EPH - International Journal of Applied Science
ISSN (Online): 2208-2182
Volume 01 Issue 03-August-2015

DOI: https://doi.org/10.53555/eijas.v1i3.15

ON USING NONPARAMETRIC REGRESSION METHODS TO ESTIMATE CATEGORICAL OUTCOMES MODELS WITH MIXED DATA TYPES

Ahmed M. Mami[1*], Ayman Ali Elberjo[2]
[*1] *Department of Statistics, Faculty of Science, Benghazi University, Benghazi, Libya*
[2] *Department of Statistics, Faculty of Science, Benghazi University, Benghazi, Libya P.O.Box 9480*

[*]***Corresponding Author:-***
E-mail:- ahmed.mami@uob.edu.ly

Abstract:-
*Many data analysis methods are sensitive to the type of data under study. When we begin any statistical data analysis, it is very important to recognize the different types of data. Data can take a variety of values or belong to various categories, whichever numerical or nominal. However, there are two types of data, quantitative and qualitative (Categorical) data. The general and powerful methodological approaches for the analysis of quantitative data have been widely taught for several decades. While the analysis for qualitative data analysis have blossomed only in the past 25 years. The need for analysis of categorical data techniques has increased steadily in recent years, in economic, health, social science. However, analysis of categorical data models when the dependent variable binary or multinomial outcomes with mixed explanatory variables are complex. The main goal of this paper is to estimate a nonparametric regression model of the binary and multinomial outcomes models with mixed explanatory variables, it is based on nonparametric conditional CDF method and (**PDF**) method of bandwidth selection, presented by Li and Racine (2008). Then we have compared it with one of the most common method of parametric regression (the logistic regression model). The comparisons will be based on two criteria depends on their classification ability through Correct Classification Ratio **CCR** as well as their log likelihood value **LLK**. We conducted several simulation studies using generated random data (categorical discrete and continues) in order to investigate the performance of both the parametric model and the nonparametric model for binary and multinomial outcomes. Interesting results have been achieved in this work. Application on real-data have also been applied when there exist mixed variables. We make use of dataset of the Household Expenditure Survey (**HES**).*

Keywords:-*Nonparametric Regression, logistic regression model, a conditional cumulative distribution function(**CDF**), conditional probability density functions (**PDF**) , bandwidth selection, Cross-validation criterion (**CV**), Correct Classification Ratio (**CCR**), log likelihood (**LLK**), Household Expenditure Survey (**HES**).*

INTRODUCTION

1. Logistic Regression Model

The logistic regression models can be classified upon the scale of the dependent variable. If the dependent variable has only two levels (categories), then the Binary Logistic Regression Model is used. If the dependent variable has more than two levels (categories), then the Multinomial Logistic Regression is used.

1.1 Binary Logistic Regression Model

The important difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable [3]. Recently, the binary logistic regression model has become a popular tool in most business applications [2].

Many categorical dependent variable *y* have only two categories, we denote the two possible outcomes, Success "1" and Failure "0". The distribution of *y* is specified by probabilities for one outcome.

The **S** shaped curves in Figure 1 are typical.

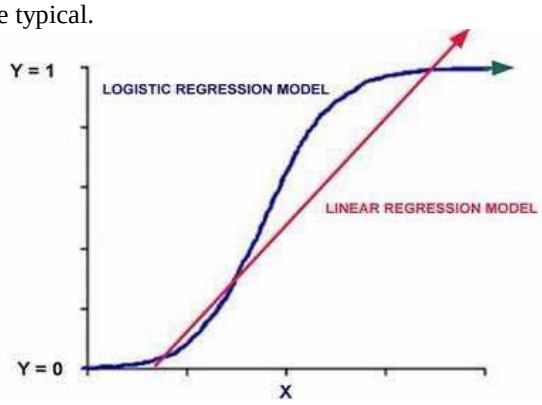

Figure 1: The graphical representation of a typical linear regression model and a typical binary logistic regression model.

The most important mathematical function with this shape has formula

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad (1)$$

This is called the logistic regression function, so the binary logistic regression model is

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \quad (2)$$

It is a special case of Generalized Linear Model (GLM), random component for (success, failure). The outcome has a binomial distribution, where $\frac{\pi(x)}{1-\pi(x)}$ is called odds ratio and the logit function

$\log\left[\frac{\pi(x)}{1-\pi(x)}\right]$ is called *logit* model. Therefore, $\frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + \beta_1 x)$

$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log[\exp(\beta_0 + \beta_1 x)] = \beta_0 + \beta_1 x \quad (3)$$

Where πis restricted to (0 –1) range, and the logit can be any real number with the possible range for linear predictors such as $\beta_0 + \beta_1 x$. If the parameters β> 0 then $\pi(x)$ increases as x increases. If β< 0, then $\pi(x)$ decreases as x increases. Thus, β determines the rate of increase or decrease of the curve, and if β=0, then the curve flattens to a horizontal straight line. The logistic regression model can be extend to other models with multiple explanatory variables, where the formula for π (x), becomes

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \quad (4)$$

Let *p* denotes number of explanatory variables for a binary dependent *y* by $X_1, X_2, \ldots, X_p$, the Model for log odds is

$$Logit\,[p(y=1)] = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \quad (5)$$

Where the parameters of logistic regression models are estimated by the Maximum Likelihood Estimation (MLE) method.

Where the likelihood equation

$$l(\beta) = \sum_{i=1}^{N} y_i \left(\sum_{p=0}^{P} x_{ip}\beta_p\right) - n_i \log\left(1 + \exp\left(\sum_{p=0}^{P} x_{ip}\beta_p\right)\right) \quad (6)$$

To find the critical points of the log likelihood function, we first set the derivative with respect to each β equal to zero. It yields a (p+1) terms of the non-linear equations that cannot be solved only through an iterative algorithm

Where $\hat{\pi} = \exp(\sum_{p=0}^{P} x_{ip}\hat{\beta}_p) \Big/ [1+\exp(\sum_{p=0}^{P} x_{ip}\hat{\beta}_p)]$ is maximum likelihood estimate of $\pi(x)$. Agresti (1990) had presented the maximum likelihood estimator $\hat{\beta}_{ML}$ of β and shows that $\hat{\beta}_{ML}$ is the value of β that maximizes equation (6)

Lastly, the Binary logistic regression model does not assume (as the linear regression model) normality, linearity, and homoscedasticity.

1.2 Multinomial Logistic Regression

The Multinomial logistic regression (**MLR**) model is a simple extension of the binary logistic regression that allows for more than two categories of the dependent variable.

If we have *n* independent observations with *p* explanatory variables, and the qualitative dependent variable has *J* categories. Then, to construct the logits in the multinomial case, one of the categories must be considered the base level and all the logits are built relative to it. Any category can be taken as the base level, so we will take category *J* as the base level. Then there is no ordering, it is apparent that any category may be considered *J*. Let π_j denote the multinomial probability of an observation falling in the i^{th} category. Thus, to find the relationship between this probability and the *p* explanatory variables $X_1, X_2,...,X_p$ the multiple logistic regression model then be

$$\log\left[\frac{\pi_j(X_i)}{\pi_k(X_i)}\right] = \beta_{0i} + \beta_{1j}X_{1i} + \beta_{2j}X_{2i} + ...+ \beta_{pj}X_{pi} \qquad (6)$$

Where $j = 1, 2, \ldots, (k-1), i = 1, 2, ..,n$. Since all the π's add to unity, this reduces to

$$\log(\pi_j(X_i)) = \frac{\exp(\beta_{0i} + \beta_{1j}X_{1i} + \beta_{2j}X_{2i} + ...+ \beta_{pj}X_{pi})}{1+\sum_{j=1}^{k-1}\exp(\beta_{0i} + \beta_{1j}X_{1i} + \beta_{2j}X_{2i} + ...+ \beta_{pj}X_{pi})} \qquad (7)$$

For $j = 1,2, \ldots,(k-1)$ and $i = 1, 2, ..,n$

The multinomial logistic regression (**MLR**) model uses the maximum likelihood estimation (**MLE**) method to evaluate the probability of categorical membership in the same manner as the binary logistic regression model.

It is important to note that, as in the binary logistic regression model, the **MLR** model does not assume (as in linear regression models) normality, linearity, and homoscedasticity.

2. Nonparametric Estimation of Conditional CDF with Both Categorical and Continuous data

The Nonparametric Regression methods are simply alternative statistical approaches used when some assumptions valid for parametric Regression methods are not provided. They are effective methods for data, which have low sample size or inconsistent sample.

It is more suitable to use nonparametric estimators when there is no parametric form for the regression function, because when the parametric model is valid, nonparametric models will be less efficient. Furthermore, nonparametric models can be used to test the validity of parametric models [4].

The estimation of a regression function (i.e., a conditional mean) is often the most common econometric application of nonparametric techniques. However, it is of interest usually to model conditional quantiles (e.g., the median), particularly when it is felt that the conditional mean is not representative of the impact of the explanatory variables on the dependent variable. Furthermore, the quantile regression function in general provides a much more comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function [6].

The conditional quantile function can be easily modeled by inverting the conditional Cumulative Distribution Function (**CDF**) at the desired quantile. Of course, the conditional **CDF** is unknown and must be estimated. The nonparametric estimation of conditional **CDF**s has received recently much attention.

The crucial issue in the nonparametric estimation is the selection of smoothing parameters (bandwidths). Unfortunately, there is no an automatic data-driven method for selecting the optimal smoothing parameters when estimating a conditional **CDF**.

However, Li and Racine (2008) had presented methodology for conditional **CDF** estimator and they adopted the conditional Probability Density Function (**PDF**) method of bandwidth selection proposed by Hall et al. (2004) in the context of estimating a conditional **CDF**.

2.1 The Conditional Cumulative Density Function Estimator

We use $F(y|z)$ to denote the conditional **CDF** of Y given $Z = z$,

$$\tilde{F}(y \mid z) = \frac{n^{-1}\sum_{i=1}^{n} I(Y_i \le y)K_\gamma(Z_i, z)}{\hat{\mu}(z)}, \qquad (8)$$

Let,, (z = z^c, z^d)= where $z^c \in R^q$ is a q-dimensional continuous random vector, and where z^d is an r-dimensional discrete random vector. $\hat{\mu}(z) = n^{-1}\sum_{i=1}^{n} K_\gamma(Z_i, z)$ is the kernel estimator of $\mu(z)$. Also, we define that $K_\gamma(Z_i z) = W_h(Z_i^c, z^c)\ L_\lambda(Z_i^d, z^d)$ as a product of two kernels defined as $W_h(Z_i^c, Z_j^c) = \prod_{s=1}^{q} h_s^{-1} w\left(\frac{Z_{is}^c - Z_{js}^c}{h_s}\right)$ and $L_\lambda(Z_i^d, Z_j^d) = \left[\prod_{s=1}^{r} \lambda_s^{|z_{is}^d - z_{js}^d|}\right]\left[\prod_{s=r+1}^{r} \lambda_s^{I(z_{is}^d = z_{js}^d)}\right]$ with λ_s and h_s played the role of bandwidths.

• Smoothing the Dependent Variable

When the dependent variable Y is a continuous random variable, one can use an alternative estimator that also smoothing the dependent variable y. Thus, we can estimate Thus, we can estimate $F(y \mid z)$ by

$$\hat{F}(y \mid z) = \frac{n^{-1}\sum_{i=1}^{n} G\left(\frac{y - Y_i}{h_0}\right) K_\gamma (Z_i, z)}{\hat{\mu}(z)}, \qquad (9)$$

Where $G(v) = \int_{-\infty}^{v} w(u)du$ is the distribution function derived from the density function w(·), and h0 is the bandwidth associated with Yi

2.2 Selection of Smoothing Parameters

Li and Racine (2008) have affirmed that there does not exist an automatic data-driven method for optimally selecting bandwidths when estimating a conditional **CDF** in the sense that a weighted mean integrated square error (MISE) is minimized. However, there do exist well developed automatic datadriven methods for selecting bandwidths when estimating the closely related conditional **PDF**. In particular, Hall et al. (2004) have considered the estimation of conditional probability density functions when the conditioning variables are a mix of categorical and continuous data types. Let $f(y|z)$ denote the conditional probability density function of Y given $Z = z$, and let $g(y,z)$ denote the joint density of (Y,Z). We estimate $f(y|z)$ by $f(y \mid z) = \hat{g}(y \mid z)/\hat{\mu}(z)$, where

$$\hat{g}(y \mid z) = n^{-1}\sum_{i=1}^{n} w_{h_0}(Y_i, y)K_\lambda(Z_i, z)$$

$$w_{h_0}(Y_i, y) = h_0^{-1} w\left(\frac{Y_i - y}{h_0}\right)$$

is a kernel estimator of g(y,z), and where

Hall et al. (2004) have shown that a Weighted Square (WSQ) difference between $\hat{f}(y \mid z)$ and $f(y \mid z)$, i.e.,

$$WSQ_f \overset{def}{=} \int \left[\hat{f}(y \mid z) - f(y \mid z)\right]^2 \mu(z)s(y, z)dzdy$$

The cross-validation objective function $CV_f(h,\lambda)$, which can be defined as

$$CV_f(h \mid \lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{G}_{-i}(Z_i)s(Y_i, Z_i)}{\hat{\mu}_{-i}(Z_i)^2} - \frac{2}{n}\sum_{i=1}^{n} \frac{\hat{g}_{-i}(Y_i, Z_i)s(Y_i, Z_i)}{\hat{\mu}_{-i}(Z_i)} \qquad (10)$$

Where $\hat{\mu}_{-i}(Z_i) = (n-1)^{-1}\sum_{j \ne i} K_\gamma(Z_i, Z_j)$ and $\hat{\mu}_{-i}(Z_i) = (n-1)^{-1}\sum_{j \ne i} w_{h_0}(Y_i, Y_j)K_\gamma(Z_i, Z_j)$ Leave-one-out estimators of $\mu(Z_i)$ and $g(Y_i, Z_i)$ respectively, and

$$\hat{G}_{-i}(Z_i) = \frac{1}{(n-1)^2}\sum_{j \ne i}^{n}\sum_{l \ne i}^{n} K_\gamma(Z_i, Z_j)K_\gamma(Z_i, Z_l)\int w_{h_0}(y, Y_j)w_{h_0}(y, Y_i)dy$$

For comprehensive details of this theoretical development are in [6].

The Simulation Study

To compare the performance of the two methods of regression modelling when the dependent variable is (binary or multinomial outcomes) with mixed explanatory variables namely, the parametric method and the nonparametric method.

3.1 Description of the Experiment.

In this study, the goal is to compare the performance of the two different regression models namely, the parametric and the nonparametric. First, a random data has been generated. However, such a study is necessarily to be restrictive, because there are many possibilities for the number of explanatory variables *P*, and sample size (*n*). In the first set of experiments, the binary outcomes are utilized with the existence of different kinds of explanatory variables. However, in the second set of experiments the multinomial outcomes are utilized with the existence of different kinds of explanatory variables. Four choices of the number of explanatory variables (*p*=2, *p*=3, *p*=4, and *p*=5) are considered.

We have also chosen three different sample sizes (*n*=50, *n*=100 and *n*=300).

24 simulation studies were conducted, 12 of them were conducted on the binary

Outcome context and other 12 were conducted on multinomial outcome context. Each simulation study involves L=500 repetitions.

3.1.1 The Numerical Summary for the Binary Outcome

Estimate model of binary outcome when P=2							
Estimation Method		Parametric Logit Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.639 0.062	0.625 0.048	0.615 0.028	0.718 0.120	0.674 0.087	0.631 0.043
LLK	Mean (S.D)	-31.563 2.105	-64.459 2.902	-196.406 4.613	-29.104 4.958	-61.273 6.551	-192.347 8.700
Estimate model of binary outcome when P=3							
Estimation Method		Parametric Logit Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.784 0.059	0.772 0.042	0.767 0.025	0.842 0.081	0.802 0.057	0.774 0.027
LLK	Mean (S.D)	-23.623 3.989	-50.055 5.217	-154.909 9.02917	-19.161 6.349	-44.571 8.056	-148.165 12.398
Estimate model of binary outcome when P=4							
Estimation Method		Parametric Logit Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.794 0.060	0.788 0.042	0.779 0.024	0.868 0.084	0.826 0.063	0.789 0.028
LLK	Mean (S.D)	-22.630 4.371	-47.603 5.846	-150.102 9.382	-17.08 6.420	-40.577 9.602	-140.776 14.021
Estimate model of binary outcome when P=5							
Estimation Method		Parametric Logit Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.852 0.054	0.840 0.036	0.838 0.020	0.918 0.066	0.882 0.053	0.846 0.025
LLK	Mean (S.D)	-17.06 4.631	-38.230 6.223	-122.21 10.128	-11.081 5.705	-29.383 9.427	-110.586 15.802

Table (1): The numerical summary of results obtained using two different methods (parametric, nonparametric) from the 1st set of Simulation studies for binary outcome. That includes (Means, Standard Deviations) for both criteria CCR and LLK respectively of 500 simulation runs.

From Table (1), we have noticed the following remarks. First, the mean values for **CCR** in nonparametric method always higher than their corresponding counterparts mean values for **CCR** in Parametric method for all possible choices of sample size. Secondly, the values of standard deviations for **CCR** in nonparametric method always higher than their corresponding counterparts standard deviation values for **CCR** in parametric method, the main reason behind that the **CCR1** values have much less variation than the **CCR2** values, which in some cases the **CCR2** values can reach 100% classification ratio. Thirdly, the mean values in both **CCR1** and **CCR2** are increasing as the number of explanatory variables increases particularly at small sample size where we noticed that **CCR2** reach to **91%** at **n**=50. Fourthly, the mean values in both **CCR1** and **CCR2** are decreasing as the sample size *n* increases. Fifthly, the results obtained for the mean values of **CCR** are exactly applied the mean values of **LLK**. Lastly, the numerical results obtained using the Nonparametric **CDF** Model are superior comparing with their corresponding counterparts using the parametric Logit Model.

3.1.2 The Numerical Summary for Multinomial Outcomes

Estimate model of multinomial outcome when P=2							
Estimation Method		Parametric **MLR** Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.478 0.055	0.444 0.043	0.418 0.026	0.557 0.143	0.510 0.117	0.447 0.066
LLK	Mean (S.D)	-50.512 2.387	-104.682 2.617	-320.730 3.827	-49.007 5.854	-101.992 6.844	-316.942 9.430
Estimate model of multinomial outcome when P=3							
Estimation Method		Parametric **MLR** Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.606 0.066	0.591 0.0457	0.576 0.028	0.686 0.127	0.627 0.079	0.585 0.032
LLK	Mean (S.D)	-41.535 4.644	-86.924 5.714	-269.412 9.453	-37.541 7.708	-81.921 10.378	-263.980 13.293
Estimate model of multinomial outcome when P=4							
Estimation Method		Parametric **MLR** Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.616 0.064	0.587 0.045	0.567 0.029	0.713 0.139	0.641 0.096	0.583 0.036
LLK	Mean (S.D)	-39.548 4.592	-85.045 5.593	-266.185 9.803	-34.464 8.764	-78.464 11.876	-258.442 14.491
Estimate model of multinomial outcome when P=5							
Estimation Method		Parametric **MLR** Model			Nonparametric **CDF** Model		
Sample size		n=50	n =100	n =300	n =50	n=100	n =300
CCR	Mean (S.D)	0.653 0.065	0.605 0.049	0.57925 0.02659	0.77124 0.142029	0.67778 0.10764	0.608606 0.046804
LLK	Mean (S.D)	-34.591 5.010044	-78.661 6.361	-250.71 9.334	-29.053 9.202	-71.053 13.330	-239.475 17.109

Table (2): The numerical summary of results obtained using two different methods (parametric, nonparametric) from the 2nd set of Simulation studies for multinomial outcome. That includes (Means, Standard Deviations) for both criteria CCR and LLK respectively of 500 simulation runs.

From Table (2), we have noticed the following points. Firstly, in this simulation study the mean values for **CCR** in nonparametric method always higher than their corresponding counterpart mean values for **CCR** in parametric method for all possible choices of sample size. However, we have noted that the **CCR** mean values were smaller in both estimated models with their corresponding counterparts in case of the binary outcome. Secondly, the values of standard deviations for **CCR** in nonparametric method always higher than their corresponding counterparts standard deviation values for **CCR** in parametric method. The main reason behind that the **CCR1** mean values have much less variation than the **CCR2** mean values, which in some cases the **CCR2** values can reach up to 100% classification ratio. Thirdly, the mean values in both **CCR1** and **CCR2** are increasing as the number of explanatory variables increases particularly at small sample size where we noticed that **CCR2** reach to **77%** at **n**=50. Fourthly, the mean values in both **CCR1** and **CCR2** are decreasing as the sample size *n* increases. Fifthly, the remarks noted about the mean values of **CCR** are exactly applied the mean values of **LLK**. Lastly, the numerical results obtained using the Nonparametric **CDF** Model are superior comparing with their corresponding counterparts using the parametric Logit Model.

4. Applications on Real Data Sets

The **HES** data in years 2009/2010 used to compile information on the level and patterns of consumption expenditure of private households, where the survey covered the following topics:

- Geographic Coverage: The survey was covered both urban and rural areas (Strata) in Malaysia. Those Strata consists of 16 states all over Malaysia.
- The survey covered the Type of Living Quarters which we classified into three different categories
- The Status of Living Quarters also are included which also classified into three different categories.
- The Number of household members is also considered.
- The Household expenditure, which contained the consumption and the non-consumption expenditure Coverage.

The Household Expenditure Survey 2009/10, was carried out for a period of 12 months, started from April 2009 to March 2010.

However, we have considered six variables.
Then, we begin by estimating the model and compute the classification matrices by applying the two different methods (parametric, nonparametric) for the four different random samples size ($\mathbf{n_{1=264}}$, $\mathbf{n_{2=306}}$, $\mathbf{n_{3=964}}$, and $\mathbf{n_{4=1622}}$) that have been drawn from survey data. It is worth noting that in the parametric method, we use the Logit model and then we compute the classification matrix, which in turn use to compute the Correct Classification Ratio (**CCR**). However, in the nonparametric method we use the conditional density estimator and then compute the conditional mode, which in turn use to compute the **CCR**. Finally, we will be able to compare their Correct Classification Ratio (**CCR**) via their classification matrices. The second criterion of comparisons is based on calculating the Log Likelihood value (**LLK**) of both the Logit model and the conditional density function.

4.1 The case of Binary outcome

The variable "Strata" as the dependent variable in the binary outcome case have two categories (Urban, Rural). Then, we fit the variable "Strata" with five explanatory variables, namely: States, Total members, Type of Living Quarters, Status of Living Quarters, and Total Expenditure 01-12.
By applying the two different methods (parametric, nonparametric), we obtained the results that have been tabulated in Table (3). Table (3) displays the numerical results of the two criteria CCR and LLK obtained by utilizing two different methods (parametric, nonparametric) of binary outcome for four different random samples drawn from the HES data.

Estimation Method	Parametric **Logit** Model				Nonparametric **CDF** Model			
Sample size	$\mathbf{n_1=264}$	$\mathbf{n_2=306}$	$\mathbf{n_3=964}$	$\mathbf{n_4=1622}$	$\mathbf{n_1=264}$	$\mathbf{n_2=306}$	$\mathbf{n_3=964}$	$\mathbf{n_4=1622}$
CCR	0.788	0.784	0.774	0.760	0.841	0.830	0.791	0.799
LLK	-117.3	' -167.5	-462.2	-790.5	-99.4	-133.5	-429.0	-681.3

Table (3). The numerical results of the two criteria CCR and LLK obtained by utilizing two different methods (parametric, nonparametric) of binary outcome by utilizing different random samples drawn from the HES data.

Having a close look at Table (3), where the comparisons based on their classification ability as well as their log likelihood value, we have noted that the proposed nonparametric method performed well comparing with the parametric method in the light of the results obtained in this work.
The largest value for **CCR** and **LLK** obtained at the $\mathbf{n_1}$ and $\mathbf{n_2}$ sample size. As the sample size increases, we noticed that both methods gets closer to each other. This agrees with the simulation results obtained in the binary outcome case.

4.2 The case of multinomial outcome

In this case we considered that the dependent variable multinomial outcome which is the "Status of Living Quarters" have three categories with five explanatory variables (strata, States, Total members, Type of Living Quarters,, Total Expenditure 01-12). However, by applying the two different methods, we obtained the results that have been tabulated in table 4 that involves two different criteria (**CCR**, **LLK**).

Estimation Method	Parametric **Logit** Model				Nonparametric **CDF** Model			
Sample size	$\mathbf{n_1=264}$	$\mathbf{n_2=306}$	$\mathbf{n_3=964}$	$\mathbf{n_4=1622}$	$\mathbf{n_1=264}$	$\mathbf{n_2=306}$	$\mathbf{n_3=964}$	$\mathbf{n_4=1622}$
CCR	0.723	0.703	0.693	0.701	0.784	0.843	0.759	0.759
LLK	-168.9	-206.9	-672.6	-1138.0	-150.4	-158.2	-556.2	-922.1

Table (4). The numerical results of the two criteria CCR and LLK obtained by utilizing two different methods (parametric, nonparametric) of multinomial outcomes by utilizing different random samples Drawn from the HES data.

From Table (4) we can clearly see that the nonparametric method performed well than the parametric method in the light of the results obtained in this application. We can see that it gives the largest value for (**CCR, LLK**) of $\mathbf{n_1}$, $\mathbf{n_2}$ sample size and both methods are closed to each other as the sample size increases, and this confirms the results obtained from the simulation.

5. Summary and Conclusion

From the simulation study, the evaluation of our two criteria support the superiority of the nonparametric method over the parametric method for all possible sample sizes *n* as well as for all possible number of explanatory variables *p*. However, in case of small samples the nonparametric method is better able to predict the **CCR** values than its corresponding counterpart of the parametric method. Moreover, the nonparametric method performance very well as the

number of explanatory variables increases, where it reaches its highest level in cases of p=5. An interesting feature to note is that, values of **CCR** in both methods getting closer to each other as the sample sizes increase.

In the multinomial outcomes, we have noted that the values of **CCR** was lower than their corresponding counterparts the binary outcome case. However, the nonparametric method still provides much better performance in case of small samples size with p=4, p=5, where it reaches its highest level in cases p=5 with n=50. Again, we have noted that values of **CCR** in both methods getting closer to each other as the sample sizes increase.

In our application, we have considered four different random samples drawn from dataset, namely the Household Expenditure Survey 2009/10 **HES** for 6495 Household. In this application, we have noticed that the nonparametric method is better able to predict the **CCR** value than its corresponding counterpart of the parametric method. All results in this application have supported the following

Conclusion:

When sample size is **n1=264**, the nonparametric model performance very well in prediction comparing with the parametric model. Also, as the sample size increases to **n2=306** we have noticed slightly drop. However, as the sample size increases to **n3=964, n4=1622**, both models tend to be close to each other in performance.

Finally, we are hoping from this work to present an alternative approach to analyze the categorical data that may be useful in solving similar problems in many fields that uses the categorical data. Lastly, we may see the results obtained in this study are interesting because of the lack of research on this area.

References

[1].Agresti. A. (1990). Categorical Data Analysis. Wiley, New York.

[2].Agresti, A. (2002). Categorical Data Analysis 2nd edition. John Wiley & Sons, Inc. Hoboken, New Jersey.

[3].David W. Hosmer (2000) Applied logistic regression, 2nd ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

[4].Eubank, Randall L. (1988) Spline Smoothing and Nonparametric Regression. New York: Marcel Dekker 1st ed.

[5].Hall, P., Racine, J. & Li, Q. (2004) Cross-validation and the estimation of conditional probability densities, Journal of the American Statistical Association **99**(2), 1015–1026.

[6].Li Q, Racine J (2008) Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data. Journal of Business and Economic Statistics, 26(4), 423–434.