4                          | 5.656    | 49    |
| 3  | 41                        | 41.487   | 7                          | 6.513    | 48    |
| 4  | 37                        | 40.520   | 11                         | 7.480    | 48    |
| 5  | 45                        | 40.201   | 4                          | 8.799    | 49    |
| 6  | 39                        | 37.607   | 9                          | 10.393   | 48    |
| 7  | 33                        | 35.142   | 15                         | 12.858   | 48    |
| 8  | 33                        | 32.590   | 16                         | 16.410   | 49    |
| 9  | 24                        | 27.217   | 25                         | 21.783   | 49    |
| 10 | 19                        | 17.506   | 33                         | 34.494   | 52    |
| 1  | 48                        | 47.541   | 1                          | 1.459    | 49    |
| 2  | 46                        | 46.044   | 3                          | 2.956    | 49    |
| 3  | 45                        | 44.258   | 4                          | 4.742    | 49    |

|        |    |    |        |    |        |    |
|--------|----|----|--------|----|--------|----|
| Step 3 | 4  | 42 | 42.494 | 7  | 6.506  | 49 |
|        | 5  | 38 | 40.385 | 11 | 8.615  | 49 |
|        | 6  | 35 | 37.855 | 14 | 11.145 | 49 |
|        | 7  | 38 | 34.830 | 11 | 14.170 | 49 |
|        | 8  | 35 | 30.562 | 14 | 18.438 | 49 |
|        | 9  | 21 | 23.564 | 28 | 25.436 | 49 |
|        | 10 | 12 | 12.465 | 36 | 35.535 | 48 |
|        | 1  | 49 | 47.842 | 0  | 1.158  | 49 |
|        | 2  | 46 | 46.529 | 3  | 2.471  | 49 |
|        | 3  | 44 | 45.010 | 5  | 3.990  | 49 |
| Step 4 | 4  | 41 | 43.026 | 8  | 5.974  | 49 |
|        | 5  | 42 | 40.864 | 7  | 8.136  | 49 |
|        | 6  | 35 | 38.401 | 14 | 10.599 | 49 |
|        | 7  | 38 | 35.098 | 11 | 13.902 | 49 |
|        | 8  | 33 | 29.732 | 16 | 19.268 | 49 |
|        | 9  | 20 | 22.247 | 29 | 26.753 | 49 |
|        | 10 | 12 | 11.251 | 36 | 36.749 | 48 |
|        | 1  | 49 | 48.133 | 0  | .867   | 49 |
|        | 2  | 46 | 46.830 | 3  | 2.170  | 49 |

Contingency Table for Hosmer and Lemeshow Test

|        |   | Previously defaulted = No |          | Previously defaulted = Yes |          | Total |
|--------|---|---------------------------|----------|----------------------------|----------|-------|
|        |   | Observed                  | Expected | Observed                   | Expected |       |
| Step 4 | 3 | 44                        | 45.262   | 5                          | 3.738    | 49    |

|    |    |        |    |        |    |
|----|----|--------|----|--------|----|
| 4  | 43 | 43.399 | 6  | 5.601  | 49 |
| 5  | 41 | 41.369 | 8  | 7.631  | 49 |
| 6  | 36 | 38.425 | 13 | 10.575 | 49 |
| 7  | 39 | 34.448 | 10 | 14.552 | 49 |
| 8  | 34 | 29.443 | 15 | 19.557 | 49 |
| 9  | 17 | 21.868 | 32 | 27.132 | 49 |
| 10 | 11 | 10.823 | 37 | 37.177 | 48 |

Classification Table<sup>a</sup>

| Observed |                      |     | Predicted                   |     |                    |                                 |     |
|----------|----------------------|-----|-----------------------------|-----|--------------------|---------------------------------|-----|
|          |                      |     | Selected Cases <sup>b</sup> |     |                    | Unselected Cases <sup>c,d</sup> |     |
|          |                      |     | Previously defaulted        |     | Percentage Correct | Previously defaulted            |     |
|          |                      |     | No                          | Yes |                    | No                              | Yes |
| Step 1   | Previously defaulted | No  | 340                         | 20  | 94.4               | 150                             | 7   |
|          |                      | Yes | 95                          | 34  | 26.4               | 42                              | 12  |
|          | Overall Percentage   |     |                             |     | 76.5               |                                 |     |
| Step 2   | Previously defaulted | No  | 335                         | 25  | 93.1               | 147                             | 10  |
|          |                      | Yes | 74                          | 55  | 42.6               | 39                              | 15  |
|          | Overall Percentage   |     |                             |     | 79.8               |                                 |     |
| Step 3   | Previously defaulted | No  | 337                         | 23  | 93.6               | 144                             | 13  |
|          |                      | Yes | 68                          | 61  | 47.3               | 36                              | 18  |
|          | Overall Percentage   |     |                             |     | 81.4               |                                 |     |
| Step 4   | Previously defaulted | No  | 337                         | 23  | 93.6               | 143                             | 14  |

|                    |    |    |      |    |    |
|--------------------|----|----|------|----|----|
| Yes                | 69 | 60 | 46.5 | 35 | 19 |
| Overall Percentage |    |    | 81.2 |    |    |

Classification Table<sup>a</sup>

| Observed |                      | Predicted                     |      |
|----------|----------------------|-------------------------------|------|
|          |                      | Unselected Cases <sup>b</sup> |      |
|          |                      | Percentage Correct            |      |
| Step 1   | Previously defaulted | No                            | 95.5 |
|          |                      | Yes                           | 22.2 |
|          | Overall Percentage   |                               | 76.8 |
| Step 2   | Previously defaulted | No                            | 93.6 |
|          |                      | Yes                           | 27.8 |
|          | Overall Percentage   |                               | 76.8 |
| Step 3   | Previously defaulted | No                            | 91.7 |
|          |                      | Yes                           | 33.3 |
|          | Overall Percentage   |                               | 76.8 |
| Step 4   | Previously defaulted | No                            | 91.1 |
|          |                      | Yes                           | 35.2 |
|          | Overall Percentage   |                               | 76.8 |

a. The cut value is .500

b. Selected cases validate EQ 1

c. Unselected cases validate NE 1

d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

| Variables in the Equation |          |        |      |         |    |      |        |                     |
|---------------------------|----------|--------|------|---------|----|------|--------|---------------------|
|                           |          | B      | S.E. | Wald    | df | Sig. | Exp(B) | 95% C.I. for EXP(B) |
|                           |          |        |      |         |    |      |        | Lower               |
| Step 1 <sup>a</sup>       | debtinc  | .129   | .016 | 61.777  | 1  | .000 | 1.138  | 1.102               |
|                           | Constant | -2.500 | .228 | 119.948 | 1  | .000 | .082   |                     |
| Step 2 <sup>b</sup>       | employ   | -.131  | .022 | 34.850  | 1  | .000 | .877   | .840                |
|                           | debtinc  | .140   | .018 | 61.974  | 1  | .000 | 1.150  | 1.111               |
| Step 3 <sup>c</sup>       | Constant | -1.695 | .258 | 43.051  | 1  | .000 | .184   |                     |
|                           | employ   | -.194  | .029 | 44.691  | 1  | .000 | .824   | .778                |
|                           | income   | .017   | .005 | 12.880  | 1  | .000 | 1.017  | 1.008               |
|                           | debtinc  | .147   | .018 | 63.826  | 1  | .000 | 1.159  | 1.117               |
|                           | Constant | -2.064 | .285 | 52.349  | 1  | .000 | .127   |                     |
| Step 4 <sup>d</sup>       | employ   | -.194  | .030 | 43.110  | 1  | .000 | .823   | .777                |
|                           | address  | -.060  | .022 | 7.770   | 1  | .005 | .942   | .903                |
|                           | income   | .021   | .005 | 16.810  | 1  | .000 | 1.021  | 1.011               |
|                           | debtinc  | .152   | .019 | 64.793  | 1  | .000 | 1.164  | 1.122               |
|                           | Constant | -1.822 | .296 | 37.781  | 1  | .000 | .162   |                     |

| Variables in the Equation |                     |
|---------------------------|---------------------|
|                           | 95% C.I. for EXP(B) |

|                     |          | Upper |
|---------------------|----------|-------|
| Step 1 <sup>a</sup> | debtinc  | 1.175 |
|                     | Constant |       |
| Step 2 <sup>b</sup> | employ   | .916  |
|                     | debtinc  | 1.191 |
| Step 3 <sup>c</sup> | Constant |       |
|                     | employ   | .872  |
| Step 4 <sup>d</sup> | income   | 1.027 |
|                     | debtinc  | 1.201 |
|                     | Constant |       |
|                     | employ   | .873  |
|                     | address  | .982  |
|                     | income   | 1.031 |
|                     | debtinc  | 1.208 |
|                     | Constant |       |

a. Variable(s) entered on step 1: debtinc.

b. Variable(s) entered on step 2: employ.

c. Variable(s) entered on step 3: income.

d. Variable(s) entered on step 4: address.

Variables not in the Equation

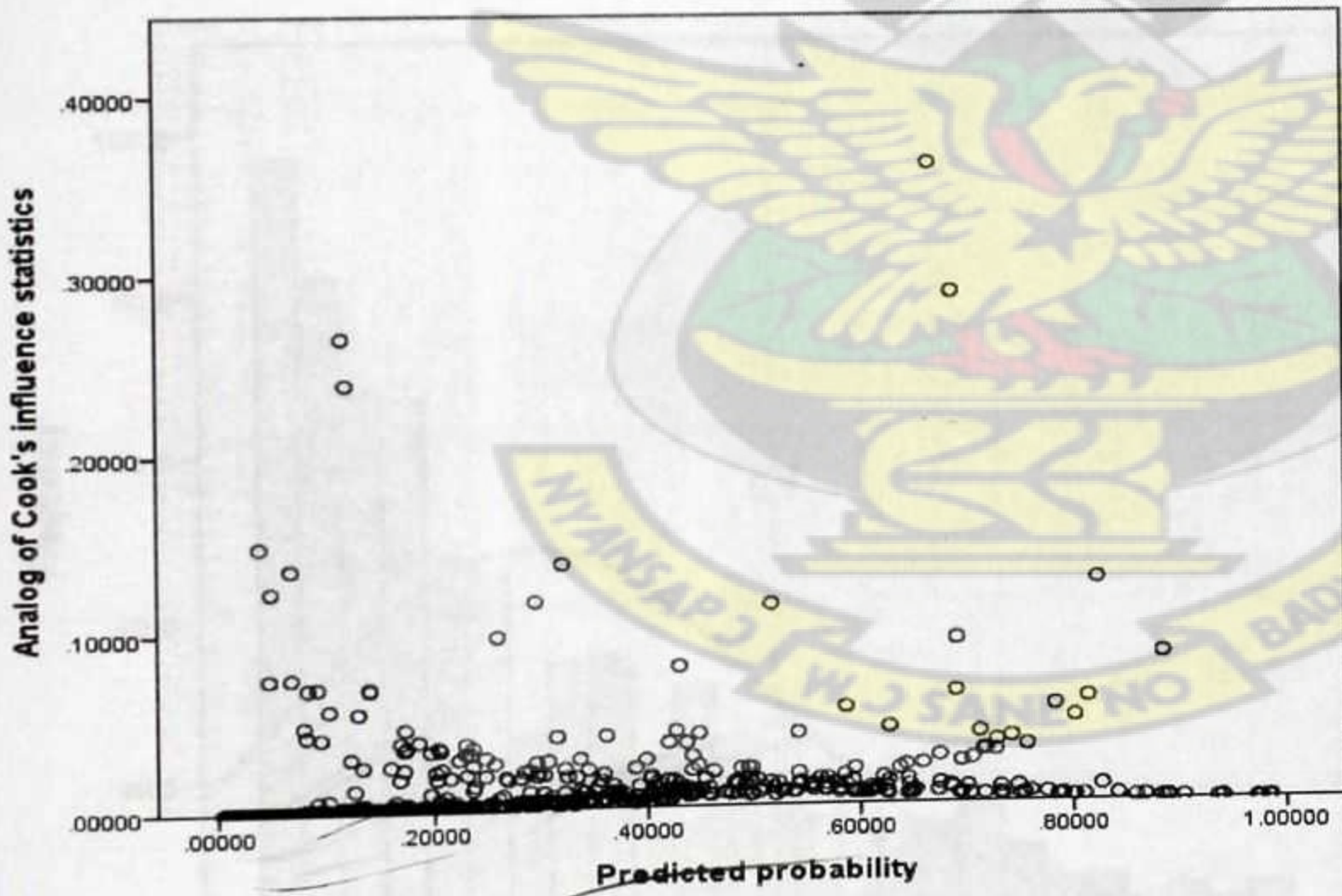
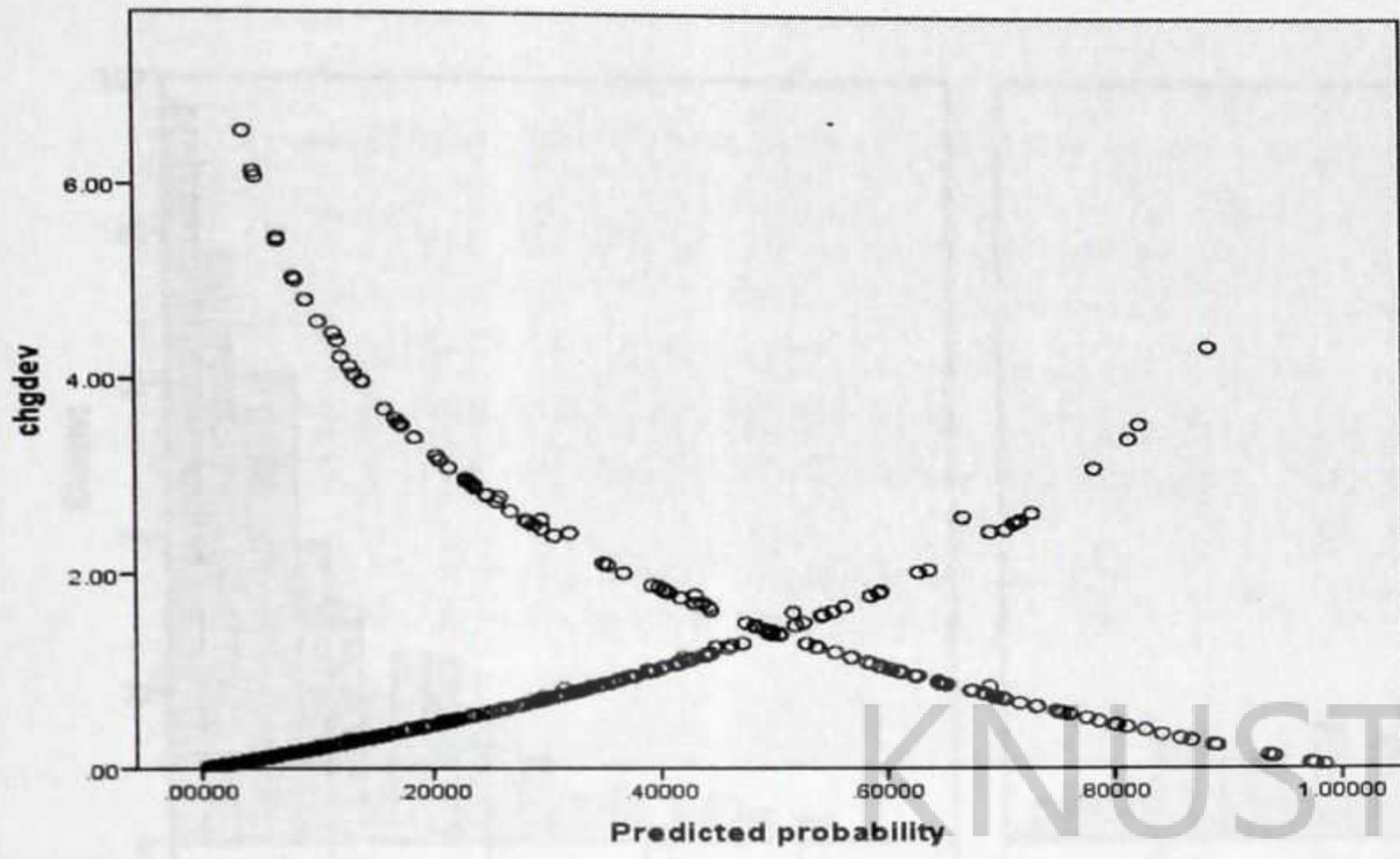
|        |           |     | Score   | df | Sig. |
|--------|-----------|-----|---------|----|------|
| Step 1 | Variables | age | — 9.193 | 1  | .002 |

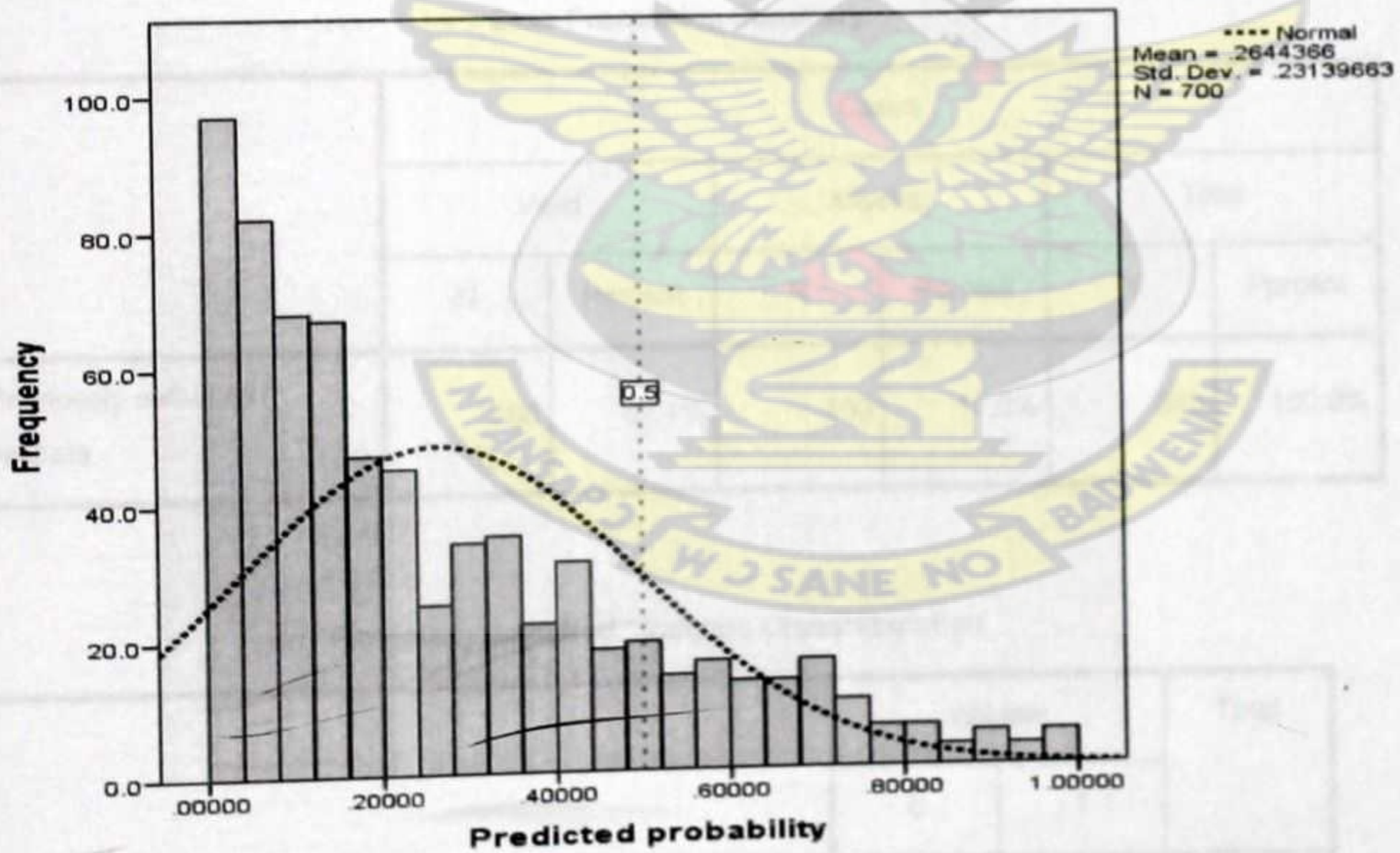
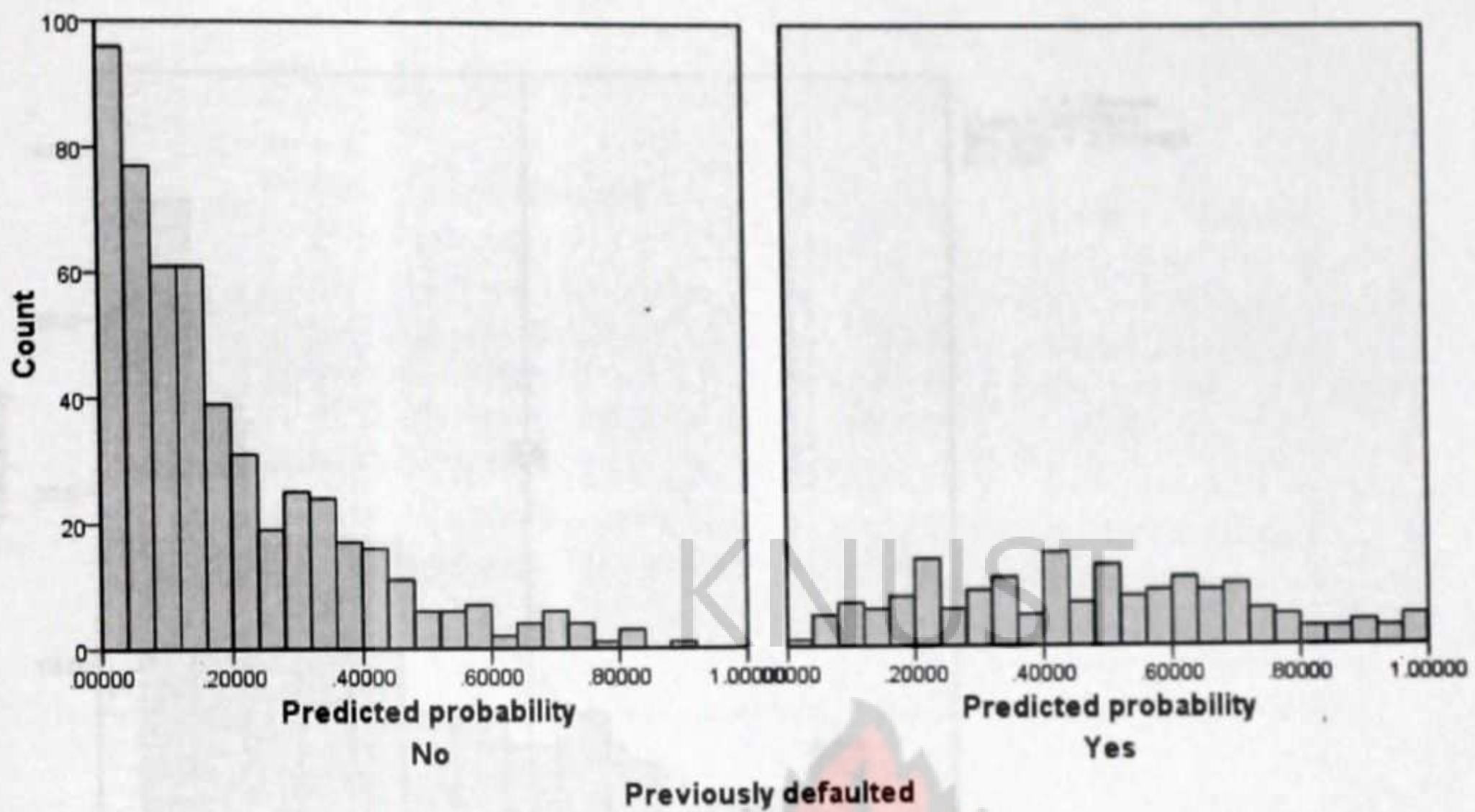
|        |                    |           |        |    |      |
|--------|--------------------|-----------|--------|----|------|
|        |                    | ed        | 11.254 | 4  | .024 |
|        |                    | ed(1)     | 9.069  | 1  | .003 |
|        |                    | ed(2)     | 2.466  | 1  | .116 |
|        |                    | ed(3)     | 1.916  | 1  | .166 |
|        |                    | ed(4)     | 3.350  | 1  | .067 |
|        |                    | employ    | 39.428 | 1  | .000 |
|        |                    | address   | 10.841 | 1  | .001 |
|        |                    | income    | .744   | 1  | .388 |
|        |                    | numdepend | .484   | 1  | .487 |
|        |                    | othdebt   | 4.963  | 1  | .026 |
|        | Overall Statistics |           | 65.774 | 10 | .000 |
|        |                    | age       | .107   | 1  | .744 |
|        |                    | ed        | 4.682  | 4  | .322 |
|        |                    | ed(1)     | 3.324  | 1  | .068 |
|        |                    | ed(2)     | .444   | 1  | .505 |
|        |                    | ed(3)     | 1.963  | 1  | .161 |
|        |                    | ed(4)     | .581   | 1  | .446 |
| Step 2 | Variables          | address   | 3.240  | 1  | .072 |
|        |                    | income    | 26.794 | 1  | .000 |
|        |                    | numdepend | .213   | 1  | .644 |
|        |                    | othdebt   | 4.986  | 1  | .026 |
|        | Overall Statistics |           | 42.919 | 9  | .000 |
|        |                    | age       | .381   | 1  | .537 |
| Step 3 | Variables          | ed        | 2.029  | 4  | .730 |

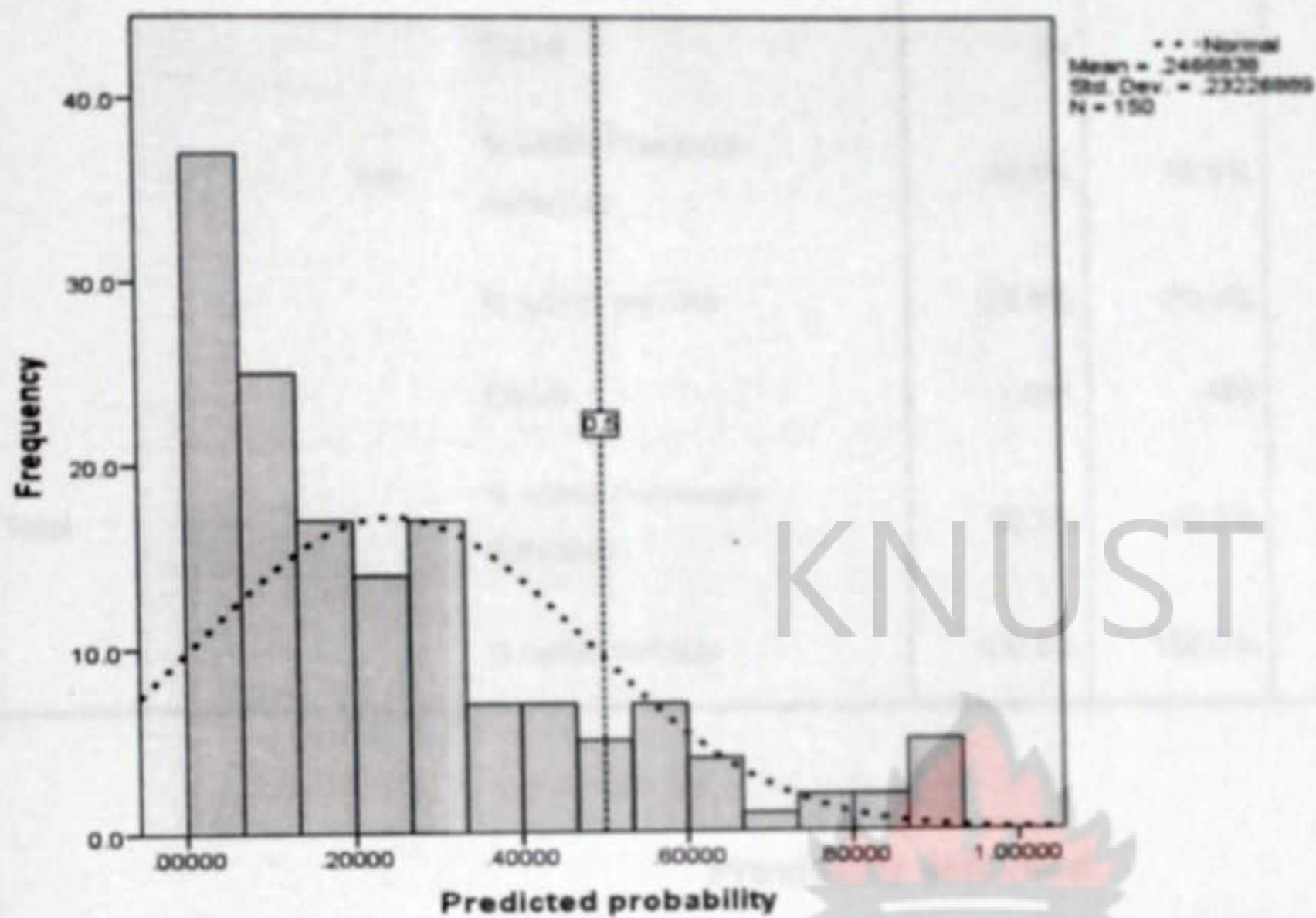
|                    |        |   |      |
|--------------------|--------|---|------|
| ed(1)              | .272   | 1 | .602 |
| ed(2)              | .306   | 1 | .580 |
| ed(3)              | .285   | 1 | .593 |
| ed(4)              | .167   | 1 | .682 |
| address            | 7.955  | 1 | .005 |
| numdepend          | .078   | 1 | .780 |
| othdebt            | .357   | 1 | .550 |
| Overall Statistics | 12.327 | 8 | .137 |

#### Variables not in the Equation

|                    |           | Score | df | Sig. |
|--------------------|-----------|-------|----|------|
| Step 4             | Variables |       |    |      |
|                    | age       | 1.477 | 1  | .224 |
|                    | ed        | 2.054 | 4  | .726 |
|                    | ed(1)     | .370  | 1  | .543 |
|                    | ed(2)     | .201  | 1  | .654 |
|                    | ed(3)     | .462  | 1  | .497 |
|                    | ed(4)     | .047  | 1  | .829 |
|                    | numdepend | .733  | 1  | .392 |
|                    | othdebt   | .116  | 1  | .734 |
| Overall Statistics |           | 4.318 | 7  | .742 |







Case Processing Summary

|                                    | Cases |         |         |         |       |         |
|------------------------------------|-------|---------|---------|---------|-------|---------|
|                                    | Valid |         | Missing |         | Total |         |
|                                    | N     | Percent | N       | Percent | N     | Percent |
| Previously defaulted *<br>validate | 700   | 82.4%   | 150     | 17.6%   | 850   | 100.0%  |

Previously defaulted \* validate Crosstabulation

|                      |                               |  | validate |       | Total  |
|----------------------|-------------------------------|--|----------|-------|--------|
|                      |                               |  | 0        | 1     |        |
| Previously defaulted | Count                         |  | 157      | 360   | 517    |
|                      | % within Previously defaulted |  | 30.4%    | 69.6% | 100.0% |

|       |     |                               |        |        |        |
|-------|-----|-------------------------------|--------|--------|--------|
| Total | Yes | % within validate             | 74.4%  | 73.6%  | 73.9%  |
|       |     | Count                         | 54     | 129    | 183    |
|       |     | % within Previously defaulted | 29.5%  | 70.5%  | 100.0% |
|       |     | % within validate             | 25.6%  | 26.4%  | 26.1%  |
|       |     | Count                         | 211    | 489    | 700    |
|       |     | % within Previously defaulted | 30.1%  | 69.9%  | 100.0% |
|       |     | % within validate             | 100.0% | 100.0% | 100.0% |

