Indian Scholar
ISSN 2350-109X
www.indianscholar.co.in
An International Multidisciplinary Research e-Journal

APPLICATION OF DISCRIMINANT ANALYSIS ON DISCRIMINATORY FACTORS OF HYPERTENSION AMONG ADULTS

A.E. Anieting
Department of Mathematics and Statistics
University of Uyo,Uyo
Nigeria

Mosugu, J. K.
National Open University of Nigeria
Abuja, Nigeria

Abstract

This study applied discriminant analysis to examine the discriminatory factors in Adults. Three risk factors- Age, Body Mass Index and Waist circumference were considered. The result showed that the three predictors jointly and individually discriminate between hypertensive and non-hypertensive. The result also revealed that Age and BMI were the major discriminating factors.

1.0 INTRODUCTION

The basic purpose of discriminant Analysis is to estimate the relationship between a single categorical dependent variable and a set of quantitative independent variables. It is capable of handling either two groups or multiple groups. When three or more classifications are identified the technique is referred to as Multiple Discriminant Analysis. It also involves deriving a variate the linear combination of two or more independent variables that will discriminate best between defined groups. Discriminant analysis was devised in 1930 by the combined effect of Fisher, Hoteling and Mahalanobis in UK, US and India respectively. As a graphical version of MANOVA, it can be used to compliment the findings of cluster analysis and Principal Component Analysis.In an investigation, Olubuyide et al (2008) used discriminate function analysis to investigate the relative value of six biochemical parameters in the diagnose of liver disease. Four groups totally 70 subjects were used. It was discovered that about 66% of all individuals could be correctly assigned to one of the four groups using biochemical markers alone. The aim of this study is to determine the discriminatory variable for hypertension, to build a discriminant function for hypertension using age, BMI and Waist circumference as

predictors and to develop a classification criteria for hypertension. The data collected involved 490 adults in Uyo as recorded by the records department of the University of Uyo teaching hospital.

2.0 METHODOLOGY

2.1 Box's M Test for the Equality of Covariance Matrices

One of the assumptions of discriminant analysis is the equality of covariance matrices. In order to check this assumption, the Box's M Test was used.

Test Statistic:

$$M = (N-1)\log_e|S_{pl}| - \sum_{i=1}^{2}(n_i - 1)\log_e|S_i| \qquad (2.1)$$

In which S_i is the covariance matrix of the i^{th} sample and S_{pl} is the pooled sample covariance matrix.

$$S_{pl} = \frac{\sum_{i=1}^{k} \upsilon_i S_i}{\sum_{i=1}^{k} \upsilon_i} = \frac{E}{\upsilon_E}$$

The F- approximation for the distribution of M can be referred to as the Box's M-test. It is calculated as follows:

$$c_1 = \left[\sum_{i=1}^{k}\frac{1}{\upsilon_i} - \frac{1}{\sum_{i=1}^{k}\upsilon_i}\right]\left[\frac{2p^2+3p-1}{6(p+1)(k-1)}\right]$$

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k}\frac{1}{{\upsilon_i}^2} - \frac{1}{\left(\sum_{i=1}^{k}\upsilon_i\right)^2}\right]$$

$$a_1 = \frac{1}{2}(k-1)p(p+1), \quad a_2 = \frac{a_1+2}{|c_2 - c_1^2|}, \quad b_1 = \frac{1 - c_1 - \frac{a_1}{a_2}}{a_1},$$

$$b_2 = \frac{1 - c_1 - \frac{2}{a_2}}{a_2}$$

$$if\ c_2 > c_1^2,$$

$$F = -2b_1 \ln M \text{ is approximately } F_{a_1,a_2}$$

$$if\ c_2 < c_1^2,$$

$$F = \frac{2a_2 b_2 \ln M}{a_1(1 + 2b_2 \ln M)} \text{ is approximately } F_{a_1,a_2}$$

In either case, H_o is rejected if $F > F_\alpha$. If $\upsilon_1 = \upsilon_2 = \cdots = \upsilon_k = \upsilon$, then c_1 and c_2 simplifies to

$$c_1 = \frac{(k+1)(2p^2+3p-1)}{6k\upsilon(p+1)}, \quad c_2 = \frac{(p-1)(p+2)(k^2+k+1)}{6k^2\upsilon^2}$$

Decision Rule: Reject H_0 if p<0.05 otherwise accept H_0 at the 5% level of significance. It is possible to have classifications into two or more multivariate normal populations, but in this case, we shall limit ourselves to classifications into two normal populations denoted by $\pi_1$and π_2. Suppose we have two multivariate normal populations with equal variance-covariance matrices,

$$N(\mu_1, \sum) \; and, N(\mu_2, \sum) where, \mu_i (i = 1,2), (\mu_1, \mu_2, \dots, \mu_p)'$$

is the vector of means of the i^{th} population and is the variance-covariance matrices of the two populations. The pdf of i^{th} population is given below:

$$P_i(X) = \frac{1}{(2\pi)^{\frac{P}{2}} |\sum|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(X-\mu_i)' \sum^{-1} (X-\mu_i)\right] \qquad (2.2)$$

The ratio of the densities of two multivariate normal populations is given below;

$$exp\left[-\frac{1}{2}\{(X-\mu_1)'\Sigma^{-1}(X-\mu_1) - (X-\mu_2)'\Sigma^{-1}(X-\mu_2)\}\right] \geq k$$

$$\frac{P_1(X)}{P_2(X)} = \frac{exp\left[-\frac{1}{2}(X-\mu_1)'\Sigma^{-1}(X-\mu_1)\right]}{exp\left[-\frac{1}{2}(X-\mu_2)'\Sigma^{-1}(X-\mu_2)\right]} \geq k$$

Taking the natural logarithms of the first inequality above; which is monotone increasing we have:

$$-\frac{1}{2}\{(X-\mu_1)' \sum^{-1} (X-\mu_1) - (X-\mu_2)' \sum^{-1} (X-\mu_2)\} \geq \log k \qquad (2.3)$$

The second term of (2.3) above is the Mahalanobis square distance between and for k suitably chosen (which of course can be one and then log k will be zero), the LHS of (2.3) can be expanded and rearranged to obtain the following:

$$X' \sum^{-1} (\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)' \sum^{-1} (\mu_1 - \mu_2) \geq \log k \qquad (2.4)$$

The first term of the inequality in (2.4) is the well-known Fisher's linear discriminant function which is linear in the component of the observation vector.

2.2 Wilk's Lambda Test.

This is done to determine the difference between the two set of variables

Test Statistic:

$$\wedge = \frac{|E|}{|E+H|} \qquad (2.5)$$

$$p = number\ of\ variables$$
$$v_H = degrees\ of\ freedom\ for\ hypothesis$$
$$v_E = degrees\ of\ freedom\ for\ error$$

Where;

$$v_H = k - 1 \; and \; v_E = \sum_{i=}^{k} n_i - k = N - k$$

Between Group Matrices (H)

$$\mathrm{H} = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i.} - \bar{y}_{..})'$$

With only two groups H becomes;

$$\mathrm{H} = n\sum_{i=1}^{2}(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i.} - \bar{y}_{..})'$$

This can be expressed as;

$$H = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_{1.} - \bar{y}_{2.})(\bar{y}_{1.} - \bar{y}_{2.})' \qquad (2.6)$$

Between Group Matrix (E)

$$\mathrm{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{ij} - \bar{y}_{i.})(\bar{y}_{ij} - \bar{y}_{i.})'$$

$$= \sum_{i=1}^{k}(n_i - 1)\mathrm{S}_i$$

$$E = (n_i - 1)S_1 + (n_i - 1)S_2 + \ldots + (n_i - 1)S_i \qquad (2.7)$$

The hypothesis is rejected when the value of Λ exceeds the upper α-level percentage point of the F-distribution, with degrees of freedom as shown below;
An approximate F-statistic is given as:

$$F = \frac{1 - \wedge^{\frac{1}{t}}}{\wedge^{\frac{1}{t}}}\frac{df_2}{df_1} \qquad (2.8)$$

$$\mathrm{F}_{critical} = \mathrm{F}_{p,\upsilon_E - p + 1}$$

$$df_1 = p\upsilon_{\mathrm{H}}, df_2 = wt - \frac{1}{2}(p\upsilon_{\mathrm{H}} - 2)$$

$$w = \upsilon_{\mathrm{E}} + \upsilon_{\mathrm{H}} - \frac{1}{2}(\mathrm{p} + \upsilon_{\mathrm{H}} + 1), \mathrm{t} = \sqrt{\frac{\mathrm{p}^2\upsilon_{\mathrm{H}}{}^2 - 4}{\mathrm{p}^2 + \upsilon_{\mathrm{H}}{}^2 - 5}}$$

Decision Rule

Reject H_0 if p <0.05 ($F_{cal} > F_{critical}$) otherwise accept H_0 at the 5% level of significance. Also, Reject H_o if $\Lambda \leq \Lambda_{\alpha,p,\upsilon_H,\upsilon_E}$

.

2.3 ESTIMATING MISCLASSIFICATION RATES

Among the $n1$ observations in π_1, $n11$ are correctly classified into π_1, and $n12$ are misclassified into π_2, where $n_1 = n_{11} + n_{12}$. Similarly, of the n_2 observations in π_2, n_{21} are misclassified into π_1, and n_{22} are correctly classified into π_2 where $n2 = n21 + n22$. Thus

$$Apparent\ error\ rate = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

$$= \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

Similarly,a$pparent\ Correct\ Classification\ rate = \frac{n_{11}+n_{22}}{n_1+n_2}$

$$Apparent\ error\ rate = 1 - Apparent\ Correct\ Classification\ rate$$

3.0 The Analysis

3.1 Computations of Mean Vectors and Variance – Covariance Matrix.

$$F = -2b_1 \ln M \text{ is approximately } F_{a_1,a_2}$$

$$F = -2b_1 \ln M$$

$$= -2(0.16)(-5.195)$$

$$= 1.66$$

$$F(10,\infty) = 1.83$$

Decision: Based on the result above, the null hypothesis is not rejected since the calculated F-value (1.66) is less than the F-critical (1.83). Hence, the covariance matrices are equal

3.2 Testing for differences between groups using Wilk's Lambda

$$\Lambda = \frac{9.24 \times 10^{11}}{2.04 \times 10^{12}}$$

$$= 0.453$$

$$\Lambda_{\alpha,p,\upsilon_H,\upsilon_E} = \Lambda_{0.05,3,1,407} = 0.976$$

$$F_{p,\upsilon_E-p+1} = F_{3,405} = F_{3,\infty}^{(0.05)} = 2.6$$

Decision:

The calculated F-value is greater than that of the critical. Hence, there is a significant difference of the variables in the two groups. ($\Lambda = 0.453 < 0.976$).

3.3 Computation of discriminant function coefficient and discriminant function

The discriminant function coefficient is given as:

$$a' = (0.511 \quad 0.326 \quad 0.145)$$

Thus, the discriminant function can be written as;

$\hat{Y} = 0.511X_1 + 0.326X_2 + 0.145X_3$

Where;

X_1=Age, X_2=Body Mass Index (BMI), X_3=Waist Circumference

From the result above shows that age is the major discriminatory factor of hypertension followed by body mass index (BMI) and the waist circumference.

$$Hypertension = o.835Age + 0.238wc + 0.257BMI$$

The result age is the major discriminatory factor of hypertension followed by body mass index (BMI) and the waist circumference is also confirmed by the standardized coefficient.

3.5 Classifying hypertensive status using the discriminant model

For π_1 (hypertensive group),

$$\hat{l}_1 = a'\bar{x}_1 = (0.511 \quad 0.326 \quad 0.145)\begin{pmatrix}45.86\\29.98\\84.11\end{pmatrix}$$

$$= 0.511(45.86) + 0.326(2.98) + 0.145(84.11) = 36.60$$

Similarly, for π_2 (non-hypertensive group),

$$\hat{l}_2 = a'\bar{x}_2 = (0.511 \quad 0.326 \quad 0.145)\begin{pmatrix}22.33\\22.79\\77.19\end{pmatrix}$$

$$= 0.511(22.33) + 0.326(22.79) + 0.145(77.19) = 30.03$$

The classification rule is as follows:

Classify as group 1 (Hypertensive) if $\hat{Y} > \frac{1}{2}(\hat{l}_1 + \hat{l}_2) = 33.32$

Classify as group 2 (non-hypertensive) if $\hat{Y} < \frac{1}{2}(\hat{l}_1 + \hat{l}_2) = 33.32$

There are no new observations available; so the procedure is illustrated by classifying two of the observations each in both π_1 and π_2.For

$x_{11}' = (24 \quad 25.7 \quad 77)$, the first observation in π_1,

$$\hat{Y}_{11} = a'x_{11} = (0.511 \quad 0.326 \quad 0.145)\begin{pmatrix}24\\25.7\\77\end{pmatrix} = 31.81$$

Which would misclassifyx_{11}into π_2

For $x_{12}' = (48 \quad 29.37 \quad 88)$

$$\hat{Y}_{12} = a'x_{12} = (0.511 \quad 0.326 \quad 0.145)\begin{pmatrix}48\\29.37\\88\end{pmatrix} = 46.86$$

Which is greater than 85.98, and x_{12}would be correctly classified as belonging to π_2.

3.6. Estimation of Misclassification Rates

Apparent Correct Class rate = $\frac{n_{11}+n_{22}}{n_1+n_2}$

$$= \frac{34 + 359}{35 + 374} = 0.96$$

Apparent error rate =1 - 0.96= 0.04

Interpretation: the classification results reveal that 96% of individuals were classified correctly into 'hypertensive' or 'not hypertensive' groups.

4.0 conclusion

According to the results of this study, it is pertinent to conclude that there is a significant difference of the variables (age, BMI, waist circumference) in the two groups. The assumption of equality of covariance matrices was not violated and the main determinant of hypertension is age according to the present study. Also, the percentage (96%) of individuals being classified correctly into their respective groups is high which reveals that the discriminant model proves to be very good.

REFERENCES

Abdi, H. (2007). Discriminant correspondence analysis. In N.J.Salkind (ed) .*Encyclopedia of Measurement and Statistic*. Thousand Oaks (CA): Sage , pp. 270-275.

Alexacos, C. E. (1996). Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. *Journal of Educational Psychology* , 55, 297-306.

Alstad, K.S. & Smirk, F.H. (1948). Hypertension in New Zealand: A survey of 443 consecutive deaths in Dunedin hospital. *New Zealand Medical Journal* , 47, 298-308.

Bokeoglu, C.O. & Buyokozturk, S. (2008). Discriminant function analysis: Concept and application. *Egitim Arastirmalani diergisi* , (33), 73-92.

Dennison C. (2007). Determinants of Hypertension Care and Control Among Peri-Urban Black South Africans. *The Hihi Study, Ethnicity and Disease* , 17, 484-491.

Fisher, R. A. (1936). The use of multiple measurement in taxonomic problems. *Annals of Eugenes* , 7(2), 179-188.

Gal, M. D., Borg, W. R & Gall, J. P. (1996). *Educational research: An introduction.* (6th edition). New York: Longman Publishers.

Hardle, W. & Simar, L. (2007). Applied multivariate statistic analysis. *Springer Berlin Heidelbe* , 289-303.

Huberty, C. J. (1974). *Discriminant analysis*. Paper presented at the Annual Meeting of the American Educational Research Association (59th), Chicago, ILLINOIS, April 1974.

Mclachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. *Willey Interscience. ISBN 0-471-69115-1. MR1190419* ..

Moore, L. L., Visioni, J, & Qureshi, M. M. (2005). Weight loss in overweight adult and the long term risk of Hypertension. *Arch Intern Med* , 165 (11) 1298-1303.

Randall, M. D & Neil, K.E. (2009). *Disease management: A guide to clinical pharmacology (2nd edition).* London: Pharmacautical Press.

Report, Statistical Analysis (2007). *population and housing census results.* Ethiopia: Addis Ababa.

Stevens, J. (1996). *Applied multivariate statistics (3rd edition).* Mahwah, NJ: Lawrence Eribaum Associates, Inc.

World Health Organisation (2005). *Prevention of chronic diseases: A vital investment.* Geneva, Switzerland: WHO Global Report.