Comparative Study Of Double Discriminant Analysis And Logistic Regression Based On Binary And Contnuous Variables

By

Okonkwo, Evelyn Nkiruka
Nnamdi Azikiwe University, Awka, Nigeria

Onyeagu, Sidney I.
Nnamdi Azikiwe University, Awka, Nigeria

Okeke, Joseph Uchenna
Anambra State University, Uli, Nigeria

Nwabueze, Joy Chioma.
Micheal Okpala University of Agriculture,Umudike Nigeria

Ogbonna, Blessing
Nnamdi Azikiwe University, Awka, Nigeria

ABSTRACT

In the classification of an observation consisting of both binary and continuous variables, double discriminant analysis and logistic regression had been considered appropriate by most researchers. In this study, these two techniques were extensively discussed and compared using two real life data sets. The average value of PMC for the two data sets showed that logistic regression is optimal to double discriminant analysis in classifying objects whose exogenous variable consist of discrete and continuous variables.

Key words: Double discriminant analysis (DDA), Logistic regression, Regressor, and Probability of misclassification.

1.0 INTRDUCTION

Logistic regression allows one to predict outcome such as group membership from a set of variables that may be continuous, discrete, dichotomous, or a mix (Tabachnick and Fidell, 1996). The problem of discriminant analysis arises when one wants to predict group membership on

the basis of feature vector x. From the above two sentences, it is obvious that the same research questions can be answered by both methods. The logistic regression may be better suitable for cases when the dependant variable is dichotomous such as yes/no, pass/fail, infected/not infected, defective/good life/death, etc., while the independent variable can be on any scale. The discriminant analysis might be better suited when the dependant variable has two groups or more with the requirement that the independent variables will be normally distributed, linearly related, and each group has the same variance and covariance for the variables.

Several authors have formally compared these two techniques. For example, Halperin et al (1971) obtained results from several attributed type explanatory variable and noted only small difference in the classification ability between the two analytic procedures. Press and Wilson (1978) concluded that each analytic technique served a unique function: discriminant analysis was useful for classification of observations into one or two or more populations, whereas logistic regression was useful for relating a qualitative (binary) dependent variable to one or more explanatory variables by a logistic distribution functional form of P(E). Kleinbaum et al (1982) compared classification ability of logistic regression and discriminant analysis and noted that logistic model was slightly superior. Afuecheta et al (2010) compared these two methods on three data sets of normal and non-normal data and concluded that logistic regression is the more flexible and more robust method in the case of violation of linear discriminant assumptions.

The objective of this paper is to compare the performance of double discriminant function obtained using point-biserial model developed by Chang and Afifi (1974) , and logistic regression based on binary and continuous explanatory variables for classifying subjects into one of two populations.

2.0 DOUBLE-DISCRIMINANT ANALYSIS

An observation consisting of both binary and continuous variables may be classified into one of two populations by the double-discriminant function based on the point-biserial model. When the parameters are unknown or partially known, a sample double-discriminant function is obtained by replacing the unknown parameters by their sample estimates.

Suppose an observation $W = \begin{pmatrix} X \\ Y \end{pmatrix}$ is to be classified into one of two populations π_i, i= 1,2 where Y is px1 vector of continuous variates and X is a Bernoulli variate with P(X=1) $= \theta_i$ and P(X=0) = $1-\theta_i$ if W belongs to π_i. We assume that W follows a point-biserial model, that is, that the conditional distribution of Y given $X = x$ (0 or 1) is $N(\mu_{ix,}, \Sigma_x)$ when $W \in \pi_{i,}$ where Σ_x is a pxp positive definite matrix.. Under the point-biserial model, and given $X = x$, the likelihood ratio procedure is to classify the observation W into π_1 if

$$C_x = [Y - \frac{1}{2}(\mu_{1x} + \mu_{2x})]' \Sigma_x^{-1} (\mu_{1x} - \mu_{2x}) + \alpha_x \geq k \qquad (2.1)$$

where $\alpha_x = \ln \left\{ \left(\theta_1 / \theta_2 \right)^x \left[(1-\theta_1)/(1-\theta_2) \right]^{1-x} \right\}$, x = 0,1, and k is a given constant. Otherwise the observation is classified into π_2 .The discrininant function C_x in (2.1) is called the double-discriminant function by Chang and Afifi (1974).

If the parameters are unknown, we may replace them by their sample estimates. Let

$\left\{ W_{ij} = \begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix}, i = 1, 2 and\ j = 1, 2, ... N_i \right\}$ be two sequences of observation vectors independently drawn from π_1. We shall add a subscript x on Y_{ij} and N_i to indicate those values corresponding to $X_{ij} = x$. The sample double-discriminant function is defined according to whether α_x is known or unknown. When α_x is known but $\mu_{ix,}$ and Σ_x are unknown, the sample double-discriminant is

$$T_x = [Y - \frac{1}{2}(\bar{Y}_{1x} + \bar{Y}_{2x})]' S_x^{-1} (\bar{Y}_{1x} - \bar{Y}_{2x}) + \alpha_x \geq k \qquad (2.2)$$

where

$$\bar{Y}_{ix} = N_{ix}^{-1} \sum_{j=1}^{N_{ix}} Y_{ijx},$$

$$N_{i1} = \sum_{j=1}^{N_i} X_{ij}$$

$$N_{i0}=N_i - N_{i1},$$

$$S_x = M_x^{-1}\sum_{i=1}^{2}\sum_{j=1}^{N_{ix}}\left(Y_{ijx}-\bar{Y}_{ix}\right)\left(Y_{ijx}-\bar{Y}_{ix}\right)',$$

$$M_x = N_{1x}+N_{2x}-2.$$

N_i is number of observations in population i, and N_{ix} is the number of observations in x(1or 0) class of ith population.

When all the parameters are unknown, the double-discriminant function is

$$U_x = [Y-\frac{1}{2}\left(\bar{Y}_{1x}+\bar{Y}_{2x}\right)]'S_x^{-1}\left(\bar{Y}_{1x}-\bar{Y}_{2x}\right)+\hat{\alpha}_x \geq k \qquad (2.3)$$

$$\hat{\alpha}_x] = \ln\left[\left(N_{1x}N_2 \big/ N_{2x}N_1\right)\right].$$

It can be easily shown that T_x and U_x are invariant under any nonsingular linear transformation of the Y's.

As sample sizes N_{ix} tend to infinity, $\bar{Y}_{ix}\to\mu_{ix}$, $S_x\to\Sigma_x$, and $\hat{\alpha}_x\to\alpha_x$ in probability. Hence the limiting distribution of T_x tends to that of C_x (Tu 1978), $N\left(\frac{1}{2}D_x^2+\alpha_x, D_x^2\right)$ or $N\left(-\frac{1}{2}D_x^2+\alpha_x, D_x^2\right)$ depending on whether W comes from π_1 or π_2 =,where

$$D_x^2 = (\mu_{1x}-\mu_{2x})'\Sigma_x^{-1}(\mu_{1x}-\mu_{2x}) \qquad (2.4)$$

It should be noted that some of the N_{1x} may assume small values if θ_i is close to zero or one. Effort should be to ensure that N_{1x} > p (the number of parameters).

3.0 Logistic Regression

Logistic regression is part of a category of statistical models called generalized linear models (Agresti, 1996). Logistic regression, more commonly called logit model, deals with the binary case, when the response variable is dichotomous (i.e., binary 0 or 1). The predictor variables may be quantitative, categorical, or a mixture of the two. This model is mainly used

to identify the relationship between one or more explanatory variables X_i and response variable Y. It has been used for prediction and determining the most influencing explanatory variable(s) on the variables (Cox and Snell, 1994). Instead of a straight line, logistic regression seems preferable to fit some kind of sigmoid curve to the observed points. The tails of sigmoid curve level off before reaching P(E) =0 or P(E) = 1, so that the problem of impossible values of P(E) is avoided.

The basic form of the logistic function is

$$P = \frac{1}{1+e^{-Z}} \qquad (3.1)$$

where Z is the predictor variable(s) and e is the base of the natural logarithm, equal 2.71828… ,P is estimated probability of event occurring.
In the multivariate case, Z instead of being a single predictor variable, is a linear function of a set of predictor variables:

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p. \qquad (3.2)$$

3.1 One Regressor

Assuming that we have a single regressor, let us try to write a simple linear regression model as

$$Y = \beta_0 + \beta_1X + \varepsilon \qquad (3.3)$$

We would logically let $Y_i = 0$ if the unit does not have the characteristic that Y represents, and $Y_i = 1$ if the unit does have the characteristic.
It thus follows that ε_i can also take on only two values: $1 - \beta_0 + \beta_1X$ if $Y_i = 1$ and $-\beta_0 + \beta_1X$ if $Y_i = 0$. Therefore ε_i cannot be approximately normally distributed. Consequently, the model (3.3) is inapplicable for a binary dependent variable.
In simple linear regression, the starting point in determining a model is a scatter plot of Y. Consequently; it is necessary to consider other plots. One such plot is a plot of E(Y\X) against X. Rather than plotting points, we must postulate a relationship between Y and X variables since the ordinate of the plot is not related to the data. It is customary to let $E(Y_i\backslash X_i) = \pi_i$,which is $P(Y_i = 1)$, where Y is binomial. Here π_i represents the probability of, for example, someone dying within a stated time period who has B.P given by

X_i. Given one regressor the probability of an event, say E, for a given value of X, P(E) = $\pi_i(X)$ is

$$\pi_i(X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)} \qquad (3.4)$$

The model given by (3.4) satisfies the important requirement that $0 \le \pi_i \le 1$ and will be satisfactory model in many applications. The model in terms of Y would be written as

$$Y = \pi(X) + \varepsilon$$

It follows from (3.4) that

$$\frac{\pi}{(1-\pi)} = exp(\beta_0 + \beta_1 X)$$

so

$$\log \frac{\pi}{(1-\pi)} = \beta_0 + \beta_1 X \qquad (3.5)$$

Since (3.5) results from using a logistic transform (also called a logit transform), the model is called a logistic regression model. The left side of (3.5) is called log odds ratio, and this can be explained as follows. Since π = P(Y=1), it follows that 1-π = P(Y=0), and so $\frac{\pi}{(1-\pi)}$ is the ratio of the two probabilities, which, when stated in the form the of odds, gives the odds of having Y=1, for a given value of X. Odds are frequently stated in terms of "against" rather than "for" so the odds against having Y=1 would be obtained from $\frac{1-\pi}{\pi}$.

The absence of error on the right side of (3.5) is because the left side is a function of (Y/X), instead of Y, which serves to remove the error term.

The interpretation of β_1 is naturally somewhat different from the interpretation in the linear regression. In (3.5) β_1 obviously represents the amount by which the log odds change per unit change in X. This implies that a unit increase in X increases the odds by the multiplicative factor e^{β_1}.

4.0 DATA AND THEIR ANALYSIS

4.1 The Data for this study are two real life data sets. One set was collected from Amaku General Hospital, Awka, Anambra state, Nigeria. The data is on fasting and non-fasting blood sugar level (FBS and NFBS respectively)

of diabetics and non-diabetics patients with their gender randomly selected from the cases reported at the Hospital in 2009.
The laboratory reference ranges for adults are:

Glucose (fasting): 9 – 4mmol/l and Glucose (non-fasting):4 – 8mmol/l

The other set is on four albino CD-1 Sprague-Dawley rats at the weaning age with similar weights. The data is available in Tu and Han (1982).

4.2 FINDINGS

4.21 Result of diabetic and non-diabetic patient data

Double discriminant analysis

The sample double discriminant functions (DDF) are:

$T_0 = 1.47732y_1 + 1.32279y_2 – 19.55786$ for male.

$T_1 = 1.33301y_1 + 2.04684y_2 – 24.06988$ for female.

The above two sample double discriminant functions when applied on the original data gave the probability of misclassification as: 0.05

Result of Logistic Regression

The logistic regression model for the data is

$Z= - 960.615 + 125.379y_1 +26.1447\ y_2 – 13.6178x$

The p- value of the model is: 0.0000

The probability of correct classification is 1.00.

Here y1=fastings blood glucose level, y2= non-fastings blood glucose level, and x = sex

4.22 Result of four albino CD-1 Sprague-Dawley rats

The sample double discriminant functions (DDF) are:

$T_0 = 0.020544y_1 + 0.04306y_2 - 7.585704$ for male.

$T_1 = 0.040258y_1 + 0.214881y_2 + 3.596135$ for female.

The above two sample double discriminant functions when applied on the original data gave the probability of misclassification as: 0.327

Result of Logistic Regression

The logistic regression model for the data is

$Z= 6.93092 -0.0228665y_1 + 0.0044232\ y_2 - 2.67589X$

The p- value of the model is: 0.0000

The probability of correct classification is 0.74 with ties of 0.04 and the probability 0f misclassification is 0.22

Here y1=body weight, y2= total length, and x = sex

Summary of findings

The frequencies of misclassifications are listed in the following table.

Table 4.1

Sample data	DDF	Logistic
Diabetic and non-diabetic patients	0.05	0.00
Albino CD-1Sprague-Dawley rats	0.327	0.26
Average	0.1885	0.13

5.0 CONCLUSIONS AND DISCUSSION.

Double discriminant analysis (DDA) and logistic regression are used when the observations of exogenous variable consist of binary and continuous

variables with dichotomous dependent variable. We gave extensive discussion on the similarities and dissimilarities of the two methods in the literature. From the average value of probability of misclassification of 0.1885 and 0.13 for DDA and logistic regression analysis respectively, we can conclude that logistic regression is optimal to DDA in classifying objects whose exogenous variable consist of discrete and continuous variables.

Reference

1. Afuecheta, E.O., Ogum, G.E.O., Osuji, G.A., and Utazi, C.E. (2010). Comparison of Linear Discriminant Analysis and Logistic regression in Classification Problems. Conference Proceedings Nigeria Statistical Association, 81-89.
2. Agresti, A. (1996). *An introduction to Categorical Data Analysis*. John Wiley& Sons, Inc., New York.
3. Chang, P.C., and Afifi, A. A.(1974). Classification Based on Dichotomous and Continuous Variables. *Journal of the American Statistical Association*, 69: 336-339.
4. Cox, D.R., and Snell, E.J. (1994). *Analysis of Binary Data*. Chapman& Hall, London.
5. Halperine, M., Blackwelder W.E., and Verter, J.I (1971). Estimation of the Multivariate Logistic Risk Function and Maximum Likelihood Approach. *Journal of Chron. Dis.* 24:125-158.
6. Kleinbaum D.G., Kupper, L.L., and Morgenstern, H. (1982). Epidemiology Research: Principles and Quantitative Methods.

Toronto: Lifetime Learning 461-470.

7. Krzanowski, W.J. (1993). *Principles of Multivariate Analysis.* Oxford University Press Inc., New York. 337-345.

8. Press, S.J., and Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of American Statistical Association* 73:699-705.

9. Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics* Harper Collins, New York.

10. Tu, C.T. (1978). Discriminant Analysis Based on Binary and Continuou Variables "Unpublished Ph.D Dissertation", Iowa State University.

11 Tu, C.T., and Han C.P. (1982). Discriminant analysis Based on Binary and continuous variable. *Journal of American Statistical Association* 77:447-454.