IJST EXPLORE International Journal of Science and Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-7318

International Journal of Science and Technology Volume 1 No. 10, October, 2012

Assessment of Logistics Regression for Classification of Drug Data

S.S. Abdulkadir
Department of Statistics and Operations Research,
Federal University of Technology, Yola

ABSTRACT

The classification of individuals who involve in illicit drugs into drug peddlers and non-peddlers on the basis of oral evidence, type and the quantity of exhibit found with them usually pose problem for the purpose of prosecution. This paper uses logistic regression to classify offenders into the two dichotomous dependent variable (peddlers or non-peddlers) in order to ease the work of the agency, National Drug Law Enforcement Agency (NDLEA), responsible for illicit drugs. The data used in this study include age of offenders, type and weight of exhibit collected from the agency. The discriminant analysis was first used by Abdulkadir and Emmanuel (2010) to classify the illicit drug offenders into peddlers and non-peddlers based on the data. In literature the author discovered that the discriminant analysis cannot handle mixed data (discrete and continuous data) effectively instead logistic regression better fit the data. The model correctly classified 95.6% original grouped cases with positive and negative predictive values 97.35% and 91.55% respectively, which are higher than values obtained under the discriminant analysis. It was discovered that large quantity of exhibit has more effect than other variables, yet their inclusion significantly improves the outcome of the findings.

Keywords: *Logistic regression, classification, odds ratio, discriminant Analysis.*

1. INTRODUCTION

Logistic regression is a mathematical modeling approach that can be used to describe the relationship between independent or predictor variables to dichotomous dependent variable (Kleinbaum et al.1998). The model has traditional been appealing due to its performance in classification, the potential to use its output as probabilistic estimates since they are in the range[0,1], and interpretation of the coefficients in terms of the log-odds ratio. It is especially in biostatistical application where binary classification tasks occur frequently (Hastie et al.2001).Agresti (1996) asserted that relationships between the probability of success $\pi(x)$ and X are usually nonlinear rather than linear. Therefore a fixed change in X may have less impact when $\pi(x)$ is near 0 or 1 than $\pi(x)$ is in the middle of its range.

Logistic regression is often chosen if the predictor variables are a mixture of continuous and categorical variables and/or if they are not nicely distributed, that is, logistic regression makes no assumptions about the distribution of predictor variables. The logit analysis is usually employed if all the predictors are categorized while discriminant function analysis is used if all the predictors are continuous and nicely distributed (see Http://www.math.toronto.edu/mathnet). This paragraph motivated the author to re-analyze the data obtained on illicit drug using logistic regression analysis. The data was first used in a paper by Abdulkadir and Emmanuel (2010) using discriminant analysis while the data includes both qualitative and categorized predictor variables. In the paper the authors believed that the agency, 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- 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 involve 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 benefitted 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. (http://www.incb.org/). On this premise the author believe offenders, especially peddlers/traffickers deserve stiff penalty than the users, because without seller buyer 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

IJST © 2012 – IJST Publications UK. All rights reserved.

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 tumble 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.

Since the aim of the study is to determine whether or not an offender can be classified as a peddler/trafficker on the bases of age of offender, type, and weight of substance, therefore large quantity as never been reported in literature to have influenced the logistic regression model. However, some large quantity which could have been considered as outliers in the data collected was included/removed to determine the effect on the result of the study.

The logistic regression is employed for classification of objects into binary variable as

$$Y = \begin{cases} 1 & Drugpeddler \\ 0 & non-peddler \end{cases}$$

According to Ganesalingam (1989), classification of objects to groups is usually thought of as partition of objects into subsets in which the members are more similar. Classifying individuals into groups such that there is relative homogeneity within the groups and heterogeneity between the groups is a problem which has been considered for many years. For this paper the author intents to re classify objects into groups in which they were known to belong using logistic regression analysis.

2. LOGISTIC MODEL

Let X_1, X_2…, Xp be predictor variables which consist of qualitative and quantitative variables and Y as a dichotomous dependent variable, where Y is coded as 1 or 0 for its two categories as indicated above.

$$E(X) = \frac{1}{1+\exp[-(B_0 + \sum_{j=1}^{k} B_j x_j)]} \qquad (2.1)$$

This equation can be written in a form that describes the probability of occurrence of one of the two possible outcomes of Y, as follow

$$P(Y=1) = \frac{1}{1+\exp[-(B_0 + \sum_{j=1}^{k} B_j x_j)]} \qquad (2.2)$$

In general,

$$f(Z) = \frac{1}{1+e^{-z}} \text{ where } z = B^0 + \sum_{j=1}^{k} B_j X_j$$

The function f(z) is called logistic function. This function is well suited to modeling a probability, since the values of f(z) varies from-∞ to $+\infty$. The logistic model, therefore, is set up to ensure that, whatever estimate of risk we get, it always falls between 0 and 1. This is not true for other models, which is why the logistic model is often used when a probability must be estimated.

3. ESTIMATING THE ODDS RATIO USING LOGISTIC REGRESSION

The regression coefficients B^j in the logistic model given in (2.1) play an important role in providing information about the relationships of the predictors in the model to the exposure variable. The qualification of this relationship involves a parameter called the odd ratio (Kleinbaum *et.al* (1998))

The odd ratio is a measure of effect because it is a measure that compares two or more groups in predicting the outcome variable.

The odd of an event D = {Y=1}

$$\text{Odds (D)} = \frac{\Pr(D)}{1-\Pr(D)}$$

Any odds ratio is defined as a ratio of two odds.

The logistic regression is written as

$\text{Logit}[\Pr(Y=1)] = \log^e [\text{odds}(Y=1)]$

$$= \log^e \left[\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right].$$

Thus, equation (2.1) becomes

$$\text{Logit}[\Pr(Y=1)] = B^0 + \sum_{j=1}^{k} B_j X_j$$

The odds ratio, say, for groups A and B can be defined as

$$e^{\sum_{j=1}^{k}(X_{Aj}-X_{Bj})B_j}$$

IJST © 2012 – IJST Publications UK. All rights reserved.

In general the odds ratio for groups A and B in given by

$$\frac{\text{Odds for } X_A}{\text{Odds for } X_B} = \frac{e^{(B_0+\sum_{j=1}^{k} B_j X_{Aj})}}{e^{(B_0+\sum B_j X_{Bj})}} = e^{\sum_{j=1}^{k}(X_{Aj}-X_{Bj})B_j} \quad (3.1)$$

The constant term B_0 in the logistic model (1) drops out of the odds ratio expression in (3.1). The expression (3.1) describes a population odds ratio parameter because the B_j terms in the expression are themselves unknown population parameter. An estimate of this population odds ratio is obtained by fitting the logistic model using maximum likelihood estimation and substitute in the ML estimates B_j, together with the values of X_{Aj} and X_{Bj}, into the formula (3.1) to obtained a numerical value for the odds ratio.

4. RESULTS AND DISCUSSION

A total number of 184 cases of illicit drugs which comprises of 116 drug peddlers and 68 non-peddlers was extracted from the record of NDLEA in Yola and considered for this study. The average age of the peddlers is 28.11years with a standard deviation of 8.345 while the average age of non-peddlers is 27.67 years with standard deviation of 7.482.The exhibit type for the two groups is the same. The average weights of exhibit caught with peddlers and non-peddlers are 2288.36 grams and 122.68 grams respectively.

The table below depicts the Nagelkerke R square of 86% of the total variation in the outcome variable (drug status-peddlers and non-peddlers) explained by the logistic regression model fitted into the data.

Table 4.1:Model summary

Step	-2log likelihood	Cox & Snell R-square	Nagelkerke R –Square
1	58.989	0.631	0.862

The estimates of the logistic regression model parameters are presented in Table4.2 below. The Wald statistics for age and exhibit weights are 2.021 and 32.753 respectively. These show that the two are important risk factors to determine the status of illicit drug offenders. The exhibit weight is more important that the age going by the value of the Wald statistic, besides the Wald value for age is not significant while that of exhibit weight is highly significant (P-value of 0.000).The odds ratio for age is 1.069 meaning that an increase in age by one year age will increase the rate of peddling by 0.069(95% CI 0.975 to 1.173).

Table 4.2: Estimates of the Logistic Regression Model

	B	S.E	Wald	Df	Sig	Exp(B)	95% C.I for Exp(B) lower	Upper
Age	0.067	0.047	2.021	1	0.155	1.069	.975	1.173
Exhibit type	-1.171	1.905	0.378	1	0.839	0.310	.007	12.964
Exhibit Weight(grams)	-0.009	0.002	32.753	1	0.000	0.991	.988	.994
Constant	3.020		2.083		0.149	20.497	-	

The overall accuracy of this model (logistic regression) to predict becoming a drug peddler, with a predicted probability of 0.5 or greater) is 95.1% as shown in Table 4.3.The predictive model is z = 3.020 +0.067age -1.171Exhibit Type -0.009Exhibit Weight. The interest is to use this model (logistic regression) to predict the outcome for a new case. To determine how good the model is we computed the followings: The sensitivity from Table4.3 is 94.83%

Table 4.3: Classification Result (for logistic model)

Observed	Predicted		% Correct
	1	0	
1	110	6	94.8
0	3	65	95.6
Overall %			95.6

95.6% of original grouped cases correctly classified.
and specificity is 95.59.%. The positive predictive value (PPV) is 97.35% and the negative predictive value (NPV) is 91.55%. These results show that whenever we have a new subject, we can use logistic model to predict his probability of becoming a drug peddler. For instance, if given the age of a new subject, the weight and the type of exhibit caught with him and if the predictive model gives a low probability, it means that the subject is very

IJST © 2012 – IJST Publications UK. All rights reserved.

unlikely to become a drug peddler because the NPV reveals that we should be 91.55% confident and on the other hand if the model gives high probability it means that the subject is very likely to become a drug peddler because the PPV gives 97.35% confidence.

The summary results obtained under the discriminant analysis are presented in Tables4.4 and 4.5 to justify the need to use logistic regression when variables contain both discrete and continuous data.

Table 4.4: Test for Equality of Group Means (through Discriminant Analysis ANOVA tests)

	Wilks' Lambda	F	Df 1	Df 2	Sig
Y_1	0.993	1.792	1	260	0.182
Y_2	0.999	0.193	1	260	0.661
Y_3	0.925	21.114	1	260	000

The above Table 4.4 contains the results obtained when discriminant analysis was employed to classify the drug traffickers. In the table the Y_1 represents type of exhibit, Y_2 the age of offender and Y_3 the weight of exhibit. All the three variables used in the model except one, Y_3, is significant. This implies that only weight of exhibit can be used to predict whether a drug offender is a peddler of illicit drug or not.

Table 4.5 Classification Results (for Discriminant Analysis)

Observed	Predicted		% Correct
	1	0	
1	60	56	51.72
0	40	50	41.18
Overall %			46.45

46.45 % of original grouped cases correctly classified.

Comparing Tables 4.3 and 4.5 the overall percent grouped cases correctly classified is higher when logistic regression model was employ to classify the illicit drug offenders.The positive predictive value is 60% while the negative predictive value is 89.29%

5. CONCLUSION

The logistic regression model and discriminant analysis were used to classify data collected on illicit drug offenders. The results obtained when logistic regression was employed shows that 95.6% of the original grouped cases correctly classified while discriminant analysis correctly classified 46.45% of the same group. Moreover, the discriminant analysis results indicate that only weight of exhibit is significant in classification of offenders, but the logistic regression includes, in addition, the type of exhibit and age of offenders as important variables in the classification. Although the large quantity of exhibit has more effect than other variables, yet their inclusion significantly improves the outcome of the findings.

REFERENCES

[1] Abdulkadir,S.S.and Emmanuel,T(2010),Discriminant Analysis and Classification of Drug Peddlers and Non-Peddlers. Journal of Research in National Development.Vol8(1)

[2] Agresti,A.(1996). An Introduction to Categorical Data Analysis.John Wiley and Sons.New York

[3] Ganesalingam S., (1989), "Classification and Mixture Approaches to Clustering via Maximum likelihood."Applied Statistics,38.no 3,455 - 466

[4] Hastie, T; Tibshirani,R;and Friedman,J.H.(2001).The Elements of Statistical Learning.Springer.

[5] Izenman,A.J(2001),legal and Statistical aspect of the forensic study of illicit drugs,Statistical Science,16.

[6] Kleinbaum, D.G; Kupper, L.L;Muller, K.E,and Nizam,A(1998) Applied Analysis and Multivariate Methods. Third Edition. Duxbury Press.

[7] Odejide A.O. (1992), Drugs in the Third World. In Drugs and Society to Year 2000, Ed by Vamos and Corrivean Pg 116 – 119

[8] Http://www.math.toronto.edu/mathnet)

[9] http://www.incb.org/

IJST © 2012 – IJST Publications UK. All rights reserved.