ON THE CLASSIFICATION AND ANALYSIS OF DATA FROM TUBERCULOSIS PATIENTS

S.O. OGUNLEYE [1,+] and A.B. FAGBOHUN[2]

1. Department of Industrial Mathematics, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria.

2. Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria.

(*Submitted*: 23 January 2006; *Accepted*: 18 October 2006)

Abstract

Linear discriminant analysis and logistic regression analysis models were applied to analyse data from ninety randomly selected discharged tuberculosis patients collected from the records Department of the Obafemi Awolowo University Teaching Hospitals Complex, Ile- Ife, Osun State. The comparison of the performance of these models in terms of classification into target groups (collections of patients known to have had complications and those with no history of complications of pulmonary Tuberculosis) were also investigated.

The result revealed the significance of some associated risk-factors (predictors) in the prediction of the complication of Pulmonary Tuberculosis. The variables identified to be statistically significant among the associated risk-factors were length of time of reporting to the right hospital, previous exposure to immuno-suppressive diseases especially HIV/AIDS, social history of the patients and nature of occupation of the patients before infection. The classification results showed that logistic regression did better than discriminant analysis in terms of proportion of correct classifications.

Finally, the study showed that logistic regression analysis is more effective in detecting related health problems of tuberculosis patients in which a discrete and dichotomous data are available.

Keywords: Discriminant analysis, Logistic regression, Pulmonary Tuberculosis, Risk factor, Immuno- suppressive disease.

1. Introduction

Linear discriminant analysis and logistic regression are widely used multivariate statistical techniques for analysis of data with categorical outcome variables. Both of them are appropriate for the development of linear classification models. Nevertheless, the two techniques differ in their basic idea. Logistic regression makes no assumptions on the distribution of the explanatory variables while linear discriminant analysis is developed for normally distributed explanatory variables.

Maja *et al.* (2004) and Stephen (1997) reported that logistic regression is more flexible, statistically more robust in practice, easier to use and understand than discriminant analysis in cases of violations of these assumptions.

Truett *et al.* (1967) emphasized that the assumption of multivariate normality is unlikely to be satisfied in applications or in practice.

Halperin *et al.* (1971) concluded that "use of the maximum likelihood method would be preferable, whenever practical, in situations where the normality assumptions are violated especially when many of the independent variables are qualitative".

Mitchell *et al.* (2004) also concluded that logistic regression is preferable under non-multivariate normal condition.

Press and Wilson (1978) carried out two empirical studies on non-normal classification problems, compared logistic regression and linear discriminant analysis and found logistic regression with maximum likelihood estimators performing better than linear discriminant analysis in both cases, thus supporting the results of Halperin et al (1971) and Stephen (1997).

Flury and Riedwyl (1990), Mitchell et al (2004) and Cox (1989) also described an empirical comparison of logistic regression and discriminant analysis and found that under non-multivariate normal condition, the logistic regression method is considered appropriate.

In this work, we have chosen to compare the performance of linear discriminant analysis and logistic regression analysis models to data collected form Tuberculosis patients.

+ *corresponding author* (email: kayodeogunleye2002@yahoo.com)

2. Materials and Methods

(*a*) Data Collection Procedure

Data from ninety randomly selected discharged Tuberculosis patients from 1998 to 2001 collected from the records department of the Obafemi Awolowo University Teaching Hospitals Complex, Ile- Ife, Osun State, Nigeria were used. The Hospital diagnostic index cards and the case notes of these discharged patients were thoroughly studied with particular attention being paid to some of the factors influencing the probability of having complications of Pulmonary Tuberculosis, which formed the main focus of our study.

In the study, Tuberculosis patients were dichotomized into two groups: the C^+ (π_1) group, those who had suffered from complications of pulmonary Tuberculosis after clinical diagnosis and the C^- (π_0) group, those who had not suffered from any complications of pulmonary Tuberculosis. The dependent variable Y is defined as

Y (outcome) = 1 if C^+ (π_1) and 0 if C^-(π_0)

The predictors (explanatory variables) available for this study are described as follows:

x_1 = age, is coded to take discrete values 1-5. For example, 15-24 years is coded 1, 25-34 years is coded 2, 35-44 years is coded 3, 45 –54 years is coded 4 and age e•55 years is coded 5.

x_2 = nature of occupation of the patient before infection, is coded to take discrete values 0-3. For example, no job is coded 0, student is coded 1, unskilled workers (traders, cement or tobacco factory or quarry workers e.t.c.) are coded 2 and the skilled workers (workers in the chest hospital or Tuberculosis wards e.t.c) are coded 3.

x_3 = previous contact with a person having a chronic cough or an infected person. This takes binary values 0-1. For instance, no sign of contact is coded 0 and with sign of contact 1.

x_4 = social history of the patient (tobacco smoking and alcohol consumption). This also takes discrete values 0-3. For instance, the patient who had not smoked or drank before is coded 0, the patient who had not drank at all but had smoked before is coded 1, those who had drank before but had not smoked at all are coded 2 and those who had drank and smoked before are coded 3.

x_5 = previous exposure to immuno-suppressive disease is coded to take discrete values 0-4. The patient with no disease is coded 0, presence of HIV/AIDS as the main disease is coded 1, presence of at least one from diabetes, leprosy, cancer, malnutrition, measles or any disease that can depress immunity apart from HIV/AIDS is coded 2, presence of at least one from hypertension, pneumonia with or without malaria fever is coded 3 and presence of only malaria fever is coded 4.

x_6 = previous exposure to tuberculosis infection takes discrete values 0-2. The patient with no previous tuberculosis infection is coded 0, the patient who had been infected before but failed to complete his or her treatment is coded 1 and those who had been infected before but completed their treatment are coded 2.

x_7 = housing condition of the patient before infection is coded to take values 0-2. Homelessness is coded 0, crowded living condition is coded 1 and uncrowded living condition is coded 2.

x_8 = length of time of reporting to the right hospital after noticing persistent cough or other discomfort is coded to take values 1-5. The patient who had reported after 1-3 weeks of persistent cough or other discomfort is coded 1, the patient who had reported after 1-5 months of persistent cough is coded 2, those that had reported after 6-10 months of persistent cough are coded 3, those that had reported after 11-15 months of persistent cough are coded 4 and the patient who had reported after 16-20 months of persistent cough or other discomfort is coded 5.

(*b*) Statistical method of analyses

The following models were examined:

(i) The logistic regression model is usually formulated mathematically by relating the probability of some event occurring of the jth object i.e. P_j conditional on a vector X_j of explanatory variables, through the logistic distribution functional form. Thus, the probability of some event j,

$$p_j = p_r\left(Y_j = 1 \middle| \underline{X}_j\right) = \frac{e^{\beta_0 + \underline{X}_j \beta}}{1 + e^{\beta_0 + \underline{X}_j \beta}}$$

where

$\underline{X}_j = (x_{j1}, x_{j2}, \ldots, x_{jk})$ is the 1 x k vector of explanatory variables

and

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix}_{k\times 1}$$

are unknown regression parameters that are estimated from the data by the method of maximum likelihood estimation.

The linear logistic regression model, which was employed for this purpose, is given by

$$\log_e\left(\frac{P_j}{1-P_j}\right) = \beta_0 + \underline{X}_j\beta = \beta_0 + \sum_{i=1}^{k}\beta_i x_{ji}$$

(ii) The linear discriminant function model is formulated by assuming that the two populations (π_1, π_0) are multivariate normal with equal covariance matrices Σ and that the costs of misclassification are equal. If μ_1 and μ_0 denote the mean vectors of the two populations, a likelihood ratio test readily yields the classification procedure to classify the object into first population or group if

$$(\mu_1 - \mu_0)'\sum^{-1}\left(x - \frac{1}{2}(\mu_1 + \mu_0)\right) \geq \log\left(\frac{q_0}{q_1}\right)$$ Press and Wilson (1978)

q_0, and q_1 are prior probabilities of belonging to group 0 and group 1. Anderson (1958), suggested that the parameters q_0, q_1, μ_1, μ_0 and Σ be replaced by their sample estimates because they will be unknown in practice. The classification rule now depends on

$$(\bar{x}_1 - \bar{x}_0)'S^{-1}\left(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_0)\right)$$ which is usually called Anderson's Classification Statistic where

$$\hat{q}_0 = \frac{n_0}{n},\ \hat{q}_1 = \frac{n_1}{n},$$

$n_1 = \sum_{i=1} y_i$, $n_0 = n - n_1$, the respective numbers of observations from π_1 and π_0

$$\hat{\mu}_1 = \bar{x}_1 = \frac{\sum_{y_i=1} x_i}{n_1} \quad \hat{\mu}_0 = \bar{x}_0 = \frac{\sum_{y_i=0} x_i}{n_0}$$

$$\hat{\sum} = \left(\frac{\sum_{y_i=1}(x_i - \bar{x}_1)(x_i - \bar{x}_1)' + \sum_{y_i=0}(x_i - \bar{x}_0)(x_i - \bar{x}_0)'}{n}\right)$$

3. Results and Discussions

Table 1 shows the sresult of correlation analysis of the dependent variable and the independent variables with one another. This is done in order to bring out the best variables that could be used in classification. At 5% level of significance, some of the explanatory variables were significantly correlated with the response variable Y while some were non- significantly correlated with it. After such variables with no significant correlation with response variable Y are extracted, four variables are left in our study. These are x_2, x_4, x_5 and x_8.

The associations of these variables (x_2, x_4, x_5 and x_8) with complications of Pulmonary Tuberculosis provide some new areas of interest to the medical researcher. These associations may be studied by inspection of the estimated functions. These are for logistic regression:

$$\hat{\lambda}_j = 0.448 + 0.229\, x_{2j} + 1.172\, x_{4j} - 2.685\, x_{5j} + 2.414\, x_{8j}$$

and for discriminant analysis:

$$\hat{Z}_j = -0.908 + 0.213\, x_{2j} + 0.425\, x_{4j} - 1.061\, x_{5j} + 1.174\, x_{8j}$$

As might be expected, the two functions are quite similar. The complications of Pulmonary Tuberculosis were positively associated with x_2, x_4 and x_8 but negatively associated with x_5. We observe that absence of complications of Pulmonary Tuberculosis was influenced mainly by the presence of malaria fever than presence of complications of Pulmonary Tuberculosis. Presence of immuno- suppressive diseases most especially HIV/AIDS associated most strongly to the occurrence of complications of Pulmonary Tuberculosis. The presence of complications of Pulmonary Tuberculosis was also strongly associated with variables x_2, x_4 and x_8.

The ninety Tuberculosis patients were then classified from the estimated functions above (the functions estimated from the ninety Tuberculosis patients) and were later cross-validated using leaving-one-out method. The cross-validation was done-in-order to obtain a more realistic and reliable estimate of the misclassification rate. The classification rule used for classification is the allocation of patients with scores x_2, x_4, x_5 and x_8 to group C^+ for logistic regression if

$$\hat{\lambda}_j \geq P_0$$

and to the group C^- if

$$\hat{\lambda}_j < P_0$$

where P_o is the prior probability of group C^+ of m patients out of n patients and for discriminant analysis, the allocation is to group C^+ if

$$\hat{Z}_j \geq \ln \frac{q_0}{q_1}$$

and to the group C^- if

$$\hat{Z}_j < \ln \frac{q_0}{q_1}$$

where q_0 and q_1 are prior classification probabilities for the second and first population (C^- and C^+) respectively. In this work, 50 patients were in C^+ group while 40 were in C^- group. This result implies that the higher the score for any given observation, the higher the probability that the patient will be classified as having complication of Pulmonary Tuberculosis or vice-versa.

Table 2 shows the summary of classifications of the ninety Tuberculosis patients by logistic regression and discriminant function methods. The logistic regression classified 82 (37 + 45) of the 90 patients observed correctly, for a 91.1 percent correct classification rate. The discriminant analysis correctly classified 80 (32+ 48) of the 90 patients observed for a 88.9 percent correct classification rate. The prior probabilities used were 0.44 (40/90) of having no complications and 0.56 (50/90) of having complications, the approximate proportions of the actual cases in our data.

Table 3 shows the summary of classifications of the ninety Tuberculosis patients cross-validated using leaving-one-out method. The logistic regression classified 82 (37 + 45) of the 90 patients cross-validated correctly for a 91.1 percent correct classification rate while the discriminant analysis correctly classified 79 (31+ 48) of the 90 patients cross-validated for a 87.8 percent correct classification rate. Of particular interest in this work is the pattern of errors. There were some overlap when we looked at the cases that were misclassified by each function. Five cases (patients) were misclassified the same by both functions. The two functions misclassified two cases (observation 38 and 41) in π_1 (first group) as being in π_0 (second group) while the rest three cases (observation 64, 66 and 75) in π_0 were also misclassified as being in π_1 by the two functions as well. In addition, logistic regression misclassified three cases (observation 29, 36 and 37) in π_1 as being in π_0 that discriminant analysis classified properly. Discriminant analysis also misclassified 6 cases (observation 56, 62, 65, 70 77 and 79) in π_0 as being in π_1 that logistic regression classified properly. Thus, there is a clear difference in the types of cases misclassified by the two functions. The discriminant function misclassifies more patients (cases) into the group (π_1) than the logistic regression function.

4. Conclusion

It was observed that presence of immuno suppressive diseases most especially HIV/AIDS with a longer duration of persistent cough before reporting to the hospital contributed mostly to the occurrence of complications of Pulmonary Tuberculosis. We also observed that patients that smoke tobacco with poor socio-economic status are also prone to complications of Pulmonary Tuberculosis. Therefore, the length of time of reporting to the right hospital after noticing persistent cough, previous exposure to immuno-suppressive disease especially HIV/AIDS, social history and nature of occupation of the patients before the infection were identified to be statistically significant in the prediction of the complication of Pulmonary Tuberculosis. It was also observed that logistic regression performed better than discriminant analysis in terms of the proportion of correct classifications. This agreed with the conclusions of Stephen (1997) and Press and Wilson (1978). Thus, logistic regression model could be preferred to linear discriminant function model for analysing Tuberculosis data, which are non-normal in hospital data management.

REFERENCES

Anderson, T.W., 1958. An introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York, London. Chapter 6, pp. 126-152.

Cox, D.R. and Snell, E.J., 1989. Analysis of Binary Data 2nd edition. Chapman and Hall, London.

Flury, B. and Riedwyl, H., 1990. Multivariate Statistics (A Practical Approach). Chapman and Hall, London.

Halperin, M.B., William, C. and Verter, J., 1971. Estimation of the Multivariate Logistic Risk Function: A comparison of the Discriminant Function and Maximum Likelihood Approaches. *Journal of Chronic Diseases*, 24, 125-158.

Maja, P., Mateja, B. and Sandra, T., 2004. Comparison of Logistic Regression and Linear Discriminant Analysis: A simulation study. *Metodolosiki zvezki*, 1(1), 143-161.

Mitchell, P.B., Slade, T. and Andrew, G., 2004. Twelve-month prevalence and Disability of DSM-IV Bipolar Disorder in an Australian General Population Survey. *Psychological Medicine*, 34, 777-785.

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

Stephen, L., 1997. Note on Logistic Regression and Discriminant Analysis. University of Exeter, Dept. of Psychology, Exeter, U.K. www2.chass.NCSV.Edu/Garson/Pa765/Logistic.htm. pp. 1-9.

Truett, J., Cornfield, J. and Kannel, W., 1967. A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. *Journal of Chronic Disease*, 20, 511-524.