CLASSIFICATION RATES: NON-PARAMETRIC VERSES PARAMETRIC MODELS USING BINARY DATA

A. O. Adem[1], A. W. Gichuhi[2] and R. O. Otieno[3]
[1] Mathematics and Physics Department, Technical University of Mombasa
[2,3]Statistics and Actuarial Science Department, Jomo Kenyatta University of Agriculture and Technology
Email: aggreyadem@yahoo.com

Abstract
Estimations of the conditional mean and the marginal effects for particular small changes in the covariates have been of interest in financial, economics and even educational sectors. The standard approach has been to specify a parametric model such as probit or logit and then estimating the coefficients by maximum likelihood method. This is only applicable when the distribution form from which the data has been drawn is known. Non parametric methods have been proposed when the functional form assumptions cannot be ascertained. This research sought to establish if non parametric modeling achieves a higher correct classification ratio than a parametric model. The local likelihood technique was used to model fit the data sets. The same sets of data were modeled using parametric logit and the abilities of the two models to correctly predict the binary outcome compared. The results obtained showed that non-parametric estimation gives a better prediction rate (classification ratio) for a binary data than parametric estimation. This was achieved both empirically and through simulation. For empirical results two different data sets were used. The first set consisted of loan applications of customers and the second set consisted of approved loans. In both data sets the classification ratio for non-parametric method was found to be 1 while that for parametric was found to be 0.87 (only 87 out of the 100 observations were correctly classified) and 0.83 respectively. Simulation was done based on sample sizes of 25, 50, 75, 100,150,200,250,300 and 500. The simulated results further showed that the accuracy of both models decrease as sample size increases.

Key words: Parametric, non-parametric, local likelihood, logit, confusion matrix and classification ratio

1.0 Introduction

Regression analysis is one of the most useful and the most frequently used statistical methods (Efron and Tibshirani, 1993). The aim of the regression methods is to describe the relationship between a response variable and one or more explanatory variables. Among the different regression models, logistic regression plays a particular role. The basic concept, however, is universal. The linear regression model is, under certain conditions, in many circumstances a valuable tool for quantifying the effects of several explanatory variables on one dependent continuous variable. For situations where the dependent variable is qualitative, however, other methods have been developed. One of these is the logistic regression model, which specifically covers the case of a binary (dichotomous) response. Cramer (2003) discussed an overview of the development of the logistic regression model. He identified three sources that had a profound impact on the model: applied mathematics, experimental statistics, and economic theory. Agresti (2002) also provided details of the development on logistic regression in different areas. However, logistic regression is widely used as a popular model for the analysis of binary data with the areas of applications including physical, biomedical, and behavioral sciences. For example, Cornfield (1962) presented the preliminary results from the Framingham Study. The purpose of the study was to find the roles of risk factors of cholesterol levels (low versus high values) and blood pressure (low versus high values) in the development of coronary heart disease (yes or no) in the population of the town. The logistic regression model can be easily modified to handle the case in which the outcome variable is nominal with more than two levels (Hosmer and Lameshow, 2000).

An extension of the logistic regression model is called the multinomial logistic regression model, when the categorical dependent outcome variable has more than two levels (Chan, 2004). For example, Zocchi and Atkinson (1999) noted that in their multinomial logistic regression model on the dose level experiment to measure the influence of gamma radiation on the emergence of house flies, three disjoint outcomes occurred: death before the pupae opened, death during emergence, and life after emergence. A modification of the logistic regression model, known as the discrete choice model, was first proposed by McFadden (1974). The model is also known as multinomial or polychotomous logistic regression in the health sciences and as the discrete choice model in econometrics (Breslow and Powers, 1978).

The two main approaches that have been fronted to estimate the coefficients of the explanatory variables are parametric and non-parametric. The maximum likelihood estimation (MLE) is the most widely-used general method of parametric estimation procedures and is treated as a standard approach to parameter estimation and inference in statistics (van der Vaart, 1998). Under very general conditions, maximum likelihood estimates are consistent, asymptotically efficient, and asymptotically-normally distributed. Notice that this normality allows one to

compute the confidence interval and perform statistical tests in a manner analogous to the analysis of linear multiple regression models, provided the sample size is large. A lot of research has been done in maximum likelihood estimation method. Starting with the fundamental work of Gourieroux *et al*. (1981) and Gallant and Tauchen (1996), a variety of simulation-based methods have been recently introduced to consistently estimate parametric models for discussions. For example, Brandtand Santa-Clara (2002), Durham and Gallant (2002), Elerian *et al*. (2001) and Eraker (2001) among others, suggested simulation-based procedures for maximum likelihood estimation. Somewhat discrete is the approach in Aït-Sahalia (2002) who recommends approximations to the true, generally unknown, transition density of the discretely sampled process for the purpose of consistent likelihood estimation. Carrasco *et al*. (2002), Chacko and Viceira (2003), Jiang and Knight (2002), and Singleton (2001) suggested characteristic function based generalized method of moment (GMM) estimation. GMM-based estimation is also discussed in Conley et al. (1997), Duffie and Glynn (2004) and Hansen and Scheinkman (1995).

Adem, Gichuhi and Odhiambo (2012) applied the parametric approach to estimate the probability of loan default in Kenya's Commercial banks.

Scholars have proposed the non parametric approach to cases where the parametric approach has shown limitations. Tibshirani and Hastie (1987) introduced the concept of local likelihood estimation. Properties of local likelihood have been analyzed in Fan *et al*. (1995), Fan and Gijbels (1996) and Fan *et al*. (1998). Staniswalis (1989) considered kernel smoothers maximizing a kernel weighted likelihood function. Cessie and Houweingen (1991) proposed a test for a logistic model by kernel smoothing with an adhoc bandwidth. Royston (1992) proposed a test based on cumulative sums of residuals to be aware of lack of linearity in the logit of probability function, where large cumulative sums in absolute value at any point support the evidence of lack of fit. Fan, Heckman and Wand (1995) introduced local polynomial estimators in one-parameter exponential family and quasi-likelihood models. Recently, many other applications have been introduced and studied. For relevant references on this subject we refer to the books of Wand and Jones (1995) and Fan and Gijbels (1996). Hart (1997) provided a goodness of fit test from the variance ratio point of view.

Gozalo and Linton (2000) illustrated non parametric regression with a binary dependent variable Y and unrestricted interactions among regressors X. Pulustein and Robinson (2002) developed a test analogous to Pearson Chi squared and deviance statistics with a modification of continuous covariates. Racine and Li (2004) suggested hybrid product kernel that coalesces continuous and discrete regressors.

Fr$\ddot{O}$lich (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 is found that, the parametric logit and semi parametric Klein-Spady estimators do not detect heterogeneity. Kuo-Chin and Yi-Ju (2005) chose test statistics based on non-parametric local linear regression technique with optimal bandwidth chosen by a cross validation method to logistically fit with continuous and categorical covariates. Gourieroux and Jasiak (2010) introduced a local likelihood method of value at risk computation for univariate or multivariate data on portfolio returns. The approach relies on a local approximation of the unknown density of returns by means of a mis-specified model. The method allows one to estimate locally the conditional density of returns and to find the local conditional moments, such as a tail mean and tail variance.
The purpose of this paper is to investigate the classification rate of non-parametric methods using a binary data as compared to parametric method.

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 a 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 the Bernoulli and binomial distributions. To get more information in the stochastic structure of the data in terms of the Bernoulli and binomial distributions, and the systematic structure interms of the logit transformation see Adem, Gichuhi and Odhiambo (2012).

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

$$Y_i \sim B(n_i, \pi_i) \quad \text{..... (1)}$$

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

$$Logit(\pi_i) = X_i'\beta \quad \text{..... (2)}$$

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 (1) and (2) is a generalized model with binomial response and link logit. More often we 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. The logit of a function can also be defined as the log of odds ratio which can be expressed as

$$Logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i'\beta \quad \text{..... (3)}$$

Taking antilog on both sides enable us to define the odds for the i^{th} unit as

$$\frac{\pi_i}{1-\pi_i} = \exp(X_i'\beta) \quad \text{..... (4)}$$

This expression defines a multiplicative model for the odds. Solving for the probability π_i in the logit model in equation (4) gives the logistic model

$$\pi_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \quad \text{..... (5)}$$

Equation (5) can simplified as

$$f(y) = \frac{e^z}{1+e^z} \quad \text{..... (6)}$$

Where z the logit of y is defined as

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad \text{..... (7)}$$

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 $x_1, x_2, \ldots x_k$. 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

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

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, Mc Fadden, 1974.

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

Once a model has been selected, there is a need to estimate parameters. Assume that we are interested in estimating the conditional mean and the marginal effects for particular small changes in the covariates. The standard approach proceeds by specifying a parametric model for example, a probit or logit model, estimating the coefficients by maximum likelihood and then computing the conditional mean and marginal effects. The disadvantage of parametric estimation is its reliance on functional form assumptions which lead to inconsistent estimators if the model is not correctly specified, Fr$\ddot{o}$lich (2006). Therefore when the particular family of distributions is unknown, parametric estimation becomes limited. In such a case a non-parametric approach becomes more appealing.

To relax some of the parametric assumptions, several semi parametric estimators have been suggested. A single index restriction is often invoked which assumes some unknown function and θ as a co efficient vector. A number of $\sqrt{n}$ consistent estimators of θ have been developed including iterative methods such as Han (1987), Ichimura(1993), Klein and Spady (1993) and non-iterative methods such as the average derivative estimators of Hardle and Stocker (1989), Powell et al (1989), Stocker (1991) and Horowitz and Hardle (1996). Although less restrictive than the parametric models, a semi-parametric model restricts interactions between the regressors which is less appealing for many applications where heterogeneity in responses is often considered important, Heckman et al, (1997). Non parametric estimation is the most flexible among the three. Although it is subjective to the curse of dimensionality and does not achieve $\sqrt{n}$ convergence, it may still perform well in finite samples, Fr$\ddot{o}$lich (2006).

In non-parametric estimation we use a maximum kernel local weighted likelihood estimator to eliminate the restrictive assumptions of the parametric estimation. The most popular class of non-parametric estimators is the polynomial regression which however is not particularly suited for binary models as it doesn't incorporate the restriction $E(Y/X) \in [0,1]$, Fan and Gijbels (1996). An immediate solution is to cap the estimates at 0 and 1, which however makes the objective function non-differentiable and also implies that the estimated marginal effects may be exactly zero at some X values.

To handle binary data when the distributional form is unknown, we use local Maximum likelihood estimation which is based on the idea of local fitting, Tibshirani and Hastie (1987).

2.2 Local Likelihood

Consider a response variable Y, having a probability density function $f(y;\theta_1(x),\theta_2(x),...,\theta_v(x))$ involving v parameters depending on an explanatory variable $x=(x_1,x_2,...,x_d)$. In this generalized regression setting, the form of f is assumed to be known and the v parameters are unknown real valued functions of the covariate-vector X. In parametric maximum likelihood estimation, the functions $\theta_\ell(x)$ are modeled globally in terms of a finite number of regression coefficients. Instead of assuming a specific functional form for each $\theta_\ell(x)$, one can allow the data to describe this relationship non-parametrically, only requiring some weak smoothness assumptions. The basic idea of local likelihood is a simple extension of the local fitting technique where smoothing ideas is applied to data in which the relationship can be expressed through a likelihood function. Consider n independent realizations $y_1,y_2,...,y_n$ of the random variable Y with $Y_i \sim f(Y,\theta)$, for $i=1,2,...,n$. The function likelihood is given by

$$L(\theta_1,\theta_2...,\theta_n = \prod_1^n f(y_i,\theta_i) \quad \text{..... (8)}$$

A standard modeling procedure would assume a simple parametric form for the $\theta_i's$. If we assume that $\theta_i = \beta_0 + \beta_1 x_i$, then the parsimonious covariates form can be replaced with an unspecified smooth function $\varphi(x_i)$; $\theta_i = \varphi(x_i)$

$\varphi(x_1),\varphi(x_2),...,\varphi(x_n)$ Can be estimated by maximizing $L(\varphi(x_1),\varphi(x_2),...,\varphi(x_n))$. However this would result in an unsatisfactory estimate

due to over fitting. Alternatively suppose that the function $\varphi(x_i)$ has a $(p+1)^{th}$ continuous derivative at the point x_i. For data point x_j in a neighborhood of x_i, $\varphi(x_j)$ can be approximated via a Taylor series expansion by a polynomial of degree p

$$\varphi(x_j) \approx \varphi(x_i) + \varphi'(x_i)(x_j - x_i) + ... \frac{\varphi^{(p)}(x_i)}{p!}(x_j - x_i)p \equiv \chi^T \mathrm{B} \quad \text{..... (9)}$$

Where $\chi = (1, x_j - x_i ..., (x_j - x_i)^p)^T$ and $\mathrm{B} = (\beta_0, \beta_1, ..., \beta_p)^p$ the contribution to the log likelihood for data points (x_j, y_j) in the neighborhood of x_j is $\ell(y_j, x^T\beta)$ weighted by w_j. This leads to the local likelihood or local kernel-weighted by log – likelihood as proposed by Fan *et a*l, 1998;

$$L(\quad L(\mathrm{B} \setminus x_i) = \sum_{j=1}^{n} \ell(y_j, x^T\beta) w_j \quad \text{..... (10)}$$

Maximizing the local log likelihood with respect to B gives the vector of the estimates $\hat{\mathrm{B}} = \left(\hat{\beta}_0 - \hat{\beta}_p\right)^T$ .One of the components of the local likelihood is the smoothing technique. Several smoothing methods exists, in this paper we chose the kernel based method.

2.3 Kernel

Kernel regression is an estimation technique used to fit data when estimating a regression function or a probability density function where the underlying assumptions about the distribution function are unknown. The idea of kernel regression is putting a set of identical weighted functions called kernel (kind of a bump function) to each observational data point. The kernel basis function depends only on the radius or width (variance) from local data point X to a set of neighboring locations x. Kernel regression is a superset of a local weighted regression and closely related to moving average and K nearest neighborhood, radial basis function, neural network and support vector machine. Given a random sample $x_1, x_2, ..., x_n$, with a continuous univariate density f ,

$$\hat{f} = \lim_{h \to 0} \frac{1}{2h} \Pr(x - h < X < x + h) \quad \text{..... (11)}$$

For any given h, we can estimate $P(x-h<X<x+h)$ by the proportion of the sample falling in the interval $(x-h,x+h)$. Thus a natural estimator of $\hat{f}$ is given by choosing a small h and setting,

$\hat{f}=\frac{1}{2hn}$ [No of $x_1,x_2,...,x_n$ falling in $(x-h,x+h)$]..... (12)

This is called the naive estimator. To express the estimator more compactly, we define the weight function w by

$$w(x)=\begin{cases}\frac{1}{2}, if|x|<1\\ \text{(Otherwise zero elsewhere)}\end{cases} \quad \text{..... (13)}$$

We can then rewrite the naïve estimator as

$$\tilde{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left[\frac{x-x_i}{h}\right] \quad \text{..... (14)}$$

It follows that, the estimate is constructed by placing a 'box' of width $2h$ and height $(2nh)^{-1}$ on each observation and then summing to obtain the estimate. The estimator (naïve) is not wholly satisfactory from the point of view of using density estimates for presentation. It follows from the definition that $\tilde{f}$ is not a continuous function but has jumps at the points $x_i \pm h$ and has zero derivatives everywhere else. This gives the estimates a somewhat ragged character, which is undesirable and could also provide a misleading impression. We generalize the naïve estimator and replace the weight function w by a kernel function k, which satisfies the condition,

$$\int k(x)dx=1 \quad \text{..... (15)}$$

Usually (but not always), k will be a symmetric probability density function. The kernel estimator is therefore defined by

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}k\left[\frac{x-X}{h}\right] \quad \text{..... (16)}$$

Where h is the window width also known as the smoothing /span parameter or bandwidth. The kernel estimator can be considered as a sum of bumps placed at the observations. The kernel function k determines the shape of the bumps while the window width h determines their width. If the kernel k is everywhere non-negative and is a probability density function, then it follows that $\hat{f}$ will be a density function and inherits all the continuity and differentiability properties of the kernel k. Under mild conditions (h must decrease with n), the kernel estimate converges in

probability to the true density. There are two basic problems to consider in obtaining $\hat{f}(x)$; the choice of the kernel function and the smoothing parameter h.

The Choice of kernel function will depend on practical and theoretical considerations. The performance of a kernel is measured by the mean integrated square error (MISE) or asymptotic mean integrated square error. The performance is evaluated by the distance between the true density f and the estimator $\hat{f}$.

$$MISE(h) = E\left[\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right] \quad \text{..... (17)}$$

The Epanechnikov kernel minimizes asymptotic mean integrated error the most and is therefore considered optimal. The efficiency of other kernels is measured relative to the Epanechnikov kernel. Though the Gaussian kernel is not as efficient as the Epanechnikov, we have used it in this research as it is symmetrical and assumes a probability density function.

Since we have seen that $\int k(x)dx = 1$ and that k is everywhere non negative, then $\hat{f}$ will inherit all the continuity and differentiability properties of k, the choice of the kernel function should be based on the degree of differentiability and or the computational efforts involved. Assumption about the kernel k is that, it is a symmetric function satisfying the conditions;

$$\int k(u)du = 1 \quad \text{..... (18)}$$

$$\int uk(u)du = 0 \quad \text{..... (19)}$$

$$\int u^2 k(u)du = p \neq 0 \quad \text{..... (20)}$$

And that, the unknown density f has continuous derivatives of all orders required. Usually the kernel k will be a symmetric probability density function such as the normal density and the constant p will be the variance of the distribution with this density.

Alongside the kernel, the other tool in the smoothing technique is the bandwidth which controls the size of the local neighborhood and whose choice is more important than that of the kernel. This choice should be made such that, there is neither over smoothing nor under smoothing. Small choice of h leads to an estimator with small bias and large variance making the estimate look 'wiggly' and show spurious features. Large choice of h leads to small variance at the expense of large bias and the resultant estimate will be too smooth thus at times not revealing structural features like bimodality of the underlying density f. The appropriate choice of the smoothing parameter will be influenced by the purpose for which the

density estimated is to be used. If the purpose of the density estimation is to explore the data in order to suggest possible models and hypotheses, then it will probably be quite sufficient and indeed desirable to choose the smoothing parameter subjectively.

However, many applications require an automatic choice of smoothing parameter especially if the density estimation is to be used routinely on a large data set or as part of a larger procedure. The choice of the optimal bandwidth is still a big dispute. The main argument is on whether one should use the integrated squared error or the mean integrated squared error to define optimal band width. A lot of research has been done to develop bandwidth selection methods which try to estimate the optimal bandwidth obtained by either of these criteria, *Wand et al*, 1995.

3.0 Local Logit Estimation

For binary choice models, the local logit estimator

$$E[Y/X=x]=\frac{1}{1+\exp(-x'\theta_x)} \quad \text{.....(21)}$$

is convenient, since it imposes the range restriction and is differentiable. For a further discussion see Fan, Heckman, and Wand (1995).If the logit form is closer to the true regression curve than a constant or linear specification, the local logit estimator will be less biased than kernel or local linear regression. Local logit encompasses the global logit model (where θ_x does not vary with x) and if the global logit model were indeed correct, local logit would be unbiased, see Gozalo and Linton (2000).

With $g(x,\theta_x)$ as the local model, the conditional mean is estimated as

$$\hat{E}[Y/X=x]=g(x,\theta_x) \quad \text{.....(22)}$$

Several approaches to estimate $\hat{\theta}_x$ have been suggested, including local least squares
(Gozalo and Linton (2000)), local likelihood, (Tibshirani and Hastie (1987)) and local estimating equations (Carroll, Ruppert, and Welsh (1998)). Local least squares estimates $\hat{\theta}_x$ from a sample of $n\ iid$ observations $\{(Y_i,X_i)\}_{i=1}^n$ are given by

$$\hat{\theta}_x=\arg\min_{\theta_x}\sum_{i=1}^{n}(Y_i-g(X_i,\theta_x))^2.K_H(X_i-x) \quad \text{.....(23)}$$

Where $K_H(X_i-x)$ is a kernel function and H a vector of bandwidth values. Local likelihood estimates $\hat{\theta}_x$ are given by

$$\hat{\theta}_x = \arg\max_{\theta_x} \sum_{i=1}^{n} \ln L\left(Y_i, g\left(X_i, \theta_x\right)\right).K_H\left(X_i - x\right) \quad \text{..... (24)}$$

Where $\ln L\left(Y_i, g\left(X_i, \theta_x\right)\right)$ is the log-Likelihood contribution of observation $\left(X_i, Y_i\right)$
For H converging to infinity, the local neighborhood widens and the local estimator would converge to the global parametric estimator, Froelich (2006).
Exciting rich literatures exist on the consistency and asymptotic normality of the local likelihood estimator for binary data notable by Marc Aerts and GerdaClaeskens (1997)

3.1 Confusion Matrix

This is a contingency table used in predictive analytics to allow visualization of the performance of an algorithm or a system. It contains information about actual and predicted classifications done by a predictive system. Performance of such a system is commonly evaluated using the data in the matrix. Classification matrix is used mainly to assess the accuracy of a classification system. The table below has been used to demonstrate how to obtain the classification matrix for a two class classifier. A lot of literature on this exists on Kohavi and provost (1998) and Landis and Koch (1977) . Consider a finite set of binary data where the dependent variable has been coded as 0 and 1.

Table 1

	Predicted		
Actual		0	1
	0	W	X
	1	Y	Z

From the table above, the accuracy of the system can be computed as the proportion of the predictions that were correctly given as

$$Accuracy = \frac{W + Z}{W + X + Y + Z} \quad \text{..... (25)}$$

The proportions of zeros and ones that were correctly classified are given by equations (26) and (27) respectively

$$T(0) = \frac{W}{W + X} \quad \text{..... (26)}$$

$$T(1) = \frac{Z}{Y+Z} \quad \text{..... (27)}$$

Equations (28) and (29) give the proportions of zeros and ones that were incorrectly classified respectively

$$F(0) = \frac{X}{W+X} \quad \text{..... (28)}$$

$$F(1) = \frac{Y}{Y+Z} \quad \text{..... (29)}$$

The performance of a confusion matrix is measured by its precision. The precision of a confusion matrix is the proportion of the predicted 1's (where one denotes the success cases) cases that were correctly classified and is computed by the equation, Kohavi and John (1997).

$$P = \frac{Z}{W+Z} \quad \text{..... (30)}$$

3.2 Simulation Results

To investigate the classification rate of parametric and non-parametric estimation methods for binary data, we conducted a simulation study using a parametric logistic regression model. In our simulation study, we considered five explanatory variables $x_1, x_2, \ldots . x_5$, which are fixed and the binary response variable y, which is treated as a random variable in the logistic model. For fixed values of the intercept parameter β_0 and five other parameters $\beta_1, \beta_2, \ldots \beta_5$ we wish to compare the prediction rate of the parametric estimation method and the non-parametric estimation method using simulated data of sizes 25,50,75,100,150,200,250,300 and 500. For parametric method, ordinary logistic regression model was used and the estimates obtained through maximum likelihood method for a sample size of 100 is given in table1 below.

Table 2

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.6097	0.2894	2.107	0.0351
x_1	1.1068	0.3586	3.087	0.002024
x_2	0.3836	0.3087	1.252	0.2107

x_3	1.2761	0.3409	3.7409	0.000182
x_4	0.4701	0.2496	1.884	0.0596
x_5	1.2652	0.3262	3.878	0.000105

Using Gaussian method with a data driven bandwidth selection, we non-parametrically fitted our data. Cross validation method was used to select the bandwidths. The table shows the bandwidth for each of the variable (for sample size 100)

Table 3

Variables	x_1	x_2	x_3	x_4	x_5
Fixed Bandwidth	0.8213	0.1072	0.1237	0.2398	0.4123

3.3 The Confusion Matrix

A 'confusion matrix' is a tabulation of the actual outcomes versus those predicted by a model.

The diagonal elements contain correctly predicted outcomes while the off-diagonal ones contain incorrectly predicted (confused) outcomes. The confusion matrices obtained through parametric and non-parametric methods are given below:

Confusion matrix (Parametric)

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 25 & 11 \\ 4 & 60 \end{bmatrix} \end{array}$$

This implies that the overall classification rate for a sample of size 100 is 85%.i.e. out of a 100 simulated values, only 85 were correctly classified as 0's and 1's.

Confusion matrix (non-parametric)

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 32 & 0 \\ 0 & 68 \end{bmatrix} \end{array}$$

The overall classification rate is 100%. All the 0's and all the 1's were correctly classified. The correct classification ratio by outcome for non-parametric method will be 1 for either case (both 0 and 1). A summary of the overall classification rate for various sample sizes is given in the table 4 below

Table 4 (overall classification rate)

Sample size	Parametric model	Non parametric model
25	0.88	1
50	0.88	1
75	0.87	1
100	0.85	1
150	0.84	1
200	0.84	0.96
250	0.83	0.89
300	0.833	0.87
500	0.826	0.866

From the simulated data it has been shown that non parametric estimation method gives a better prediction rate than the parametric estimation method.

4.0 Empirical Results

Under empirical results, we considered two sets of data. The first set of data consisted of 37,609 loan applicants out of whom 31,805 were approved and 5,804 were declined. The explanatory variables were age, gender (coded as 1 for female, 0 for male), occupation, amount of loan, salary, marital status (coded as 1 for married, 0 for singles) and term of loans. The variable occupation was further classified into several sectors. The response variable was loan application status coded as 1 for approved loan application and 0 for declined loan applications. Out of this data a random sample of size 100 was selected consisting of 86 loan applications which were approved and 14 loan applications which were declined.

The second set of data consisted of 15,000 applicants whose loans were approved within the year 2007 in one of the Kenya's commercial Banks. The data contained 1558 defaulters and 13442 non defaulters. The explanatory variables were age, gender (coded as 1 for female, 0 for male), occupation, amount of loan, salary, marital status (coded as 1 for married, 0 for singles) and term of loans. The variable occupation was further classified into several sectors. The response variable was

loan status coded as 1 for defaulters and 0 for non-defaulters. Out of this data a random sample of size 100 was selected consisting of 15 defaulters and 85 non defaulters. Using ordinary logistic regression model (equation 4) the estimates of the variables for both cases are given in table 3 and table 4 respectively.

Table 6: (loan application approvals)

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-7.311e-01	1.151e-01	-6.350	2.15e-10
Age	4.748e-02	2.085e-03	22.775	<2e-16
Gender(1)	6.213e-01	3.790e-02	16.394	<-2e-16
Business	-1.310e01	1.970e+02	-0.066	0.9469
CIVIL SERVANT	-2.735e-01	8.116e-02	-3.370	0.000751
CLERGY	-1.246	2.605e-01	-4.7831	1.73e-06
Education	3.738e-01	7.228e-02	5.172	2.32e-07
FARMER	9.571	8.492e+01	0.113	0.910262
FINANCE	-3.321e-01	9.155e-02	-3.628	0.000286
Management	-1.102	8.738e-02	-12.612	<2e-16
Amount of loan	-3.96e-07	7.97e-08	-4.969	6.73e-07
Salary	5.229e-0.06	6.614e-07	7.906	2.66e-15
Marital status (1)	1.999e-01	5.338e-02	3.745	0.000181
Term of loan	1.366e-02	1.633e-03	8.366	2e-16

Table 7 (default and non-default cases)

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
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
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

The classification matrix for parametric estimation method for the two sets are given below

Confusion matrix (approved loan applications verses declined applications)

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 3 & 11 \\ 2 & 84 \end{bmatrix} \end{array}$$

The overall classification ratio is 0.87.

Confusion matrix (default and non-defaults)

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 82 & 3 \\ 14 & 1 \end{bmatrix} \end{array}$$

The overall classification ratio is 0.83.

To fit our model non-parametrically we used the Gaussian kernel and our bandwidth selection was data-driven. Using cross-validation bandwidth selection, we obtained the following bandwidths for each of the variables for both sets of data;

Table 9: (showing bandwidth; approved and declined loans applications)

Variables	Status	Age	Amount	Gender	Marital	Occupation	Salary	Term
Fixed Bandwidth	0.100	0.27079	6938	0.02348	0.46	0.8571	16882	5.26

Table 10 (showing bandwidth; defaults and non defaults)

Variables	Status	Age	Amount	Gender	Marital	Occupation	Salary	Term
Fixed Bandwidth	0.07794929	6.595547	40182.51	0.3287189	0.5	0.1482192	10397.49	4.141859

The confusion matrix for both sets of data are given below

Confusion matrix (approved/declined loan application)

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 15 & 0 \\ 0 & 85 \end{bmatrix} \end{array}$$

Confusion matrix (defaults/non defaults)

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 86 & 0 \\ 0 & 14 \end{bmatrix} \end{array}$$

The overall classification ratio for both models is 1

5.0 Summary and Conclusion

The aim of this research was to establish whether non-parametric methods give a higher classification ratio (correct prediction rate) for binary data than parametric methods. This has been achieved through simulation and empirical results as the confusion matrices obtained in sections 4.2 and 4.3 show that non-parametric estimation gives a better prediction rate (classification ratio) for binary data than parametric estimation. For simulated data, the classification ratio tends to decrease as sample size increases. For the empirical cases, non-parametric method achieved 100% classification rate for both data sets while parametric method classification rate was 0.87 and 0.83 respectively for both data sets. This implies that only 87 of the 100 observations, and 83 of the 100 observations were correctly classified.

Non-parametric methods may not necessarily give a classification ratio of 1. A number of factors may influence this, notably the bandwidth selection process. Examining the influence of bandwidth selection process on non-parametric classification rate may yield ground for further research.

References

Adem, A.,Gichuhi, A. and Otieno, R.(2012) Parametric Modeling of Probability of Bank Loan Default in Kenya. *Journal of Agriculture Science and Technology* **14(1)**, pp. 61-74

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc.

Aït-Sahalia, Y., (2002) Maximum likelihood estimation of discretely sampled diffusions: a closed form approximation approach. *Econometrica* **70**, pp. 223-262

Brandt, M. W., and Santa-Clara P. (2002).Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. *Journal of Financial Economics* **63**, pp. 161-210.

Breslow, N. and Powers, W. (1978). Are There Two Logistic Regressions for Retrospective Studies? *Biometrics*, **34**, pp. 100-105.

Carrasco, M., Chernov, M., Florens J.P., Ghysels, E. (2002).Estimation of jump-diffusions with a continuum of moment conditions.IDEI Working paper, no. 140.

Carroll, R., Ruppert, D., and Welsh, A.(1998) Local Estimating Equations. *Journal of American Statistical Association*, **93**, pp. 214—227

Cessie,S. and Houwelingen, V. (1991). A goodness of fit for binary regression model based smoothing methods. *Biometrics*, **47**, pp. 1267-1282

Chacko, G., Viceira, L. (2003). Spectral GMM estimation of continuous-time processes. *Journal of Econometrics*, **116**, pp. 259-292.

Chan, Y. H. (2004). Multinomial logistic regression. *Singapore Medical Journal*, **46**, pp. 259- 269

Conley, T., Hansen, L.P., Luttmer E., Scheinkman J. (1997) Short-term interest rates as subordinated diffusions. *Review of Financial Studies* **10**, pp. 525-577

Cornfield, J. (1962).Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure. *Federation Proc*., **21**, pp. 58-61

Cox, D. R. (1970).*Analysis of Binary Data.* London: Chapman & Hall

Cramer, J.S. (2003).*Logit Models from Economics and Other Fields*. Cambridge University Press.

Duffie, D., and Glynn P. (2004) Estimation of continuous-time Markov processes sampled at random time intervals. *Econometrica* **72**, pp. 1773-1808

Durham, G., Gallant R., 2002. Numerical techniques for maximum likelihood estimation of continuous time diffusion processes .*Journal of Business and Economic Statistics* **20**, pp. 279-316.

Efron, B. and Tsibirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York

Eguchi, S., K. T., and Park, B. (2003) Local likelihood method: a bridge over parametric and non-parametric regression. *Journal of Nonparametric Statistics* **15** pp. 665-683.

Elerian, O., Chib S., Shephard N., (2001) Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69**, pp. 959-1012.

Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. *Journal ofBusiness and Economic Statistics* **19**, 177-191.

Fan, J., Heckman, N., and Wand, M.(1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of American Statistical association* **90**, pp.141-150.

Fan, J., and Gijbels, I. (1996). *Local polynomial modeling and its applications*, 1st ed. Chapman and Hall, London.

Fan, J., Mark, F., and Gijbels, I. (1998) Local maximum likelihood and inference. *Journal of the Royal Statistical Society* **Series B 60** pp. 591-608.

Fr$\ddot{O}$lich, M. (2006) Non-parametric regression for binary dependent variables. *Econometrics Journal* **9**,511-540.

Gallant, A.R., Tauchen G., 1996. Which Moments to Match? *Econometric Theory* **12**, pp. 657-681.

Gourieroux, C. and Monfort, A. (1981).Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models .*Journal of Econometrics*, **17**, pp. 83-97

Gourieroux, C., and Jasiak, J. (2010) Local likelihood density estimation and value-at-risk. *Journal of Probability and Statistics*

Gozalo, P., and Linton, O. (2000) Local nonlinear least squares: using parametric information in non-parametric regression. *Journal of Econometrics* **99**, pp. 63-106.

Han, A.(1987) Non -parametric analysis of a generalized regression model. *Journal of Econometrics* **35**, pp. 303-316.

Hansen, L.P., Scheinkman, J. (1995). Back to the future: generating moment implications for continuous time Markov processes*. Econometrica* **63**, pp. 767-804.

Hardle, W., and Stocker, T.(1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of American Statistical Association* **84,** pp. 986-995.

Hart, J. D. (1997) *Non parametric Smoothing and Lack-of-Fit Tests*. New York: Springer Verlag

Heckman, J., Smith, J., and Clements, N. (1997). Making the most out of program evaluations and social experiments: accounting for heterogeneity in program impacts. *Review of Economic Studies* **64**, pp. 487-535.

Horowitz, J., and Hardle, W. (1996). Direct semi parametric estimation of single-index models with discrete covariates. *Journal of American Statistical Association* **91**, pp. 1632-1694.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. Second Edition, John Wiley & Sons Inc., New York

Ichimura, H.(1993) Semi parametric least squares and weighted semi parametric estimation of single -index models*. Journal of Econometrics* **58**, pp. 71-120.

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

Jiang, G.J., Knight J., 2002. Estimation of continuous-time processes via the empirical characteristic function. *Journal of Business and Economic Statistics*, pp. 198-212

Klein, R., and Spady, R.(1993)An efficient semi parametric estimation for binary response models. *Econometrica* **61**, pp. 387-421.

Kohavi, R., and John, G. 1997. Wrappers for features subset selection. Artificial Intelligence **97**: pp. 273-324

Kohavi, R., and Provost, F. 1998. On Applied Research in Machine Learning. *Machine Learning* **30** pp.127-132

Kuo, C. L. and Yi-Ju, C.(2005) Testing the goodness of fit of logistic models based on local linear smoothing. *Journal of Information and Management Sciences* **16**

Landis, J., and Koch, G. (1977). The measurement of observer Agreement for categorical data. Biometrics, **33(1)** pp.159-174

Marc Aerts and GerdaClaeskens .(1997) Local Polynomial Estimation in Multi parameter likelihood models. *Journal of the American Statistical Association* **92**, pp 1536-1545

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics*, Edited by Paul Zarembka. New York: Academic Press

Powell, J., Stocker. J., and Stocker, T.(1989) Semi parametric estimation of index coefficients. *Econometrica* **57**, pp. 1403-1430.

Pulkstein, E. and Robinson, T. (2002).Two goodness of fit tests for logistic regression models with continuous covariates. *Statistics in Medicine* **21**, pp. 79-93

Racinne, J., and Li, Q. (2004) Non parametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**, pp. 99-130.

Royston, P. (1992). The use of cusums and other techniques in modeling continuous covariates in logistic regression. *Statistics in Medicine* **11**, pp. 1115-1129

Ruppert, D., and Wand, M.P. (1994) Multivariate locally weighted Least Squares Regression. *Annals of Statistics*, **22**, 1346-1370.

Singleton, K. (2001) Estimation of affine pricing models using the empirical characteristicfunction. *Journal of Econometrics* **102**, pp. 111-141

Staniswalis, J. (1989) The kernel estimates of a regression function in likelihood-based models.*Journal of American Statistics Association*.**84** pp. 276-283

Stocker, T. (1991) Equivalence of direct, indirect and slope estimators of average derivatives. *In non-parametric and semi parametric methods in economics and statistics*. Cambridge: Cambridge University Press.

Tibshirani, R.(1984).*Local Likelihood Estimation*. Stanford Linear Acceleration Centre.

Tibshirani, R., and Hastie, H.(1987) Local likelihood estimation. *Journal of American Statistical association* **82**, pp. 559-567.

Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press

Wand, M.P., and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman and Hall.

Zocchi, S. S. and Atkinson, A. C. (1999).Optimum Experimental Designs for Multinomial Logistic Models. *Biometrics*, **55**, pp. 4