IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 9, Issue 4 (Jan. 2014), PP 22-26
www.iosrjournals.org

A Study of Classification and Discriminant Analysis of Infants at Birth in Nigeria

Adejumo, A. O and Onyenekwe, C. E
Department of Statistics, University of Ilorin, Ilorin, Nigeria

Abstract: *In Nigeria there is no recognized scientific method of discriminating and classifying babies statistically into groups of study.*

The purpose of this study includes to set up a discriminant function and classification rule that can be used to classify babies into two groups; to estimate the proportion of observations in each of the prior group; and to estimate the probability of correct classification and misclassification respectively.

To this effect, a sample of 270 cases (infants) was observed with the following measurements: Age of mother (x_1), weight at 36th week (x_2), birth weight (x_3), Parity (x_4), Gestation Period (x_5), and sex of the baby (x_6). The birth weight was used to do the initial classification. Group 1 termed underweight (< 2.5kg) and Group 2 termed normal weight (≥ 2.5kg). We observed that the Discriminant Function $\mathbf{Z}= -0.02947228X_1- 0.0514773X_2- 8.130338X_3 + 0.062259X_4 + 0.0946538X_5 + 0.5888918X_6$. *Also 95.8 % of the original grouped cases were correctly classified. The percentage of misclassification is 4.15%. Conclusively the measure of the predictive ability which is the percentage of correct classification shows that discriminant analysis can be used to predict infants into two classes of weight and can also be used to predict group membership of any subject matter.*

Keywords: *Dicriminant, Classification, Multivariate, Misclassification, Gestational age, Confusion Matrix*

I. Introduction

Birth weight is an important measure for assessing future growth patterns of a child and investigating both immediate health risks and those in later life. It is thus a key variable in any longitudinal study of child health.

Birth weight is the weight of the child at the point of birth. It can also be defined as the first weight of an infant obtained within the first sixty (60) minutes after birth. A high birth weight is when the infant weight is greater than 4kg, a full size or normal infant weight is between 2.5kg and 4kg, a low birth weight is when the infant weight is less than 2.5kg. Low birth weight (i.e. less than 2.5kg) increase prenatal mortality for reasons ranging from prematurity to placental insufficiency. During pregnancy, a baby's birth weight can be estimated in different ways. The height of the fundus (the top of the mothers' uterus) can be measured from the public bone. This measurement in centimetres usually corresponds with the number of weeks of pregnancy after the 20th weeks, see [1].

Clinical representation of birth weight of the baby varies since many thing affect the baby's size at birth. Some do not imply a problem and smaller weights are not always indicative of growth retardation. A woman's first child usually weighs less than her subsequent babies also mothers who themselves were large when newly born tend to have large babies. The size of the baby's father also plays a role in its birthweight (see [2], [3], and [4]).

In Nigeria there is no recognized scientific method of discriminating and classifying babies statistically into groups of study. The suggested classification is by birth weight, this has prompted the researcher to set up the scientific method to a model for classification.

The main objectives of this research paper are to: set up a discriminant function and classification rule that can be used to classify babies into two groups (Group 1- Babies with underweight. < 2.5kg, Group 2 – Babies with normal weight ≥ 2.5kg); estimate the proportion of observations in each of the prior group; estimate the probability of correct classification; and estimate the probability of misclassification.

II. Methodology

Multivariate Analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. Measurements are referred to as variables and individuals or objects as units (research units, sampling units, or experimental units) or observations. In practice, Multivariate data sets are common, although they are not always analyzed as such. But the exclusive of univariate procedures with such data is no longer excusable, given the availability of multivariate techniques and inexpensive computing power to carry them out.

Historically, the bulk of applications of multivariate techniques have been in the behavioural and biological sciences. However, interest in multivariate methods has now spread to numerous other fields of investigations (see [5], [6], and [7]).

Multivariate Analysis can be said to be concerned with two areas, descriptive and inferential statistics. In the descriptive realm, we often obtain optimal linear combinations of variables. The optimality criterion varies from one technique to another, depending on the goal in each case. Although linear combinations may seem too simple to reveal the underlying structure, we use them for two obvious reasons:

They have mathematical tractability (linear approximations are used throughout all science for the same reason), see [8].

They often perform well in practice. These linear functions may also be useful as a follow-up to inferential procedures when we have a statistically significant test result that compares several groups. For example, we can find the linear combination (or combinations) of variables that led to the rejection of the hypothesis, then the contribution of each hypothesis is of interest (see [9], and [10]).

Multivariate inference is especially useful in curbing the researchers' natural tendency to read too much into the data. Total control is provided for experiment wise error rate: that is no matter how many variables are tested simultaneously, the value of α (the significance level) remains at the level set by the researcher [11].

Research in behavioural sciences mostly involves developing prediction and classification models. Discriminant function Analysis also called Discriminant Analysis is used to classify cases into the values of a categorical dependent, usually a dichotomy.

Discriminant analysis is a statistical technique that is used to classify the dependent variable between two or more categories. Discriminant Analysis also has a regression technique which is used for predicting the value of the dependent categorical variable. Discriminant analysis may be applied in a number of settings: in the armed forces, it is used in assigning new personnel to training programs. In the industry, it is used in assigning new employees to a particular job category. In health, it is used in classifying a patient into one of several diagnostic categories. In education, it is found useful as aids in both educational and vocational counseling. (see [1], [6], [12], [13], [14], [15], [16], and [17])

This concern for the classification ability of the linear discriminant function has obscured and even confused the fact that two very distinct purposes and procedures for conducting discriminant analysis exist. The first procedure, discriminant predictive analysis is used to optimize the predictive functions, that is the objective is to develop an equation that maximally discriminate the groups using P independent variables. The second procedure, discriminant classification analysis uses the predictive functions derived in the first procedure to either classify fresh sets of data of known group membership, there by validating the predictive function; or if the function has previously been validated (see [18], [19], and [20]).

The goal of discriminant analysis include identifying the relative contribution of the P variables to separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration of the groups (see [21], and [22]).

A discriminant function is also gotten which is a linear combination of these P variables that maximizes the distance between the two populations (groups) mean vector

Z= a^{i}x transforms each observation vector to a scalar in

$$Z_{11} = a^1 X_{1is} = a_1 X_{1i1} + a_2 X_{1i2} + a_3 X_{1i3} + \cdots + a_p X_{1ip}, \quad i = 1,2, \ldots n_1$$

$$Z_{21} = a^1 X_{2i} = a_1 X_{2i1} + a_2 X_{2i2} + a_3 X_{2i3} + \ldots + a_p X_{2ip}, \quad i = 1,2, \ldots n_2$$

Hence, the $n_1 + n_2$ original observations vectors in the two samples:

$$\begin{array}{cccc} Z_{11} & Z_{12} & \ldots & Z_i n_i \\ . & . & . & . \\ . & . & . & . \\ Z_{21} & Z_{22} & \ldots & Z_2 n_2 \end{array}$$

Note that $\overline{Z}_1 = a^1 \overline{x}_1 , \overline{Z}_2 = a^1 \overline{x}_2$

$$\sum\left(\frac{Z_{1i}}{n_1}\right) = a^1 \sum\left(\frac{x_{1i}}{n_1}\right) \text{ since } \overline{x_1} = \sum\left(\frac{x_{1i}}{n_1}\right)$$

$$\sum\left(\frac{Z_{2i}}{n_2}\right) = a^1 \sum\left(\frac{x_{2i}}{n_2}\right) \text{ since } \overline{x_2} = \sum\left(\frac{x_{2i}}{n_2}\right)$$

The vector *a* that maximizes the standardized difference $\left(\overline{Z}_1 - \overline{Z_2}\right)/S_2$ since $\left(\overline{Z}_1 - \overline{Z_2}\right)/S_2$ can be negative, the squared distance $\left(\overline{Z}_1 - \overline{Z_2}\right)/S_1^{\ 2}$ is used which can also be expressed as

$$\left(\left(Z_1 - Z_2\right)/S_1^{\ 2}\right) = \left(\left(a^1\overline{x}_1 - a^1\overline{x}_2\right)^2 / a^1 spca\right) = \left(a^1\left(\overline{x}_1 - \overline{x}_2\right)\right)^2 / a^1 spca$$

Where $spc = \left[\left(n_1 - 1\right)s_1 + \left(n_2 - 1\right)s_2\right]/\left(n + n - 2\right)$

$$a^1 = spc^{-1}\left(\overline{x}_1 - \overline{x}_2\right)$$

The confusion matrix

	π_1	π_2
π_1	Correct classification	P(2/1)
π_2	P(1/2)	Correct classification

Apparent Error rate = $\dfrac{P(2/1) + P(1/2)}{N}$

Where N = $n_1 + n_2$

P= Population Group

III. Analysis And Results

The data used for this research work comprises of 265 Delivery Mothers with the following measurements on each (Age of mother, weight at 36th week, Weight of baby at birth (kg), Parity, Gestation Period and sex).

Table 1: Summary of Variables Under Consideration for the Two Groups.

GROUP	MEAN	STANDARD ERROR	N
UNDERWEIGHT(1)			
Age of mother	28.1684	6.09962	95
Weight At 36	66.2105	10.53555	95
Birth weight	2.0377	.36902	95
Parity	1.6000	1.72220	95
Gestation Period	36.6632	1.14493	95
Sex	1.5158	.50240	95
NORMAL WEIGHT(2)			
Age of Mother	29.4412	5.12097	170
Weight At 36	69.0324	9.88620	170
Birth weight	3.1194	.36906	170
Parity	1.8353	1.70518	170
Gestation Period	36.8353	1.65945	170
Sex	1.4706	.50061	170

Group 1 (X_1)

$$\begin{array}{c|cccccc}
 & X11 & X12 & X13 & X14 & X15 & X16 \\
\hline
x11 & 37.20537514 & 3.9949608 & 0.46879843 & 3.68510638 & 0.74882419 & -0.07715566 \\
x12 & 3.99496081 & 66.9216125 & 0.55654535 & 2.03191489 & 0.28667413 & 0.46528555 \\
x13 & 0.46879843 & 0.5565454 & 0.13617330 & 0.02140426 & 0.01453080 & -0.02081411 \\
x14 & 3.68510638 & 2.0319149 & 0.02140426 & 2.96595745 & 0.24680851 & -0.07872340 \\
x15 & 0.74882419 & 0.2866741 & 0.01453080 & 0.24680851 & 1.31086226 & -0.02653975 \\
x16 & -0.07715566 & 0.4652856 & -0.02081411 & -0.07872340 & -0.02653975 & 0.25240761
\end{array}$$

Group 2 (X_2)

$$\begin{array}{c|cccccc}
 & X21 & X22 & X23 & X24 & X25 & X26 \\
\hline
x21 & 26.5236686 & 3.8381657 & -0.16538462 & 4.27810651 & -1.26627219 & 0.20118343 \\
x22 & 3.8381657 & 97.6902837 & -0.74063871 & 2.18920989 & 0.46293073 & -0.32158023 \\
x23 & -0.1653846 & -0.7406387 & 0.13612983 & 0.02984337 & 0.05624086 & 0.03045597 \\
x24 & 4.2781065 & 2.1892099 & 0.02984337 & 2.90762269 & -0.33303167 & 0.15489036 \\
x25 & -1.2662722 & 0.4629307 & 0.05624086 & -0.33303167 & 2.75753568 & -0.11103376 \\
x26 & 0.2011834 & -0.3215802 & 0.03045597 & 0.15489036 & -0.11103376 & 0.25060912
\end{array}$$

Discriminant Function

Z= $-0.02947228X_1-0.0514773X_2-8.130338X_3+0.062259X_4+0.0946538X_5+0.5888918X_6$

Classification Result

The classification rule is to assign to P1 if Z1> Z and to P2 if otherwise

Table 2: Confusion Matrix

	P1	P2
P1	94	1
P2	10	160

Table 3: Predicted Group Membership

PREDICTED GROUP MEMBERSHIP			
GROUPING	UNDERWEIGHT(1)	NORMAL WEIGHT(2)	TOTAL
UNDERWEIGHT(1)	94	1	95
NORMAL WEIGHT(2)	10	160	170
% UNDERWEIGHT(1)	98.9	1.1	100
% NORMAL WEIGHT(2)	5.9	94.1	100

The Probability of misclassification $=(1+10)/265=0.0415$

The Percentage of misclassification = 4.15%

The Probability of correct classification $=(94+160)/265=0.958$

The Percentage of correct classification= 95.8%

This implies that 95.8 % of original grouped cases correctly classified.

IV. Conclusion

Discriminant score was gotten and it was used in classifying the variables into two groups. 95.8% of the original grouped cases were correctly classified while 4.15% were misclassified.

The probability of misclassification is 0.0415. That is to say that the level of accuracy is high and significant.

In Summary, the measure of the predictive ability which is the percentage of correct classification shows that discriminant analysis can be used to predict infants into two classes of weight and can also be used to predict group membership of any subject matter.

Reference

[1] Adimora, G.N, Nigerian Journal of Clinical Practices. Vol 7.2004, pg 33- 36 Official Publication of the Medical and Dental Consultants Association of Nigeria.

[2] Deswal B. S., Singh J. V., Kumar D.- A Study of Risk Factors for Low Birth Weight, Indian J. Community Med. 24, 2008: 127-131

[3] Philip, J. Disala, *Clinical Gynaecologicon.cology*.(4th Edition). Mosby year book Inc. 1995

[4] Geoffrey .V.P Chamberline, *Gynaecology by Ten Teachers*.(16th Edition). ELBS. 1988

[5] Daniel B.Rowe, *Multivariate Bayesian Statistics.* Chapman and Hall/CRC, 2003.

[6] Anderson, T.W . *An Introduction toMultivariate Statistical Analysis.* New York: Wiley. 1984.

[7] Barnett,V. (ed.), *Interpreting Multivariate Data.* Wiley. 1981

[8] David,W.Stockburger, *Multivariate Statistics: Concepts, Models and Applications*. 1998

[9] Nwobi and Nduka . *Statistical Notes and Tables for Research*, Second Edition. Alphabet Nigeria Publishers. 2003

[10] Nwachukwu V.O, *Principles of Statistical Inference*. Second Edition. Zelon Enterprises. 2006

[11] Everitt, B.S & Dunn, G, *Applied Multivariate Statistical Analysis.* (2nd Edition). Arnold. 2001

[12] Alvin B.Rowe, *"Methods of Multivariate Analysis*, second edition" Wiley –Interscience, 2003.

[13] Lachenbruch,P.A, *Discriminant Analysis*. New York: Hafner. 1975

[14] Evans, F.B, "Physiological and Objective Factors in the Prediction of Brand Choice: Ford vs Chevrolet," Journal of Business 32 (October) 1959: pp.340-369.

[15] Albaum, G, "*Applying Discriminant Analysis to Unipolar Semantic Scaling Data*," American Institute of Decision Sciences Western Meetings. 1975

[16] Johnson, R.A and Wichern, D.W, *Applied Multivariate statistical Analysis*. (5th edition). Patience-Hall.2002.

[17] Fisher,L., and J.W. Vanness , "Admissible Discriminant Analysis," Journal of the American Statistical Association 68,1973: pp.603-607.

[18] Dillion, W.r., and M Golgstein , Multivariate Analysis Methods and applications. John Wiley & Sons Inc. Canada.1984

[19] Hand, D. J, *Discrimination and Classification*. New York; Wiley 1981.

[20] Hand,D.J, *Construction and Assessment of Classification Rules*. New York: Wiley. 1999.

[21] Huberty, Carl J. *Applied Discriminant analysis*. New York: Wiley-Interscience.1994.

[22] Mclachlan, G.J , *Discriminant Analysis and Statistical pattern recognition*. New York: Wiley. 1992.