*International Journal of Statistical Sciences* *ISSN 1683–5603*
*Vol. 5, 2006, pp 73-84*


# A comparison between Logistic Regression and Linear Discriminant Analysis for the Prediction of Categorical Health Outcomes

**Demosthenes B. Panagiotakos**
*Unit of Biostatistics & Epidemiology, Department of Nutrition - Dietetics*
Harokopio University, 70 E. Venizelou Str., 176 71, Athens, Greece
Email: dbpanag@hua.gr

[Received Janaury 27, 2006; Revised December 26, 2006; Accepted December 31, 2006]

**Abstract**

Both logistic regression and linear discriminant analysis can be used to predict the probability of a specified categorical outcome using several explanatory variables. The objective of this work is to investigate whether these two methods of analysis result similar findings in evaluating categorical health outcomes. For this purpose we use a specific health problem, i.e. modeling several characteristics of patients admitting with Acute Coronary Syndrome (ACS) on in-hospital mortality. In conclusion, logistic regression resulted in the same model as did discriminant analysis.

**Keywords and Phrases:** Categorical; Logistic; Discriminant

**AMS Classification:** 43A95; 90B06; 11R29

## 1 Introduction

Both logistic regression and linear discriminant analysis can be used to predict the probability of a specified categorical outcome using several explanatory variables. Particularly, logistic regression allows predicting an outcome, which may be continuous, discrete, dichotomous, or a mix. Logistic regression is very popular in the health sciences, since the discrete outcome could often be the presence or absence of a disease. Unlike regression analysis, logistic regression does not run into the problem of predicting negative probabilities for group membership. Moreover, logistic regression is especially useful when the relationship between the probability of group membership (e.g., probability of disease) depends nonlinearly on predictors. Logistic regression

analysis is based on the calculation of odds, which the ratio of the probability of the dependent outcome being into one group divided by the probability of being into the other group. Similarly, discriminant analysis aims to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no necessarily natural ordering on the groups. Discriminant analysis is based on the estimation of orthogonal discriminant functions that are linear combinations of the standardised independent variables, which yield the biggest mean differences between the groups. Thus, it could be suggested that discriminant analysis and logistic regression can be used to address the same types of research question.

The objective of this work is to investigate whether these two methods of analysis result similar findings in evaluating categorical health outcomes. For this purpose we use a specific health problem, i.e. modeling several characteristics of patients admitting with Acute Coronary Syndrome (ACS) on in-hospital mortality.

# 2 Methodology

## 2.1 Linear Discriminant Analysis and Logistic Regression

Discriminant analysis captures the relationship between multiple independent variables and a categorical dependent variable in the usual multivariate way, by forming a composite of the independent variables. This type of multivariate analysis can be used to determine which variable discriminates between two or more groups of subjects and to derive a classification model for predicting the group membership of new observations (Tabachnink BG, 1996). In the simplest type of discriminant analysis, i.e. the two group, a linear discriminant function that passes through the means of the two groups (centroids) can be used to discriminate subjects between the two groups. For each case, the coefficient for an independent variable is multiplied by the case's score on that variable; these products are summed and added to the constant and the result is a composite score for that case, i.e. their discriminant score. In general, the linear discriminant function (LDF) is represented by equation:

$$LDF = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} = bX$$

where $b_j$ : the value of the $j^{th}$ coefficient, j = 1, ..., k, and $x_{ij}$ : the value of the $i^{th}$case of the $j^{th}$ predictor. The LDF can also be written in standardized form, in which each variable is adjusted by subtraction of its mean value and division by its standard deviation. The standardized coefficients allow comparing variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability. From the LDF scores can be derived predicted probabilities and predicted group membership on the dependent variable. The basic rationale of this approach is that the bigger the between-groups sum of squares is relative to the within-group sum of squares, the more likely it is that the independent

and dependent variables are related. Similarly, this relationship can be indexed with the ratio of between-group divided by total sum of squares (eta-squared statistic or explained variability), or of within-group divided by total sum of squares (Wilks's lambda statistic or unexplained variability). Further, the ratio of between-group divided by within-group sum of squares can be changed into a ratio of variances that then becomes the F statistic. The F statistic reassures that the relationship is unlikely to be due to chance. In discriminant analysis we carry forward from this bivariate analysis two principles: a) the first focuses on the distance between the two group means (centroids) that is $|\overline{y}_1 - \overline{y}_2|$, and b) the second focuses on the pooled estimate of the variance $s_y^2 = \frac{\sum (y_{1j}-\overline{y}_1)^2 + \sum (y_{2j}-\overline{y}_2)^2}{n_1+n_2-2}$.

The principle by which the discriminant coefficients (or weights) are selected is that they are chosen in order the distance between the centroids is maximized. So coefficients are chosen that push the group means on the composite variable as far apart as possible, that is, that maximally discriminates between the two groups. Fisher (1938), suggested to transform the multivariate observation x to univariate observations y such that the y's derived from groups 1 and 2 were separated as much as possible. Thus, the linear combination $y = a'x$ is the one that maximizes the ratio: (squared distance between sample means)/(sample variance y). The vector of coefficients is given by the eigenvectors of the matrix: B*S$^{-1}$, where B = $(\overline{x}_1 - \overline{x}_2)'$is the between group matrix and S is an estimate of $\Sigma$. A critical feature of these composite sums of squares is that they encapsulate, not only the variability of each variable, but also their co-variability. Further, the coefficients can again be calculated in unstandardized or standardized form. However, the discriminant coefficients are less informative than those in regression, whatever their form. If we assume that there are 2 groups and $\overline{x}_1$, $\overline{x}_2$ are the means of each group, and S the pooled covariance matrix, the allocation rule based on Fisher's discriminant functions is the following:

$$\begin{array}{l} X_i \in \ group\,1,\ if\ y = (\overline{x}_1 - \overline{x}_2)'S^{-1}X_i \geq \frac{1}{2}(\overline{x}_1 - \overline{x}_2)'S^{-1}(\overline{x}_1 + \overline{x}_2) \\ X_i \in \ group\,2,\ if\ y = (\overline{x}_1 - \overline{x}_2)'S^{-1}X_i < \frac{1}{2}(\overline{x}_1 - \overline{x}_2)'S^{-1}(\overline{x}_1 + \overline{x}_2) \end{array} .$$

The logistic regression strategy retains the goal of generating predicted probabilities, but achieves it indirectly by using another probability index and a different criterion to choose the coefficients. Logistic regression is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome, based on values of a set of predictor variables (Hosmer DW, 1989). Since the probability of an event must lie between 0 and 1 (for the binary case), it is impractical to model probabilities with linear regression techniques, because the linear regression model allows the dependent variable to take values greater than 1 or less than 0. Defined as $p_1$the probability of an object is belonging to group 1 and as $p_0$the probability of an object belonging to group 0. The form of the logistic regression model is:

$$z_i = \log\left(p_{i1}/p_{i0}\right) = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}$$

where $p_{i1}/p_{i0}$ : is called the odds ratio, $b_j$ : the value of the $j^{th}$ coefficient, j = 1, ..., k, $x_{ij}$ : the value of the $i^{th}$case of the $j^{th}$ predictor. The parameters of the logistic model ($b_0$ to $b_k$) are derived by the method of maximum likelihood. From the logistic regression model we can derive the probability of an event occurring as:

$$P(Y_i = 1 \backslash X_i) = \frac{e^{b^T X_i}}{1 + \left(e^{b^T X_i}\right)} = \frac{1}{1 + e^{-b^T X_i}}$$

Using a probability cut-off of 0.5, you can classify an object to group 1 if p1 > 0.5 and to group 0 if $p_1 < 0.5$. The previous probability is of course the same to the one used in discriminant analysis presented above. The predicted probabilities and actual categories for each case are bundled up into a statistic called the log-likelihood function. So, the aim is to find the coefficients that maximize the value of the log-likelihood function. There is another method of expressing the strength of the multivariate relationship, which is not only less contentious, but also more intuitively appealing and potentially more practicable. 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. Cross-classifying cases according to their assigned and actual categories provide another picture of how well the independent variables predict the dependent variable. Since the predicted probabilities are decimal values between 0 and 1, they need to be dichotomized so that they can be compared with the actual 0 and 1 categories in a 2 × 2 table.

Hence, the two methods do not differ in functional form; they only differ in the methods used for the estimation of coefficients as well as for the rules used for decision making. Moreover, there are basic differences in the statistical assumptions, which underlie those two methods.

With discriminant analysis, the assumptions are: a) the data for the independent variables represent a sample from a multivariate normal distribution; therefore, predictor variables should be interval or ratio variables, b) The variance of the predictors and the correlations among the predictors within each group should be the same (equal variance/covariance matrices), and c) The predictors are not highly correlated with each other. Moreover, it is empirically recommended that the sample size in a discriminant analysis should provide at least 20 cases for each independent variable. The dependent variable in a discriminant analysis should be categorical and may have any number of categories. The categories should be mutually exclusive and jointly comprehensive, allowing each case to be assigned to a single category. It is assumed that all of the independent variables are measured on at least an interval scale. However, the more non-interval variables that are included, the less trustworthy the results will be in terms of finding the optimum separation of the groups.

With logistic regression the assumptions are that a logistic regression (i.e. a sig-

moidal dependency) exists between the probabilities of group memberships and a linear function of the predictor variables. It is also assumed that observations are independent. Moreover, in the logistic regression context, the more unequal the numbers in the categories, the more cases are needed. Add to all this the problem of missing data because of list-wise deletion, and the desirability of having enough cases to cross-validate results on a holdout sample, and it becomes painfully clear that logistic regression typically requires cases in the hundreds to guarantee trustworthy results. Additionally, there is no formal requirement for multivariate normality, homoscedasticity, or linearity of the independent variables within each category of the dependent variable. However, someone may note that satisfying these conditions among the independent variables for the whole sample may enhance robustness of the results. Thus, the problem of multi-co linearity could apply to logistic regression and the assumption of independence of cases remains in place, since case-wise exploration of the residuals may reveal patterns suggesting non-independence and may identify outliers for whom the model provides notably poor predictions.

Furthermore, in order to evaluate both approaches sensitivity, specificity and accuracy will be also calculated. The sensitivity of a binary classification test with respect to some class is the probability that the test correctly classifies cases of that class. That is, it is the proportion of true positives of all positive cases in the population. In addition, the specificity of a with respect to a given class is the probability that the test correctly classifies cases not belonging to that class. That is, it is the proportion of true negatives of all negative cases in the population. Finally, accuracy is the degree of veracity. In other words the degree of conformity of a measured or calculated quantity to its actual, true value. It is calculated as the ratio of true positive and true negative cases among all potential results.

## 2.2 Logistic regression or linear discriminant analysis?

In conclusion, linear discriminant analysis and logistic regression can be used to address the same types of research question. Similarly to logistic regression, in discriminant analysis the variable generated by the composite cannot be a predicted score on the dependent variable. Instead it is a LDF score that then feeds into calculations that produce the predicted probability of a case being in a particular category of the dependent variable. This predicted probability is then used to generate a predicted category for each case. Thus, the strategy is very similar to logistic regression in which the composite variable generates logits, which produce predicted probabilities, which produce predicted categories. It could be suggested that discriminant analysis is a more appropriate method when explanatory variables are normally distributed. In the case of categorized variables, discriminant analysis remains preferable and fails only when the number of categories is really small (2 or 3). The results of logistic regression, however, are in all these cases constantly close and a little worse than those of discriminant analysis. But whenever the assumptions of discriminant analysis are not met, the use

of discriminant analysis is not justified, while logistic regression gives good results since it can handle both categorical and continues variables, and the predictors do not have to be normally distributed, linearly related or of equal variance within each group regardless of the distribution as suggested by Pohar M., et al. (2004).

# 3 Application

## 3.1 Use of epidemiologic data to predict a health outcome

In this study, we compared the results of discriminant and logistic regression in predicting in-hospital mortality among patients presenting with a range spectrum of acute coronary syndromes. Between October 1, 2003 and September 30, 2004 (12 months) we enrolled almost all consecutive patients (participation rate = 98%) that entered in the cardiology clinics or the emergency units of six major General Hospitals, in Greece. During the study period 2,172 patients were admitted for ACS in the selected hospitals, 1649 (76%) of them were men and 523 (24%) were women. Further details about the data used may be found elsewhere (Pitsavos et al., 2005). The independent variables which were available as potential predictors for in-hospital mortality was history of coronary heart disease (yes or no), hypertension (yes or no) and diabetes mellitus (yes or no), sex (male or female), age in years, body mass index in $kg/m^2$, smoking habits (yes or no), initial level of systolic blood pressure in mm Hg, the estimated creatinine clearance rate in ml/min and the maximum level of MB isoenzyme of creatinine kinase (CKMB) in ng/ml.

Initially, we entered in both discriminant and logistic regression models only the predictors, which were statistically significant in univariate analysis. We used the standardized canonical discriminant function coefficients for discriminant analysis and z statistic (standardized coefficients, Wald statistic) for logistic regression, to evaluate the contribution of each one variable to the discrimination between two groups. The larger the standardized coefficients, the greater are the contribution of the respective variable to the discrimination. We, also, compared the sign and magnitude of coefficients. Secondarily, we performed stepwise discriminant and logistic regression analysis including all available predictors mentioned above. For discriminant analysis, the selection criterion for entry was the Wilks' Lambda, with a value F-to-enter of 3.84 and a value F-to-remove of 2.71. For logistic regression, was used the set of 0.05 significance levels for entry and 0.1 for removal of variables; these p values were selected to approximate the F values used in the discriminant analysis. We compared the variables selected, the order of selection and the sign and magnitude of coefficients. Equality of the covariance matrices was checked with the Box's M test and it was revealed that they were not equal ($p < 0.001$), thus this assumption for discriminant analysis was not met.

Response operating characteristics (ROC) curves were plotted for each model. An ROC curve graphically displays sensitivity and 100% minus specificity (false positive rate) at several cut-off points. By plotting the ROC curves for two models on the same axes, one is able to determine which test is better for classification, namely, that test whose curve encloses the larger area beneath it.

To approach further statistical generalization a non-parametric bootstrap technique was applied. In particular, we created 1000 random samples from the same dataset and we run both logistic regression and discriminant analysis models. In each sample the correct classification rate was calculated for both statistical methods. All analyses were performed using the STATA version 8.0 software (STATA Corp., College Station, TX, USA).

## 3.2 Results

Univariate analysis revealed that the CPKMB levels, the systolic blood pressure, the creatinine clearance, gender, age, and diabetes, contribute significantly in the discrimination of patients in those dying during their hospitalization and those surviving. Using in discriminant and logistic regression only these variables, both techniques revealed that CPKMB levels, systolic and diabetes were the most important contributors (Table 1).

Table 1. Variables, standardized and un-standardized coefficients for the discriminant analysis model and logistic regression models.

| Predictors | Logistic Regression | | Discriminant analysis | |
|---|---|---|---|---|
| | b coefficients | z- statistic | Unstandardized coefficients | Standardized coefficients |
| CKMB in ng/ml | 0.005 | 4.86 | 0.007 | 0.649 |
| Systolic blood pressure in mmHg | -0.021 | 3.49 | -0.015 | -0.390 |
| Diabetes (yes vs. no) | 1.076 | 3.22 | 0.812 | 0.375 |
| Creatinine clearance in ml/min | -0.020 | 2.55 | -0.04 | -0.196 |
| Age in years | 0.04 | 2.14 | 0.029 | 0.372 |
| Male vs. female | -0.63 | 1.87 | -0.625 | -0.264 |

Moreover, we observe that the direction of the relationships was the same and there were not extreme differences in the magnitude of the coefficients. The overall correct classification rate was 79% for discriminant analysis and 96.6% for logistic regression. However, when we used not equal prior probabilities for the two groups the overall

correct classification rate for discriminant analysis was increased in 96.3%. Table 2, presents sensitivity, specificity, and accuracy (i.e., correct classification rates) of both approaches at various cut-offs of the probability of dying within hospital. Although some differences were observed between the two approaches regarding the classification ability, Figure 1 that illustrates the ROC curves of the aforementioned models, clearly indicates that the logistic model is similar to the discriminant analysis model (i.e., no difference in the area under the curve, AUC, 81.8% vs. 81.1%, p = 0.9).

Table 2. Sensitivity, specificity and accuracy of logistic regression and discriminant analysis models, at various cut-off points for the probability of having the disease.

| Cut-off value* | Logistic regression | | | Discriminant analysis | | |
|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | Accuracy | Sensitivity | Specificity | Accuracy |
| 0.05 | 61 | 85 | 84.3 | 100 | 5 | 8.1 |
| 0.10 | 37 | 95 | 93.3 | 100 | 17 | 19.5 |
| 0.25 | 13 | 99 | 96.4 | 93 | 48 | 49.5 |
| 0.50 | 2 | 99 | 96.8 | 65 | 78 | 77.2 |
| 0.75 | 0 | 100 | 96.8 | 46 | 93 | 91.7 |
| 0.90 | 0 | 100 | 96.8 | 11 | 99 | 95.8 |

* P (death); values less than or equal to the cut-off value indicate that the person is alive; those greater than the cut-off value indicate that the person would die during hospitalization.

Furthermore, the stepwise approach revealed that both models selected the same variables, with the same order of entry (Table 3). Furthermore, the sign of the coefficients were the same and a slight difference was observed in the magnitude of the coefficients. The correct classification rate was 81.4% for discriminant analysis and 96.8% for logistic regression. Figure 2, shows that the logistic model is slightly superior in its classification ability compared to discriminant analysis model.

Table 3. Variables, standardized and un-standardized coefficients for the discriminant analysis model and logistic regression models, after stepwise approach in the original dataset.

| Predictors | Logistic Regression | | Discriminant analysis | |
|---|---|---|---|---|
| | b coefficients | z- statistic | Unstandardized coefficients | Standardized coefficients |
| CKMB in ng/ml | 0.005 | 5.22 | 0.007 | 0.692 |
| Age in years | 0.071 | 3.96 | 0.036 | 0.457 |
| Systolic blood pressure in mmHg | -0.027 | 3.63 | -0.017 | -0.411 |
| Males vs. Females (1 vs. 0) | -0.938 | 2.48 | -0.805 | -0.340 |
| Diabetes (1: yes vs. 0: no) | 0.901 | 2.42 | 0.594 | 0.274 |

### 3.3 Validation of the findings

In order to validate our findings we performed a bootstrap simulation technique. In particular, we re-sampled our original dataset 1000 times and we run both logistic regression and discriminant analysis. The results (i.e., bootstrap estimates, standard errors, and bias) are presented in Table 4.

Table 4. Variables, estimates (standard errors), and bias, in logistic regression and discriminant analysis, after non-parametric bootstrap re-sampling method.

| Predictors | Logistic Regression | | Discriminant analysis | |
|---|---|---|---|---|
| | b-coefficients | bias | Un-standardized coefficients | bias |
| CKMB in ng/ml | 0.003 (0.001) | -0.0001 | 0.004 (0.002) | -0.00001 |
| Age in years | 0.056 (0.002) | 0.002 | 0.029 (0.001) | 0.00005 |
| Systolic blood pressure in mmHg | -0.02 (0.006) | -0.001 | -0.012 (0.004) | -0.0001 |
| Males vs. Females (1 vs. 0) | -0.78 (0.36) | -0.018 | -0.78 (0.09) | -0.0002 |
| Diabetes (1: yes vs. 0: no) | 0.97 (0.31) | -0.012 | 0.77 (0.06) | -0.0001 |

Furthermore, the 95% CI of the correct classification rate using the information from the bootstrap re-sampling method was from 79% to 84% for discriminant analysis and from 94% to 98% for logistic regression.

## 4 Discussion

In general, results from the logistic model agreed with those of discriminant analysis. Both techniques selected the same variables when we performed the stepwise approach, while entering all significant variables from the univariate analysis in these two methods, only slight differences was observed in the order of predictors (from the most important for the discrimination between the two groups to the less important) between those methods. The overall correct classification rate was good for both, and either would be useful for the prediction of the in-hospital mortality of patients presenting with acute coronary syndromes. Moreover, although the assumption of equal covariance was not hold in this dataset, both methods had similar results. All of the sequential strategies, both hierarchical and statistical, can be used in discriminant analysis, though the statistical approach using such techniques as stepwise analysis is the most common application. When discriminant analysis is applied to more than two groups, the major consequence is that more than one discriminant function can be calculated. Each function will have its own set of coefficients and each will generate

a discriminant score for every case. Mathematically, it is possible to derive as many functions as there are groups minus 1. So for a four-group analysis, there will be a maximum of three functions, and each case will potentially have three discriminant scores. However, the fact that three functions can be derived does not mean that all are necessary in order to achieve maximum discrimination between the groups. This may be achievable with only one, or perhaps two, of the available functions. Not surprisingly then, the major new issue that arises when the dependent variable has more than two categories is how many functions are worth retaining from those that are available.

Moreover, Brenn T.and Arnesen E (1985), used discriminant analysis, logistic regression and Cox's models to select risk factors for total and coronary deaths among 6595 men aged 20-49 followed for 9 years. Groups with mortality between 5 and 93 per 1000 were considered. Discriminant analysis selected variable sets only marginally different from the logistic and Cox methods, which always selected the same sets. In addition the researchers observed that a time-saving option, offered for both the logistic and Cox selection, showed no advantage compared with discriminant analysis, since analysing more than 3800 subjects, the logistic and Cox methods consumed, respectively, 80 and 10 times more computer time than discriminant analysis. Thus the researchers concluded that discriminant analysis is advocated for preliminary or stepwise analysis, otherwise Cox's method should be used.

The presented work has some limitations. First, a violation of the assumption of equal covariance matrices as showed by the Box - M test, may modest our findings from the discriminant analysis, however, the multivariate Box M test for homogeneity of covariances is particularly sensitive to deviations from multivariate normality (which in our case is true for CKMB levels). However, this is not so important in this case since the violation from homoscedacity is due to some outlier observation in our data set. Moreover, although the presented results are not based only on one dataset, but have also been simulated using appropriate re-sampling methods, the confirmation of our findings in other original datasets is considered essential. In conclusion, logistic regression resulted in the same model and with a better correct classification rate, as did discriminant analysis. Taking into account the assumptions made before, one may decide which method should use to analyze the data.

## References

[1] Brenn T, Arnesen E. Selecting risk factors: a comparison of discriminant analysis, logistic regression and Cox's regression model using data from the Tromso Heart Study. Stat Med. 1985;4:413-23.

[2] Fisher, RA. The statistical utilization of multiple measurements. Annals of Eugenics 1938;8:376-386.

[3] Godfrey K. Statistics in practice. Comparing the means of several groups. N Engl J Med. 1985;313:1450-1456.

[4] Halperin M, Blackwelder WC, Verter JI. Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. J Chronic Dis. 1971;24:125-158.

[5] Hosmer, DW, Lemeshow, S. (1989): Applied Logistic Regression. New York: Wiley

[6] Montgomery ME, White ME, Martin SW. A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows. Can J Vet Res 1987;51:495-8.

[7] Pitsavos C, Panagiotakos DB, Antonoulas A, Zombolos S, Kogias Y, Mantas Y, Stravopodis P, Kourlaba G, Stefanadis C; Greek study of acute Coronary Syndromes study investigators. Epidemiology of acute coronary syndromes in a Mediterranean country; aims, design and baseline characteristics of the Greek study of acute coronary syndromes (GREECS). BMC Public Health. 2005;5:23.

[8] Pohar M, Blas M, Turk S. (2004): Comparison of logistic regression and linear discriminant analysis: A simulation study. Metodoloski Zvezki. 1;143-161

[9] Shott S. Logistic regression and discriminant analysis. J Am Vet Med Assoc 1991;198:1902-5.

[10] Tabachnink B.G, Fidell L.S (1996): Using Multivariate Statistics, Harper Collins, New York.