Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.5, No.5, 2015

An Assessment Of The Performance Of Discriminant Analysis And The Logistic Regression Methods In Classification Of Mode Of Delivery Of An Expectant Mother

O.S. Balogun[1*], T.J. Akingbade[2], P.E. Oguntunde[3]

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

[2]Department of Mathematical Science, Kogi State University, Anyigba, Kogi State. Nigeria.

[3]Department of Mathematics, Covenant University, Ota, Ogun State. Nigeria.

[*]E-mail: stapsalms@yahoo.com

Abstract: The study compares two statistical methods: Discriminant analysis and the Logistic regression model in predicting Mode of Delivery of an expectant mother, Natural birth and Caesarian section. Of the 184 cases examined for Mode of Delivery of an expectant mother, Discriminant Analysis classified the Natural birth correctly (64.6%) while it recorded (64.7%) success rate in classifying the Caesarian section. In the case of the Logistic regression, it recorded (76.8%) and (52.9%) success rate in classifying the Natural birth and Caesarian section respectively. The overall predictive performance of the two models was high with the Logistic regression having the highest value (64.7%) and (65.8%) for Discriminant Analysis. Among the five characteristics examined, Mothers height, Baby's weight and gender were not significant variables for identifying Mode of delivery by both methods while Mothers weight is important identifying variable for both except Mothers age which was significant in the Discriminant analysis. 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 Mode of delivery for an expectant mother. However, given the failure rate to meet the underlying assumptions of Discriminant Analysis, Logistic Regression is preferable.

Keywords: Logistic Regression, Classification, Mode of delivery, Discriminant Analysis.

1. INTRODUCTION

Child birth poses considerable risk to the lives of both mother and child particularly in situations where complication arises. Child birth is defined as the complete expulsion or extraction of a fetus from its mother.

Child birth is preceded by a period known as the Gestation period. It has been of interest to researchers to know the mode of delivery a mother is likely to use. Under normal conditions, a mother is expected to give birth by natural birth otherwise known as safe delivery, but in certain cases complications may arise leading to the use of Caesarian section. Caesarian section poses considerable risk.

West/Central Africa accounts for more than 30% of global maternal deaths, and 162, 000 women died of pregnancy or childbirth related causes in 2005. The maternal mortality ration is substantially higher here than in any other region, at 1100 maternal deaths per 100, 000 live births. Furthermore, no discernible progress has been made in reducing the ratio since 1990. Of the 23 countries in the region with comparable estimates every country but Cape Verde has an MMR of at least 500, and a third of these countries have an MMR of 1, 000 or greater. Almost two thirds of maternal deaths in the region occur in the Democratic Republic of Congo, Niger and Nigeria, which together account for approximately 20 per cent of all maternal deaths world-wide. (UNICEF 2008: Progress for Children Report)

Several factors influence the high rate of maternal mortality in Nigeria, but the most common causes are lack of access to ante – natal care, inadequate access to skilled birth attendees, delays in the treatment of complications of pregnancy, poverty and harmful traditional practices.

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 = Natural birth and 1 = Caesarian section. If a low percentage of mode of delivery of expectant mothers based on the mode of delivery characteristics that has been properly classified, then the original selected expected mothers 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 dairy 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* (1982) 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.

Balogun *et al* (2014) compared the two methods in classifying and assessing drug offender characteristics, but concluded that the two methods gives closely 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 mode of delivery. This work determined if there is convergence between the two methods of analysis in classifying the subject (mode of delivery of an expectant mother) into one of the two populations (Natural birth and Caesarian section) 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 mode of delivery of an expectant mother. 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 Expectant mother characteristics (independent variables), and mode of delivery (dependent variable). The mode of delivery and expectant mother characteristics are listed in Tables 1and 2 respectively.

DISCRIMINANT ANALYSIS

Given a set of p independent variables $X_1, X_2, ..., X_p$, (Expectant mothers characteristics in this case), the technique attempt to derive a linear combination of these variables (Expectant mothers characteristics) which best separate or discriminates the two groups (Mode of delivery 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:

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

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 Mode of delivery into one of the two groups.

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 (Mode of delivery) 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:

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

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 $(Y = Natural\,birth, Y = Caesarian\,section)$. 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 (Mode of delivery).

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 β_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 O – 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:

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

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 in Table 3.

Table 3 shows the classification performances of the two methods. Of the 99 cases of Natural birth, discriminant analysis predicted correctly 64(64.6%) and misclassified 35(35.4%), while the logistic regression classified correctly 76(76.8%) and misclassified 23(23.2%). In the case of the prediction of the group membership of Caesarian section which contains of 85 cases, the discriminant analysis classified correctly 55(64.7%) of the cases and misclassified 30(35.3%) while the logistic regression classified 45(52.9%) cases correctly and

misclassified 40(47.1%) of the cases. The overall percentage correct classification of the Drug offenders was 64.7% and 65.8%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 mode of delivery. Table 4, since ($p > 0.05$) it is significant which suggest that model fitted the data well.

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, two of the independent variables- Mothers height, Baby's weight and gender were not significant ($p > 0.05$), while the remaining variable- Mothers weight and age 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 (Mode of delivery). The Wald statistic is used to test the null hypothesis that the coefficients of independent variables in the model are zero. From the table, only one of the Expectant mothers characteristics Mothers weight is significant with an associated $p < 0.05$. However, the four other variables Mothers Height, Age, Babies weight and gender were not significant.

However in comparison, both methods identified almost the same variable. Mothers Weight is significant for both methods, while Mothers height, Baby's weight and gender were equally not significant for the two methods. Both methods however differ in the estimation of Mothers age. 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.

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 some the classification variables were less than 0.05 while others were greater than 0.05, indicating that some of the variables were not normally distributed and others are normally distributed. The Box's M test value was (37.533, $p < 0.002$), indicating a valuation of the assumption of the discriminant Analysis.

4. CONCLUSION AND RECOMMENDATIONS

Using Mode of delivery of an expectant mother 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 (64.7% and 65.8%), and almost selected the same set of variables (Mothers Weight) is very significant to identifying expectant mothers mode of delivery. The finding agrees with Montgomery *et al* (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.

Balogun, O.S., Balogun, M.A., Abdulkadir, S.S. and Jibasen, D. (2014), A Comparison of the performance of Discriminant Analysis and the Logistic Regression methods in Classification of Drug Offenders in Kwara State. International Journal of Advanced Research, Vol. 2, Issue 10, page 280-286.

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 associated With 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 withDiscriminant Model; Tourism and Hospitality Research Journal, Vol. 10, Pp 171-187.

Kleinbaum, .D.G, Kupper, .L.L, Muller, .K.E., and Morgens-Tern, H. (1982), Epidemiologic Research: Principles and Quantitative Methods. Van Nostrand Reinhold Company, New York, p. 281-417.

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.

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.

UNICEF, (2008)**:** Progress for Children Report.

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.

Table 1: Mode of delivery (Dependent variable)

Group code	Mode of delivery
1	Natural birth
2	Caesarian section

Table 2: Expectant mothers characteristics (Independent variables)

Variable code	Description
X_1	Mothers height
X_2	Mothers weight
X_3	Mothers age
X_4	Baby's weight
X_5	Baby's gender

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

		Predicted Group Membership			
Actual Group	**No. of cases**	**Discriminant Analysis**		**Logistic Regression**	
		1	2	1	2
1	99	64(64.6%)	35(35.4%)	76(76.8%)	23(23.2%)
2	85	30(35.3%)	55(64.7%)	40(47.1%)	45(52.9%)
Overall % correctly classified		64.7%		65.8%	

Table 4: Hosmer-Lemeshow

Step	**Chi-square**	**Df**	**Sig**
1	6.667	8	0.573

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	-	8.089	-	3.454	0.021	0.063
Mothers height	0.990	-0.074	0.180	4.179	0.012	0.041
Mothers weight	0.974	0.047	0.028	1.806	0.032	0.179
Mothers age	0.977	0.083	0.042	0.144	0.289	0.704
Baby's weight	1.000	-0.199	0.940	1.710	0.310	0.191
Baby's gender	0.992	-0.783	0.226	0.849	3.570	0.357

Table 6: Test of Normality and equal covariance matrices

Group		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	**d.f.**	**Sig**	**Statistic**	**d.f.**	**Sig**
Mothers height	1	0.091	99	0.040	0.980	99	0.146
	2	0.080	85	0.200	0.975	85	0.095
Mothers weight	1	0.154	99	0.000	0.847	99	0.000
	2	0.089	85	0.095	0.921	85	0.000
Mothers age	1	0.121	99	0.001	0.911	99	0.000
	2	0.134	85	0.001	0.959	85	0.009
Baby's weight	1	0.092	99	0.037	0.984	99	0.264
	2	0.098	85	0.042	0.958	85	0.008
Baby's gender	1	0.353	99	0.000	0.635	99	0.000
	2	0.373	85	0.000	0.630	85	0.000