ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# A COMPARISON OF THE PERFORMANCE OF DISCRIMINANT ANALYSIS AND THE LOGISTIC REGRESSION METHODS IN CLASSIFICATION OF DRUG OFFENDERS IN KWARA STATE

**[1]Balogun, O.S., [2]Balogun, M.A., [1]Abdulkadir, S.S. and [1]Jibasen, D.**

**1.** Department of Statistics and Operations Research, Modibbo Adama University of Technology, P.M.B. 2076, Yola, Adamawa State, Nigeria.

**2.** Department of Agricultural Extension and Rural Development, University of Ilorin, P.M.B. 1515, Ilorin, Kwara State, Nigeria.

## *Manuscript Info*

***Manuscript History:***

Received: 26 August 2014
Final Accepted: 20 September 2014
Published Online: October 2014

***Key words:***

Logistic Regression, Classification, Drug Offenders, Discriminant Analysis.

***Corresponding Author***

**Balogun, O.S.**

## *Abstract*

The study compares two statistical methods:- Discriminant analysis and the Logistic regression model in predicting Drug Offenders, Drug peddlers and Non-drug peddlers. Of the 262 cases examined for Drug Offenders, Discriminant Analysis classified the Drug Peddlers correctly (56.3%) while it recorded (84.6%) success rate in classifying the Non-drug Peddlers. In the case of the Logistic regression, it recorded (92.4%) and (97.9%) success rate in classifying the Drug Peddlers and Non-drug Peddlers respectively. The overall predictive performance of the two models was high with the Logistic regression having the highest value (95.4%) and (71.8%) for Discriminant Analysis. Among the four characteristics examined, exhibit type and age were not significant variables for identifying Drug Offenders by both methods while exhibit weight is important identifying variable for both except gender which was significant in the Logistic model. The study shows that both techniques estimated almost the same statistical significant coefficient and that the overall classification rate for both was good while either can be helpful in selection of Drug Offenders. However, given the failure rate to meet the underlying assumptions of Discriminant Analysis, Logistic Regression is preferable.

.



## 1. Introduction

The National Drug Law Enforcement Agency charged with the responsibility of dealing with drug and drug related offences, in some occasion, may not be able to distinguish between Drug peddlers and Non-drug peddlers on the basis of oral evidence on possession and dealing with illicit drugs. Therefore, there is need for a scientific method to employ in order to classify future offenders into Peddler or Non-Peddler if some variables such as age of offenders, length of dealing in illicit drugs, type of exhibit, weight of exhibit and so on are known.

The involvement in illicit drug has so many social implications some of which include prostitution, theft, sexual assaults on female folks. According to Odejide (1992), those involved in peddling are ignorant of the problem emanating from it. It is true that a small number of people, mainly those organizing the illicit drug trade, make large profits from illicit crop cultivation, but the vast majority of people, including most of those benefiting from such trade, are adversely affected by the illicit activity. In the long term, the illicit industry causes major problem that eventually affect the economic development of the country concern. On this premise the authors believe offenders, especially peddlers/traffickers deserve stiffer penalty than the users, because without sellers, buyers will not exist

and invariably reduces or eliminate the activity in the system. Therefore, the laws concerning the illicit drug need to be reformed to reflect the new idea of treating Peddlers with stiffer penalty than Non-Peddlers. This will be the responsibility of the legislators to appropriate the enabling laws.

In USA the amount of quantity of illicit drug determines the length of sentence pass on the offenders. However, the large quantity or numbers of item seized usually make the calculation of total quantity cumbersome, hence the use of statistical sampling such as multistage, composite and simple random samplings have been adopted. A rule of thumb developed by Izenman (2001) for determination of sample size is square root of N, $\sqrt{N}$ , where N is the number of items in a container is popular. A 95% confidence interval is developed to reduce the error for using the rule.

To investigate differences between or among groups, and classify cases into groups can be done using statistical methods. This method can complement oral method of classifying the drug offenders. With this technique, the drug data to which a particular data belongs can be identified using the Drug offenders' characteristics. To predict such group membership; the dependent variable is a nominal variable with two levels or categories with say 0 = drug Peddlers and 1 = Non-drug Peddlers. If a low percentage of Drug Offenders based on the Drug Offenders characteristics have been properly classified, then the original selected drug data forms have been poorly selected, but if the success rate is high, then the drug data form would have been properly selected. According to Lin Wang, *et al* (1999), if the dependent variable is nominal variable, the researcher has two choices either to use discriminant analysis or a logistic regression analysis.

Logistic regression and linear discriminant analyses are multivariate statistical methods and are two of the most popular methodologies for solving classification problems involving dichotomous class variable, Yarnold, *et al* (1994). The logistic regression predicts the probability of group membership in relation to several variables independent of their distribution. The logistic regression is based on calculating the odds of having the outcome divided by the probability of not having it. Logistic regression is non-parametric and assumed a distribution free sample. The Discriminant analysis on the other hand is used to determine which set of variables discriminates between two or more naturally occurring groups and to classify an observation into these known groups. It is a parametric method and assumes that the sample comes from a normally distributed population and that the covariance matrices of the independent variables are the same for all groups.

Several authors have formally compared the two techniques. For example, Halperin, *et al* (1971) compared the two methods and noted only small differences in the classification ability between the analytical procedures. Dattalo (1995) found that both methods performed well as classification technique but concluded that the logistic was more parsimonious and easier to interpret. Hyunjoon, *et al* (2010) also found that the two models are equally effective in predicting restaurant bankruptcy, but concluded that the logit model is preferred for restaurant bankruptcy prediction because of its theoretical soundness. George Antonogeorgos, *et al* (2009) in evaluating factors associated with asthma prevalence among 10-12 years old children concluded that the two methods resulted in similar result while Montgomery, *et al* (1987) in prediction of coliform mastitis in diary cows, concluded that both techniques selected the set of variable as important predictors and were of nearly equal value in classification performance. Press et al (1978) concluded that each analytical technique served a unique function. Discriminant analysis was useful for classification of observations into one of two populations whereas logistic regression was useful for relating a qualitative (binary) dependent variable to one or more independent variables by a logistic distribution. Kleinbaum, *et al* (1998) cited in Montgomery, *et al* (1987) compared the classification ability of both methods using data set which met the assumption of discriminant analysis and noted that logistic regression model was slightly superior. Edokpayi, *et al* (2013) compared the two methods in classifying and assessing the relative importance of the fruit form characteristics, but concluded that the two methods were of nearly equal value but logistic regression would be preferable whenever the normality assumption are violated.

Based on the above arguments, the aim of this work is to compare the two analytical methods using data set on drug offenders. This work determined if there is convergence between the two methods of analysis in classifying the subject (drug offenders) into one of the two populations (Drug Peddlers and Non-drug Peddlers) and also determined the tenability of the assumption underlying the two methods.

In choosing between the two methods, the study applied the following criterion, the prediction of group membership and the assessment of its success i.e. determine which between the two methods provides a higher accuracy in classifying the drug offenders. Determine which variables appears significant in classifying the dependent variable by inspection of the coefficients and testing the assumption of normality and equal covariance required for the validity of the discriminant analysis.

The outcome will not only complement the breeders' current practices but will also assist the research scientists to make appropriate choice in their application of these two techniques.

## 2. Materials and Methods

The data consists of four Drug Offenders characteristics (independent variables), and drug offender (dependent variable). The Drug Offenders and the Drug Offenders characteristics are listed in Tables 1and 2 respectively.

**Table 1**: **Drug Offenders (Dependent variable)**

| Group code | Drug Offenders |
|---|---|
| 1 | Drug Peddlers |
| 2 | Non-Drug Peddlers |

**Table 2**: **Drug Offenders characteristics (Independent variables)**

| Variable code | Description |
|---|---|
| $X_1$ | Exhibit type |
| $X_2$ | Age |
| $X_3$ | Exhibit Weight |
| $X_4$ | Gender |

**Discriminant Analysis**

Given a set of p independent variables $X_1, X_2, ..., X_p$, (Drug Offenders characteristics in this case), the technique attempt to derive a linear combination of these variables (Drug Offenders characteristics) which best separate or discriminates the two groups (Drug Offenders in this case). The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases with measurements for the predictor variables, but unknown group membership.

In general form, the Discriminant function is expressed as:

$$Z = a + W_1X_1 + W_2X_2 + ... + W_kX_k \quad .....(1)$$

Where: Z = discriminant score; a = discriminant constant; $W_k$ = discriminant weight or coefficients; $X_k$ = an independent variable or predictive variables.

The procedure automatically chooses a first function that will separate the groups as much as possible, it then chooses the second function that is both uncorrelated with the first function and provides as much further separation as possible. The procedure continues adding functions in this way until reaching the maximum number of functions as determined by the number of predictors and groups in the dependent variable. In two group discriminant function, there is only one discriminant function. The discriminant score obtained from the discriminant function is used to classify the Drug Offenders into one of the two drug data.

The importance of the derived discriminant function for the study was assessed using the canonical discriminant function coefficients, Wilks' Lambda, and an associated chi square and the percentage of the drug offenders correctly classified into group, Mbanasor, *et al* (2008). In testing the classification performances of the discriminant function, we use the overall hit ratio which is the same thing as percentage of the original group cases correctly classified. The relative classifying importance of the dependent variables (Drug Offenders) was assessed using the standardized discriminant coefficients. The greater the magnitude of the coefficients, the greater the impact of the variable as an identifying variable. However, to test the significance of the discriminant function as a whole we used the Wilks' Lambda. A significant lambda means one can reject the null hypothesis that the groups have the same discriminant function scores. The ANOVA table for the discriminant function score is another overall test of the discriminant analysis model. It is an F test, where a 'sig.' p-value < .05 means the model differentiates between the groups significantly better than chance.

**Classification rule**

We define the cut off as:

$$C = \frac{Z_1 + Z_2}{2} \quad .....(2)$$

Where, C = Cut off, Z =Group Centroids.

We first of all compute $Z_1(1.500)$ *and* $Z_2(1.500)$ which denote the functions at group centroids. Thus, the discriminating procedure is as follows. Assign a drug offender to group 1 if the discriminant score is > than the cut off (1.500) and group 2 if the discriminant score > the cut off (1.500), Efimafa, *et al* (2009).

**Logistic Regression**

Let Y denote the drug data which is categorical and can take one of the two possible values, denoted 1 and 2 $(Y = Drug\ \ Peddlers, Y = Non-drug\ \ Peddlers)$. Let $X = (x_1, x_2, ..., x_6)$, be the explanatory variables (Drug Offenders characteristics). This method uses the predicted probabilities to assign cases into the categories of the dependent variable and then compares the results with their actual categories. It can also be used to explain the effects of the explanatory on the dependent variables (Drug Offenders).

The logistic regression model can be defined mathematically as:

$$P = \frac{1}{1-\ell}\beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

Where P is the probability of the event occurring (i.e. the probability of selecting a particular Drug Offenders). $X_1 + X_2 + ... + X_6$ are the independent or predictor variables, and $\beta_1, \beta_2, ..., \beta_6$ are the coefficients representing the effects of the predictor variables and $\beta_0$ is the intercept (the value of the equation when all the X's are zero)

**Evaluation of the Logistic Regression Model**

In assessing the logistic regression model involves an overall evaluation of the model, the statistical significance of the individual regression coefficients, the goodness of fit statistics and the validation of predicted probabilities. A logistic model is said to provide a better fit if it demonstrates an improvement over the intercept –only model. An improvement over this baseline is examined by using three inferential statistical tests: the likelihood ratio, score and Wald tests. The statistical significance of individual regression coefficients $(i.e.\beta)$ is tested using the Wald chi-square statistic. The Hosmer-Lemeshow (H - L) is the inferential goodness of fit test used to assess the fit of a logistic model against actual outcome. The H – L statistic is a Pearson Chi-square statistic. If ($p > 0.05$) it is insignificant it suggests that the model fitted the data well. But if ($p < 0.05$) it is significant suggesting that the model did not fit the data.

**A Test of Assumption of Multivariate Normality and equal Covariance Matrices of the Discriminant Analysis**

Since in most studies, comparison of the logistic regression and discriminant analysis gives almost similar results, in order to decide which method to use, we consider the assumptions for the application of each one. In the case of discriminant analysis a normal distribution of the data and equal covariance matrices and that the violation of this assumption will render unreliable or invalid interpretation and inference of the result of the analysis.

**Normality Assumption**

The simplest method of assessing normality is by producing a histogram. The normal plot, P – P or Q – Q plot can also be used to assess the normality of a distribution. It is also possible to use Kolmogorov-Smirnov test if a sample size is greater than 50 or Shapiro-Wilk test if sample size is smaller than 50. In the present analysis, since the sample size is greater than 50 the Kolmogorov-Smirnov test used. The convention is that a significant value greater than 0.05 indicates normality of the distribution, Normadiah, *et al* (2011).

**Assumption of equal Covariance Matrices**

The hypothesis of interest is:

$$H_0 : V_1 = V_2 \ \ vs \ \ H_1 : V_1 \neq V_2$$

The assumption is that covariance matrices of the independent (classification) variables is the same for the two groups. Box's M test is used to test the equality of covariance matrices. If ($p > 0.05$), we do not reject the hypothesis that the two covariance matrices is equal but if ($p < 0.05$) the hypothesis that the two covariance is equal is rejected.

## 3. Result and Discussion

The results of the discriminant analysis and logistic regression model are presented below.

**Table 3: Classification of Drug Offenders by Logistic Regression and Discriminant Function Methods**

| | | Predicted Group Membership | | | |
|---|---|---|---|---|---|
| **Actual Group** | **No. of cases** | **Discriminant Analysis** | | **Logistic Regression** | |
| | | 1 | 2 | 1 | 2 |
| 1 | 119 | 67(56.3%) | 52(43.7%) | 110(92.4%) | 9(7.6%) |
| 2 | 143 | 22(15.4%) | 121(84.6%) | 3(2.1%) | 140(97.9%) |
| Overall % correctly classified | | 71.8% | | 95.4% | |

Table 3 shows the classification performances of the two methods. Of the 119 cases of Drug Peddlers, discriminant analysis predicted correctly 67(56.3%) and misclassified 52(43.7%), while the logistic regression classified correctly 110(92.4%) and misclassified 9(7.6%). In the case of the prediction of the group membership of Non-drug Peddlers which contains of 143 cases, the discriminant analysis classified correctly 121(84.6%) of the cases and misclassified 22(15.4%) while the logistic regression classified 140(97.9%) cases correctly and misclassified 3(2.1%) of the cases. The overall percentage correct classification of the Drug offenders was 71.8% and 95.4% for the discriminant analysis and the logistic regression method respectively. The results have therefore shown that the overall classification rate for both methods was good and either can be helpful in predicting the possibility of detecting or selecting drug data.

**Table 4: Hosmer-Lemeshow**

| **Step** | **Chi-square** | **d.f.** | **Sig** |
|---|---|---|---|
| 1 | 14.015 | 8 | 0.081 |

Table 4, since ($p > 0.05$) it is insignificant which suggest that model fitted the data well.

**Table 5: Variables and Coefficients for the Discriminant Analysis and the Logistic Regression models**

| | **Discriminant Analysis** | | | **Logistic Regression** | | |
|---|---|---|---|---|---|---|
| **Independent Variable** | **Wilks' Lambda** | **Canonical Coefficient** | **P-value** | **Wald Statistic** | **Coefficient** | **P-value** |
| Constant | - | 1.653 | - | 0.299 | 1.108 | 0.585 |
| Exhibit type | 0.993 | -1.373 | 0.172 | 0.187 | -0.726 | 0.666 |
| Age | 1.000 | -0.022 | 0.785 | 2.345 | 0.063 | 0.126 |
| Exhibit Weight | 0.930 | 0.104 | 0.000 | 43.462 | -0.010 | 0.000 |
| Gender | 1.000 | 0.151 | 0.937 | 8.880 | 2.651 | 0.003 |

Table 5, the Wilks' lambda was used to test which independent variables contributes significantly to the discriminant function. The F test of the Wilks' lambda shows that, three of the independent variables-the Exhibit type, Age and Gender were not significant ($p > 0.05$), while the remaining variable-Exhibit Weight is highly significant at ($p < 0.05$). For logistic regression the coefficient for the classification equation and is used to assess the relative classifying importance of the dependent variable (Drug Offenders). The Wald statistic is used to test the null hypothesis that the coefficients of independent variables in the model are zero. From the table, two of the Drug Offenders characteristics Exhibit weight and Gender were significant with an associated $p < 0.05$. However, the two other variables Exhibit type and Age were not significant.

However in comparison, both methods identified almost the same variable. Exhibit Weight is significant for both methods, while Exhibit type and Age were equally not significant for the two methods. Both methods however differ in the estimation of Gender. The direction of relationship was the same, but there were some extreme differences in the magnitude of the coefficients. According to Andrew, *et al* (1986), for purposes of parameter estimation, logistic

regression is more robust than discriminant analysis. But as observed by Press, *et al* (1978), if the populations are normal with identical covariance matrices, discriminant analysis estimators are preferred to logistic regression estimators.

**Table 6: Test of Normality and equal covariance matrices**

| Group | | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | | **Statistic** | **d.f.** | **Sig** | **Statistic** | **d.f.** | **Sig** |
| Exhibit type | 1 | 0.541 | 119 | 0.000 | 0.201 | 119 | 0.000 |
| | 2 | 0.535 | 143 | 0.000 | 0.309 | 143 | 0.000 |
| Age | 1 | 0.138 | 119 | 0.000 | 0.896 | 119 | 0.000 |
| | 2 | 0.153 | 143 | 0.000 | 0.893 | 143 | 0.000 |
| Exhibit Weight | 1 | 0.541 | 119 | 0.000 | 0.201 | 119 | 0.000 |
| | 2 | 0.535 | 143 | 0.000 | 0.309 | 143 | 0.000 |
| Gender | 1 | 0.530 | 119 | 0.000 | 0.344 | 119 | 0.000 |
| | 2 | 0.531 | 143 | 0.000 | 0.338 | 143 | 0.000 |

The result of the test of normality is presented in Table 6. When the assumption for normality and equal covariance matrices were tested using the Kolmogorov-Smirnov test and Box's M test respectively. The significant value of all the classification variables were less than 0.05, indicating that the variables were not normally distributed. The Box's M test value was (1108.671, $p < 0.000$), indicating a valuation of the assumption of the discriminant Analysis.

## 4. Conclusion and Recommendations

Using Drug offender data, the study has compared empirically the logistic regression and linear discriminant analysis, in both the classification performances of the two methods and in assessing the relative importance of the drug data characteristics in classification performance, both methods were of nearly equal value (71.8% and 95.4%), and almost selected the same set of variables (Exhibit Weight) is very significant to identifying drug data. The finding agrees with Montgomery, *et a*l (1987) and George Antonogeorgos, *et al* (2009) that the two methods result in similar results. A test of assumptions of multivariate normality and equal covariance matrices of the discriminant analysis were not satisfied. We thus agree with the conclusion of Press, *et al* (1978) that the use of logistic regression would be preferable whenever practical in situations where the normality assumptions are violated.

## References

Andrew, W. Lo (1986), Logit versus Discriminant analysis: A Specification test and application to corporate Bankruptcies. Journal of Econometrics, Vol. 31, Issue 2, pp 151-178.

Dattalo, P. (1995), A Comparison of Discriminant Analysis and Logistic regression: Journal of Social Services Research, Volume 19, Issue 3-4, pages 121-144.

Erimafa J.T. , Iduseri, A. and Edokpa, I.W. (2009), Application of Discriminant Analysis to predict the class of Degree for graduating students in a university system: International Journal of Physical Sciences, Vol. 4(1),Pp 016 – 021.

Edokpayi, A.A., Agho, C., Ezomo, J.E., Edosomwan, O.S. and Ogiugo, O.G. (2013). A Comparison of the Classification Performance of Discriminant Analysis and the Logistic Regression Methods in Identification of Oil Palm fruit Forms. A Paper Presented at the Annual Conference of Nigerian Statistical Association. 11-13th, September, 2013,pp 20-26.

George Antonogeorgos, Demosthenes .B. Panagiotakos, Kostas .N. Priftis and Anastasia Tzonou (2009), Logistic Regression and Discriminant Analysis in evaluating factors associatedWith Asthma Prevalence among 10-12 year old children: Divergence and Similarity of the two Statistical Methods: International Journal Pediatrics. Volume 2009, pp 1-7.

Halperin, M. Blackwelder, Weverter, J.I. (1971): Estimation of the Multivariate Logistic risk function: A

Comparison of the discriminant function and Maximum Likelihood Approaches. J. Cnron. Dis. 24.125-158.

Hyunjoon Kim and Zheng Gu (2010), Predicting Restaurant bankruptcy: A Logit model in Comparison with Discriminant Model; Tourism and Hospitality Research Journal, Vol. 10, Pp 171-187.

Izenman, .A .J. (2001), Legal and Statistical aspect of the forensic study of illicit drugs, Statistical Science, 16.

Kleinbaum, D.G; Kupper, L.L;Muller, K.E,and Nizam,A (1998) Applied Analysis and

Multivariate Methods. Third Edition. Duxbury Press.

Lin Wang, Xitao Fan (1999), Comparing Linear Discriminant Function with Logistic Regression for two groups Classification problem: Journal of Experimental Education. Vol. 67

Mbanasor, J.A. and Nto, P.O.O., (2008), Discriminant Analysis of Livestock farmers' credit worthiness: Journal Of Nigeria Agriculture. Vol.1,pp 1-7.

Montgomery, N.E., White, M.E. and Martin, S.W. (1987), A Comparison of Discriminant analysis and Logistic Regression for the prediction of Coliform Mastitis in dairy Cows: Canadian Journal of Veterinary Research, 51(4) Pp 495-498.

Normadiah, M.R. and Yap, B.W. (2011), Power Comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors And Anderson-Darling test. Journal of Statistical Modeling and Analytics, Vol. 2, No. 1, 21-33.

Odejide, A.O. (1992), Drugs in the Third World. In Drugs and Society to Year 2000, Ed by Vamos andCorrivean,Pp 116 – 119.

Press, J. and Wilson, S. (1978), Choosing Between Logistic Regression and Discriminant Analysis: Journal of the American Statistical Association. Vol. 73, No. 364, pp 699-705.

Yarnold, P.R., Hart, L.A., and Soltysik, R.C. (1994), Optimizing the Classification Performance of Logistic Regression and Fisher's Discriminant Analysis: Journal of Educational and Psychological Measurement, 54, 73-85.