PARAMETRIC MODELING OF PROBABILITY OF BANK LOAN DEFAULT IN KENYA

O. Adem[1], A. W. Gichuhi[2] and R. O. Otieno[3]
[1]Mombasa Polytechnic University College
[2, 3]Statistics and Actuarial Science Department, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
E-mail: aggreyadem@yahoo.com

Abstract
Commercial banks in Kenya are the key players not only in the financial market but also in spurring the economic growth that has been witnessed in the country in the recent past. Besides Safaricom and East Africa Breweries, the other top ten most profitable companies in Kenya are the Commercial banks. The biggest part of these huge profits emanates from the interests charged on loans they advance to their customers. If these loans non-perform, these blue chip companies will come tumbling down and the entire economy will be threatened. This makes the study on probability of a customer defaulting very useful while analyzing the credit risk policies. In this paper, we use a raw data set that contains demographic information about the borrowers. The data sets have been used to identify which risk factors associated with the borrowers contribute towards default. These risk factors are gender, age, marital status, occupation and term of loan. Results show that male customers have high odds (1.91) of defaulting compared to their female counter parts, single customers have a higher likelihood (odds of 1.48) of defaulting compared to their married customers, younger customers have high odds of defaulting unlike elderly customers, financial sector customers have equal likelihood of default as support staff customers and long term loans have less likelihood of defaulting compared to short term loans.

Key words: The logistic model, the logit transformation, parameter estimation

1.0 Introduction

The Basel capital accord encourages financial institutions to develop and strengthen risk management systems. As a result banks are interested in obtaining a more objective rating of loan portfolios. High levels of indebtness imply a higher incident of default and increasing risk for lenders. Since unsecured personal loans' policies change rapidly, stringent measures have been enhanced to curb this dynamism. Quantifying financial risks and developing an efficient portfolio management strategy are important objectives of the banks. Banks consequently devote many resources to developing internal risks models. By developing an accurate credit risk rating system, banks will be able to identify loans that have lower probability of default verses loans that have higher probability of default. Thus they will better rate the loans and ultimately price the loans. In this study, we focus on Kenyan Commercial banks. Most banks in Kenya assess their loan applications based on the customers' ability to pay. This is pegged on the income Vis a Vis the expenditures. In most Kenyan commercial Banks, net income is like the core factor in loans approval.

The loan products are structured such that, there is a minimum net income requirement. This in most cases locks out those customers who earn less than the minimum income band yet they will service their loans had they been given a chance to borrow. The maximum amount of loan one can borrow is determined based on his other monthly commitments and the proposed new loan repayment amounts. In most banks the maximum ratio between the total monthly commitments and the net monthly salary is fixed at fifty percent. Therefore, Commercial Banks in Kenya do not put into consideration some of the customers' demographic attributes. In this study we focus on the customers' attributes that are statistically significant with respect to their ability to repay their loans and show their contributions towards the probability of default.

To achieve this we used the logistic regression model which is widely employed to model the outcomes of a categorical dependent variable. For categorical variables it is inappropriate to use linear regression because the response values are not measured on a ratio scale and the error terms are not normally distributed. In addition, the linear regression model can generate as predicted values any real number ranging from negative to positive infinity, where as a categorical variable can only take on a limited number of discrete values within a specified range.

The financial distress literature has been focused on finding explanatory variables that have discriminating power to differentiate financially sound companies. Initiated by Beaver (1966), Altman (1968) and Ohlson (1980), academic studies to measure financial vulnerability continued for three decades. Beaver found that the cash owed to debt ratio was the best single ratio predictor of distress in his univariate discriminant analysis.

Altman's Z-score model used multivariate discriminant analysis to select the five most significant variables for measuring the financial distress of firms. Ohlson's O-score model used a logit analysis to generate a one year prediction model. Other streams of financial modeling have been utilizing various statistical methods to predict various patterns. A few significant methods are: multinomial choice models such as logit and probit models. However, there is no consensus as to the best statistical model. Other scholars have contributed immensely to develop the earlier works and make them better.

Martin (1977) employed the application of binary choice model (logit) to the probability of a bank failure. He used a two year horizon between the statement year for the financial ratio data and the observation year of the banks situation (failed or operating). Wiginton (1980) discovered that the logit model results are more superior to discriminant analysis for consumer credit scoring. Lawrence *et al* (1992) used the analysis of default risk in mobile home credits in the US between 1974 and 1980.

Westgaards and Wijst (2001) employed a logit model for analysis of default affecting Norwegian limited liability companies during the years 1995-1999. They found that a two year period between the firm status and a firm's accounting data is optimal. Kolari *et al*. (2002) used a logit approach to modeling probability of default for US banks between 1989 and 1990.

Soest *et al* (2003) introduced a model that combines a clustery procedure with logit model fitting. They examined the extent to which macroeconomic variables are helpful in predicting bank defaults. They found that a preliminary expert clustering or automatic clustering improves the predictive power of the model and incorporation of macro variables into the model is useful. They suggested heuristic criteria to help compare model performance from the perspective of investors or bank supervision authorities. They analyzed Russian Banking system trends after the 1998 Russian crisis with rolling regressions. The variables studied included real GDP index, consumer price index, deflation, unemployment, unemployment rate, index of investments in capital, exchange rate, export/import ratio, increase of industrial production and change in real income.

Froelich (2006) applied non-parametric regression for binary dependent variables. In this paper the local logit regression is used to analyze heterogeneity effects of children on female labor supply. A comparison is made on parametric, semi parametric and non parametric modeling and it's found that, the parametric logit and semi parametric Klein-Spady estimators do not detect heterogeneity.

Jeong (2010) used the maximum entropy method which is one of the nonparametric methods used to estimate the probability of default from binary

option prices. Very little has been done in terms of studying the behavioral attributes of the loan applicants in Kenya. Therefore, in this research we examined and exploited the properties of a logit model and extended its application to banking by parametrically modeling default rate in Kenyan commercial banks. The data used to construct the model was obtained from Barclays Bank of Kenya, one of the Kenyan commercial banks.

2.0 The Model

In this study we used the logistic model that caters for categorical variables in a way roughly analogous to that in which the linear regression model is used for continuous variables. Logistic regression has proven to be one of the most versatile techniques in the class of generalized linear models as we see in the next sections.

2.1 The Logistic Model

The logistic model is one of the regression models for dichotomous data. It is appropriate when the response variable takes one of the only two possible outcomes representing success and failure, or more generally the presence or absence of an attribute of interest. Consider our set of data consisting of successful loan applicants whose applications were done in vintage. The behavior of these applicants can only take two forms; either they pay or default in payment. The concept of the logistic model is based on Bernoulli and Binomial distributions. To bring out this clearly we discuss the stochastic structure of the data in terms of Bernoulli and Binomial distributions, and the systematic structure in terms of the logit transformation. The result will be a generalized linear model with a binomial response and link logit.

2.2 The Binomial Distribution

Consider a case where the response y_i is binary assuming only two values that for convenience are coded as one or zero. For example we could define

$y_i = 1$ .....(1)

if the i^{th} individual pays otherwise 0. We view y_i as a realization of a random variable Y_i that can take the values one and zero with probabilities π_i and 1-π_i, respectively. The distribution of Y_i is called the Bernoulli distribution with parameter π_i and can be written in compact form as

$P\{Y_i = y_i\} = \pi^{y_i}(1-\pi_i)^{1-y_i}$ .....(2)

for y_i=0, 1. Note that if y_i=1, we obtain π_i=1 and if y_i=0 we obtain 1-π_i. It is known that the mean and variance of the Bernoulli distribution is given as $\mu_i = \pi_i$ and $\sigma_i^2 = \pi_i(1-\pi_i)$

Note that the mean and the variance depend on the underlying probabilityπ_i. Any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggests that a linear model that allows predictors to affect the mean but assumes that the variance is constant will not be adequate for

the analysis of binary data. Suppose now that the units under study can be classified according to the factors of interest into k groups in such a way that all individuals in a group have identical values of all covariates. In our case, successful applicants may be classified into different groups in terms of age, occupation, gender, loan term, repayment amount etc. Let n_i denote the number of observations in group i, and let y_i denote the number of units who have the attribute of interest in group i. For example, let y_i=number of defaulters in group i, we view y_i as a realization of a random variable Y_i that takes the values 0, 1... n_i . If the n_i observations in each group are independent and they all have the same probability π_i of having the attribute of interest, then the distribution of Y_i is binomial with parameters π_i and n_i . The probability distribution function of Y_i is given by

$$P\{Y_i = y_i\} = \begin{bmatrix} n_i \\ y_i \end{bmatrix} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \quad \text{.....(3)}$$

for y_i =1,0,..., n_i . Here $\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures in some specific order and the combinatory coefficient is the number of ways of obtaining y_i successes in n_i trials. The mean and variance of Y_i is given as $\mu_i = n_i\pi_i$ and $\sigma_i^2 = n_i\pi_i(1-\pi_i)$ respectively. From a practical point of view, it is important to note that if the predictors are discrete factors and the outcomes are independent, we can use the Bernoulli distribution for the individual zero to one data or the binomial distribution for grouped data consisting of counts of success in each group. The two approaches are equivalent in the sense that, they lead to exactly the same likelihood function and therefore the same estimates and standard error.

2.3 The Logit Transformation

We wish to have the probabilities π_i depend on a vector of observed covariates xi. The simplest idea would be to let π_i be a linear function of the covariates, say

$$\pi = X'_i\beta \quad \text{.....(4)}$$

where β is a vector of regression coefficients. This sometimes is called the linear probability model. This model is often estimated from individual data using ordinary least squares method. One problem with this model is that the probability π_i on the left hand side has to be between zero and one but the linear predictor $X'_i\beta$ on the right hand side can take any real value, so there is no guarantee that the predicted values will be in the correct range unless complex restrictions are imposed on the coefficients. A simple solution to this problem is to transform the probability to remove the range restrictions and model the transformation as a

linear function of the covariates. This is done in two steps. The first step involves defining the odds, defined as

$$odds = \frac{\pi_i}{1-\pi_i}$$

.....(5)

This is the ratio of the probability to its compliment or the ratio of the favorable to unfavorable cases. In some context the language of odds is more natural than the language of probabilities. The odds however, have no ceiling restrictions. Secondly we take logarithms, calculating the logit or log odds as

$$\eta = \log it(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$$

.....(6)

which has the effect of removing the floor restriction. It is worth noting that as the probability goes down to zero the odds approach zero and the logit approaches $-\infty$. At the other extreme, as the probability approaches one, the odds approach $+\infty$ and so does the logit, thus the logit map probabilities from the range (0, 1) to the entire real line. Note that if the probability is $\frac{1}{2}$, the odds are even and the logit is zero. Negative logit represent probabilities below one half and positive logit correspond to probability above one half. The logit transformation is one to one. The inverse transformation is sometimes called the antilog, and allows us to go back from logit to probabilities. Solving for π_i in equation 6, gives

$$\pi_i = \log it^{-1}(\eta_i) = \frac{e^{\eta}}{1+e^{\eta}} \quad \text{.....(7)}$$

We are now in a position to define the logistic regression model by assuming that the logit of the probability π_i rather than the probability itself follows a linear model.

2.4 The Logistic Regression Model

Suppose that we have k independent observations $y_1, y_2, ...y_k$ and that the i^{th} observation can be treated as a realization of a random variable Y_i. We assume that Y_i has a binomial distribution

$$Y_i \sim B(n_i, \pi_i) \quad \text{.....(8)}$$

with binomial denominator n_i and probability π_i. Suppose further that the logit of the underlying probability π_i is a linear function of the predictors

$$\log it(\pi_i) = X'\beta \quad \text{.....(9)}$$

where X_i is a vector of covariates and β is a vector of regression coefficients.

This defines the systematic structure of the model. The model defined in equations 8 and 9 is a generalized linear model with binomial response and link logit. Note that incidentally, it is more natural to consider the distribution of the response Y_i than the distribution of the implied error term $Y_i - \mu_i$. The regression coefficients β can be interpreted along the same lines as in linear models, bearing in mind that the left hand side is a logit rather than a mean. Thus β_j represents the change in the logit of the probability associated with a unit change in the j^{th} predictor holding all other predictors constant. Exponentiating equation 9 we find that the odds for the i^{th} unit are given by

$$\frac{\pi_i}{1-\pi_i} = \exp(X'_i\beta) \quad \text{.....(10)}$$

This expression defines a multiplicative model for the odds. Solving for the probability π_i in the logit model in equation 9 gives the more complicated model

$$\pi_i = \frac{\exp(X'_i\beta)}{1+\exp(X'_i\beta)} \quad \text{.....(11)}$$

Equation 11 can be written in a more familiar form as

$$f(y) = \frac{e^z}{1+e^z} \quad \text{.....(12)}$$

where z is the logit of y defined as

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad \text{.....(13)}$$

In summary, the logistic regression model relates the log of the odds to the explanatory variables. The logistic regression model further describes the relationship between a dichotomous response variable Y coded to take the values 0 or 1 for success and failure respectively and k explanatory variables $x_1, x_2 ..., x_k$. The explanatory variables can be quantitative or indicator variables referring to the levels of categorical variables. The logit model is preferred for binary data due to the following strengths

(i) It is easy to estimate due to the functional form of the logistic distribution.

(ii) It can be motivated as a model of choice between alternatives with random utilities where the randomness comes from independent data drawn from a Weibull distribution, McFadder, 1974

(iii) It gives rise to a linear log-odds ratio which leads to a simple interpretation of the parameters

2.5 Parameter Estimation

Since we have identified logistic regression as our desired model, there is a need to obtain precise estimate of all the parameters of the model before it is fitted. The problem of estimation is how to achieve the precision of the estimates. The main approaches to estimation are parametric and non parametric. Non parametric estimation is used when the family from which the data is drawn is unknown. Parametric estimation is considered when it is assumed that the data is drawn from one of a known parametric family of distributions. The commonly used methods in parametric estimation are the maximum likelihood in which the estimator β is chosen to maximize the likelihood function and the least squares method which minimize the sum of squares of the residuals. In our study we have used the maximum likelihood method.

The equation to be maximized can be written as:

$$\prod_{i=1}^{n}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)^{n_i} \quad \text{.....(14)}$$

We know that the logistic regression model equates the logit transform, the log odds of the probability of a success, to the linear component as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{k=0}^{K}x_{ik}\beta_k \quad \text{.....(15)}$$

Exponentiating both sides of equation (15) give

$$\left(\frac{\pi_i}{1-\pi_i}\right)=\exp\left(\sum_{k=0}^{K}x_{ik}\beta_k\right) \quad \text{.....(16)}$$

which after solving for π_i becomes,

$$\pi_i=\left(\frac{\exp\left(\sum_{k=0}^{K}x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K}x_{ik}\beta_k\right)}\right) \quad \text{.....(17)}$$

Substituting equation (16) for the first term and equation (17) for the second term, equation (14) becomes

$$\prod_{i=1}^{n}\left(\exp\left(\sum_{\substack{k=0\\ i=1}}^{K}x_{ik}\beta_k\right)\right)^{y_i}\left(1-\frac{\exp\left(\sum_{i=1}^{K}x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K}x_{ik}\beta_k\right.}\right)^{n_i} \quad \text{.....(18)}$$

If we replace 1 with $\frac{1+\sum x\beta}{1+\sum x\beta}$ to simplify the second product, equation (18) can now be written as:

$$\prod_{i=1}^{n}(\exp(y_i \sum_{k=0}^{K} x_{ik}\beta_k)(1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k))^{-n_i} \quad \text{.....(19)}$$

This is the kernel of the likelihood function to maximize. Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log likelihood function and vice versa. Therefore, we take natural log of equation (19) to simplify its differentiation process and get the log likelihood function:

$$\ell(\beta)=\sum_{i=1}^{n} y_i(\sum_{k=0}^{K} x_{ik}\beta_k)^{-n_i}.\log(1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k)) \quad \text{.....(20)}$$

To find the critical points of the log likelihood function, set the first derivative with respect to each β equal to zero. In differentiating equation (20), note that

$$\frac{\partial}{\partial\beta_k}\sum_{k=0}^{K} x_{ik}\beta_k = x_{ik} \quad \text{..... (21)}$$

since the other terms in the summation do not depend on β_k and can thus be treated as constants. In differentiating the second half of the equation (20) taking note of the general rule that $\frac{\partial}{\partial x}\log y = \frac{1}{y}\frac{\partial y}{\partial x}$ with respect to each β_k,

$$\frac{\partial\ell\beta}{\partial\beta_k} = \sum_{i=1}^{n} y_i x_{ik} - n_i\left(\frac{1}{1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k)}\right)\frac{\partial}{\partial\beta_k}\left(1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k)\right)$$

$$= \sum_{i=1}^{n} y_i x_{ik} - n_i\left(\frac{1}{1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k)}\right)\exp(\sum_{k=0}^{K} x_{ik}\beta_k)\frac{\partial}{\partial\beta_k}\exp(\sum_{k=0}^{K} x_{ik}\beta_k)$$

$$= \sum_{i=1}^{n} y_i x_{ik} - n_i \left(\frac{1}{1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k)} \right) \exp(\sum_{k=0}^{K} x_{ik}\beta_k)(x_{ik})$$

$$= \sum_{i=1}^{n} y_i x_{ik} - n_i \pi_i x_{ik} \quad \text{.....(22)}$$

The maximum likelihood estimates for $\beta, \hat{\beta}$ can be found by setting each of the K+1 equations in equation(22) equal to zero and solving for each β_k. Each such solution, if exists, specifies either maximum or minimum a critical point. The critical point will be a maximum if the matrix of the second partial derivatives is negative definite; every element on the diagonal matrix is less than zero. This matrix forms the variance-covariance matrix of the parameter estimates. The general form of the matrix of second derivatives from equation (22) is given by

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_{k'}} = \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^{n} y_i x_{ik} - n_i \pi_i x_{ik}$$

$$= \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^{n} -n_i \pi_i x_{ik}$$

$$= -\sum_{i=1}^{n} -n_i \pi_i x_{ik} \frac{\partial}{\partial \beta_{k'}} \frac{\exp(\sum_{k=0}^{K} x_{ik}\beta_k)}{1+\exp(\sum_{k=0}^{K} x_{ik}\beta_k)} \quad \text{.....(23)}$$

By applying the quotient rule and the exponential functions rule, equation (23) can be simplified to give

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_{k'}} = -\sum_{i=1}^{n} n_i \pi_i x_{ik} (1-\pi_i) x_{ik'} \quad \text{.....(24)}$$

3.0 Results

In this study, we considered a set of data consisting of fifteen thousand applicants whose loans were approved within the year 2007 in one of Kenya's commercial Banks (Barclays Bank of Kenya). The data contained 1,558 defaulters and 13,442 non defaulters.

3.1 Significant Factors

The factors under study were; age, gender, occupation, amount of loan, salary, marital status and term of loans. The variable occupation was further classified into several sectors. The variable of interest was loan status coded as 1 for defaulters and 0 for non defaulters. Using the R programme, the following results were obtained.

Results show age of applicant, gender, occupation (support staff, management and finance), marital status and term of loan are statistically significant at five percent. This means that they influence the applicants' ability to repay their loans. Using these statistically significant factors, we can construct our model as follows;

$$P(Y=1)=\frac{\exp(-0.566-0.024x_1-0.645x_2+0.431x_3+0.683x_4+0.5x_5+0.39x_6-0.021x_7)}{1+\exp(-0.566-0.024x_1-0.645x_2+0.431x_3+0.683x_4+0.5x_5+0.39x_6-0.021x_7)}$$

where $x_1, x_2, x_3, x_4, x_5, x_6$ and x_7 represents age of applicant, gender(coded as 1 for female, 0 for male),finance, management, support staff, marital status(coded as 0 for married,1 for singles) and term of loan respectively.

Table1

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-5.663e-01	2.626e-01	-2.156	0.03105 *
Age	-2.401e-02	4.028e-03	-5.959	2.53e-09 ***
Gender(1)	-6.452e-01	7.776e-02	-8.297	< 2e-16 ***
Armed Officers	4.762e-01	2.731e-01	1.744	0.08121
Business	1.990e-01	1.881e-01	1.058	0.29019
CIVIL SERVANT	-2.090e-01	2.317e-01	-0.902	0.36708
CLERGY	2.236e-01	3.932e-01	0.569	0.56969
Education	-4.997e-05	1.916e-01	-0.000261	0.99979
FARMER	1.342e-01	2.076e-01	0.647	0.51791
FINANCE	4.309e-01	2.173e-01	1.983	0.04739 *
Management	6.826e-01	2.395e-01	2.851	0.00436 **
Professionals	3.574e-01	2.079e-01	1.719	0.08559
SUPPORT STAFF	5.006e-01	1.898e-01	2.637	0.00837 **
Amount of loan	8.727e-08	1.188e-07	0.734	0.46270
Salary	1.417e-06	9.790e-07	1.447	0.14790
Marital status (1)	3.904e-01	1.266e-01	3.084	0.00204 **
Term of loan	-2.103e-02	2.661e-03	-7.901	2.76e-15 ***

From the above model (and calculations tabulated in table two and three below) we can deduce that, younger applicants tend to default more compared with their relatively aged colleagues, male customers default more than female customers and applicants in the finance sector, management levels and support staff have relatively similar probabilities of defaulting, single customers default more than married customers and the longer the term, the less the chance of defaulting.

From the above model, we can calculate the probability of a customer defaulting. For example the probability of a lady who is 30 years, works in a financial sector, unmarried and has repayment term of 36 months defaulting using the below expression is 0.134.

$$P(Y=1)=\frac{\exp(-0.566-0.024*30-0.645+0.431+0.39-0.021*36)}{1+\exp(-0.566-0.024*30-0.645+0.431+0.39-0.021*36)}$$

By making the relevant substitutions, the same method can be employed to calculate the probability of default of a male customer who is single, aged 30, and works in the financial sector which is 0.228.

Similarly, we can calculate the probability of married customers (male and female) aged 30 years working in the financial sector as 0.166 and 0.095 respectively. These can be summarized in table 2 as follows

Table 2

Gender	**age**	**sector**	**Marital status**	**Probability of default**
Female	30	finance	single	0.134
Female	30	finance	Married	0.095
Male	30	finance	Single	0.228
Male	30	finance	Married	0.166

The probabilities of a customer defaulting based term can also be calculated by amending the model equation accordingly. Results show that the probability of default decreases as the term increases. This was calculated holding age of the customer (at 30) marital status (married) and sector (finance) constant and is summarized in table three below;

Table 3

Term of loan (in months)	**Probability of default**
5	0.277
10	0.256
15	0.237
20	0.218

25	0.201
30	0.185
36	0.166

The odds ratio of the x_i with dichotomous values can be obtained as the exponential of the coefficients. The odds ratio of x_i increases as its probability increases. In our model, the odds of a male defaulting is 1.91 times that of the female while that of a single customer defaulting is 1.48 times greater than that of a married customer.

3.2 The goodness of Fit of the Model

To test the goodness of fit of the fitted model, we compare our null deviance (the errors without the parameters) which is 9128.2 to residual deviance (errors with parameters) which is 8852.9 since the residual deviance is less than the null deviance; our model is a better fit.

4.0 Summary

Parametric estimation and logistic regression are powerful tools for researchers. The popularity of these tools results from their versatility and relative ease of interpretation. The goal of this paper was to construct a parametric logistic model that predicts the probability of a customer defaulting. This has been achieved; however, the disadvantage of parametric estimation is its reliance on functional form assumptions which lead to inconsistent estimators if the model is not correctly specified. Therefore when the particular family of distribution is unknown, parametric estimation becomes limited. In such a case, non parametric estimation becomes more appealing as it does not rely on distribution assumptions. Excellent literature exists on the mathematical basis of non parametric logistic modeling as well as its application and interpretation. This is a motivation for a research on non parametric modeling of probability of bank loan default.

References

Altman E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance,* **23**, pp 589-609.

Beaver W. H. (1968). Financial ratios as predictors of failure. *The Journal of Accounting Research,* **4**, pp 71-111.

Froelich M. (2006). Non-parametric regression for binary dependent variables. *Econometrics Journal,* **9**, pp 511-540.

Gourieroux C., and Monfort, A. (1981) Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics,* **17**,pp 83-97.

Jeong M. G. (2010) Non-parametric regression for binary dependent variables. *Int. J. Contemp. Math. Sciences*, **5**, pp 201-207.

Kolari J., D. Glennon H. S. and Caputo, M. (2002). Predicting large us commercial bank failures. *Journal of Economics and Business,* **54(4),** pp 361-387.

Lawrence C. L., L. S. and Rhoades M. (1992). An analysis of default risk in mobile home credit. *The Journal of Banking and Finance,* **16(3),** pp 299-312.

Martin D. (1977). Early warning of bank failure: A logit regression approach. *The Journal of Banking and Finance,* **1**, pp 449-470.

Ohlson J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *The Journal of Accounting Research,* **19**, pp 109-301.

Soest A. H. O., Van A. P. and Karminsky A. (2003). An analysis of ratings of Russian banks. *Tilburg University CentER Discussion Paper Series* **2003/85**

Westgaards S. and van der Wijst N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach,. *European Journal of Operational Research*, **135**, pp 338-349.

Wiginton J. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *The Journal of Finance and Quantitative Analysis*, **15(3)**, pp 757-770.