Application of Discrimination and Classification on Diabetes Mellitus Data

Michael Asamoah-Boaheng
School of Graduate Studies, Research and Innovation
Box KS 854, Kumasi Polytechnic
Kumasi, Ghana
Email: asboaheng@yahoo.com

Abstract

The assignment/allocation of individuals/observations to the various known groups with known mean vectors and distinguishing characteristics has been a major concern for years and several attempts have been made at deriving parsimonious rules that address this hurdle. In this study, Fishers Linear Discriminant Function (FLDF) was derived to provide maximum separation between Type 2 and Type1diabetes patients based on identified risk factors. The assumptions of FLDF were achieved by BoxMtest of equality of covariance matrices. A seven variate data on 620 diabetes patients obtained from Komfo Anokye Teaching Hospital (KATH) diabetes ward was obtained and used for data analyses. The derived FLDF was used to reclassify the original observation to obtain the discriminant scores from the functions and 85.3 percent correct classification was achieved. Also 84.8 percent of the cross validated grouped cases were correctly classified into either being a Type 2 or Type 1 diabetes patient group. Patients age as well as their BMI were identified to be the two major contributing variables in classifying a patient as a type1 or type 2 diabetes.

Keywords: Fisher Linear Discriminant Function, Diabetes Patients, Covariance Matrices, Cross Validation.

1.0 Introduction

Discriminant analysis is a multivariate approach for identifying the features that separate known groups or populations. In other words, discrimination is a multivariate technique concerned with separating distinct sets of observations and it is exploratory in nature. [4]. Discriminant analysis as a topic in Multivariate Statistical Analysis has attracted much research interest over the years, with the evaluation of Discriminant Functions when the covariances matrices are equal and unequal. This study is therefore aimed at using the classical method of discrimination in classifying diabetic patients as either type 1 or type 2 based on some identified anthropometric features and other factors.The problem of discrimination was first initiated by [2] in which equal covariance matrices were assumed with or without normality assumption. Fisher's approach to classification with two populations was based on arriving at a linear classification function that gave maximum separation between groups without assuming normality. Several investigations mainly with respect to multidimensional normal populations with common and unequal covariance matrices have been carried out by other authors. [6]considered the robustness of LDF under three specific distributions and the case of independent variables. These distributions were considered to be non-normal and were generated from the normal distributions using the Johnson system of transformations (i.e. log normal, inverse hyperbolic sine normal and logit normal distribution). They observed considerable decline in performance of the LDF (the log normal distribution used had extremely large skewness and kurtosis). Based on their results, Fisher's LDF was greatly affected by non-normality in the population. They concluded that, the use of Fisher's LDF under non-normality contamination situations could be badly misleading and recommended that the data be transformed to approximate normality prior to the use of the LDF.

Departures from the assumptions of linear discriminant function analysis were explored by [5], where the effects of unequal covariances on the linear discriminant method werestudied. In spite of theoretical evidence supporting the use of the QDF when covariances are heterogeneous, its actual employment has been sporadic because there are unanswered questions regarding its performance in the practical situation where the discriminant function must be constructed using training samples that do not satisfy the classical assumption of the model. [8] investigated into the application of discrimination and classification on poultry feeds data.

They employed Fisher’s linear discriminant function for providing maximum separation between the two groups of eggs of which the chicken were fed with different combinations of feeds. However they proposed a linear discriminant function for the classification of eggs based on the size and cholesterol level. The function gave a good prediction based on the estimated values obtained from the Apparent Error Rates (APER) and Absolute Error Rate (AER).

[7] applied discriminant analysis in differentiating between the signal patterns of healthy subjects and those of individuals with specific heart conditions based on diagnosis of ECG signals. An approach for classifying multivariate ECG signals based on discriminant and waveletanalyses was proposed. [3] studied the variable selection criterion for linear discriminant rule and its optimality in high dimensional and large sample data. They suggested that, a new variable selection procedure called Misclassification Error Criterion (MEC) for linear discriminant rule for high dimensional data set be set up. Their study found that the MEC not only asymptotically decomposes into 'fitting' and 'penalty' terms but also possesses an asymptotic optimality in the sense that MEC achieves the smallest possible conditional probability of misclassification in candidate variable sets. After the simulation studies, the study discovered that MEC has good performances in the sense of selecting the true variable sets.

Predicting hospitalisation of patients with diabetes Mellitus; an application of the Bayesian discriminant analysis was studied by [1]. The main objectiveof his study was to develop and test a Bayesian discrimination model for the purpose of identifying both the personal and the healthcare system characteristics predictive of hospitalisation for the treatment of patients with diabetes Mellitus or commonly observed cormorbidities associated with the disease. The model was then tested by using a logit regression technique in order to estimate the probability of one or more hospitalisation events among patients with diabetes. Claims data extracted from the Hawaii Medical Service Association (HMSA) Private Business Claims (PBS) files for the 1995 calendar year was used. The model was able to correctly classify 90 percent of the observations. The study also found that multivariate discriminant analysis using a logit regression model successfully identifies important explanatory variables predictive of hospitalisation and as well as assigns patients into 1 of 2 mutually exclusive classes.

2.0 Materials and Methodology

2.1 Data Used

A seven variate data set consisting of 620 diabetes patients either type 2 or type 1 diabetes were obtained from KATH, in Ghana and was used for data analysis. The seven measured variables included their Age, Weight (*Wt*), Height (*Ht*), Systolic Blood Pressure (*BPS*), Diastolic Blood Pressure (*DPS*), Fasten Blood Sugar (*FBS*) and Body Mass Index (*BMI*).

2.2Discrimination and Classification of Two Populations

Let $f_1(x)$ and $f_2(x)$ denote the probability density function associated with a single vector random variable ***X***forthe populations π_1 and π_2 respectively.Considering an observed value $X = (x_1,..., x_p)^T$,we assign a vector *X* to either population π_1 or π_2 . Let Ω be the set ofcollection of all possible outcomes of *X*, hence, the partition of the sample space is given as $\Omega = R_1 U R_2$ where R_1 is the subspace of outcomes which we classify as belonging to population π_1 and $R_2 = \Omega - R_1$ the subspace of outcomes classified as belonging to π_2 . Therefore the conditional probability of classifying an object as belonging to π_j when it really comes from π_i becomes:

$$P(i \mid j) = P(X \in R_j \mid X \in \pi_i) = \int_{R_j} f_i(x)dx \quad , \forall i, j \quad i \neq j \qquad (1)$$

The conditional probabilities can also be obtained for $i = j$ when $i, j = 1,2$.

Let $P_i = P(X \in \pi_i)$, $i = 1,2$ be the prior probability of π_i where $P_1 + P_2 = 1$. The overall probabilities of correctly and incorrectly classifying observations are:

P(object is correctly classified as π_i) = $P(X \in R_i \mid X \in \pi_i) = P(X \in \pi_i) = P(i \mid i)p_i$ where $i = 1,2$. *P*(object is misclassified as π_i) $= P(X \in R_i \mid X \in \pi_j) = P(X \in \pi_j) = P(i \mid i)p_j$ where $i \neq j$.

2.3 Cost of Misclassification

Let $c(i \mid j)$ denote the cost of classifying an object/observation into π_i when actually belongs to π_j. Where the Expected Cost of Misclassification (ECM) is derived as:

$$ECM = c(2 \mid 1)P(2 \mid 1)p_1 + c(1 \mid 2)P(1 \mid 2)p_2 \qquad (2)$$

With p_1 and p_2 being the prior probabilities for the two populations. The two regions R_1 and R_2 below are used to minimized the expected cost of misclassification.

$$R_1 = \left\{ x \in \Omega; \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \right\} (3)$$

$$R_2 = \left\{ x \in \Omega; \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \right\} (4)$$

[4].

2.4 Classification with Two Multivariate Normal Populations when $\Sigma_1 = \Sigma_2$

The density function of $X' = (x_1, x_2, ..., x_p)$ for the two populations π_1 and π_2 is given by

$$f_i(x) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp - \frac{1}{2}(x - \mu_i)'\Sigma^{-1}(x - \mu_i)$$

If the population parameters μ_1, μ_2 and Σ are known, then after cancellation the allocation rule after minimising the Expected Cost of Misclassification (ECM) becomes
Allocate *x* to π_1 if

$$(\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \left[\ln\left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \right] \qquad (5)$$

[4].

2.5 Inferential Procedures in Discriminant Analysis

Several inferential procedures exists in discriminant function analysis. The basic ones are discussed here.

2.5.1 Test for $H_0 : \mu_1 = \mu_2$ when $\Sigma_1 = \Sigma_2$ using Hoteling's T^2-test

We assume that two independent random samples $y_{11}, y_{12}, ... y_{1n_1}$ and $y_{21}, y_{22}, ... y_{2n_2}$ are drawn from $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$ where Σ_1 and Σ_2 are known. In order to obtain a T^2 test we assume that $\Sigma_1 = \Sigma_2 = \Sigma$. From the samples, we calculate $\bar{y}_1, \bar{y}_2$, $W_1 = (n_1 - 1)S_1$ and $W_2 = (n_2 - 1)S_2$.

A pooled estimator of the covariance matrix is calculated as $S_{pl} = \dfrac{W_1 + W_2}{n_1 + n_2 - 2}$ for which $E(S_{pl}) = \Sigma$ hence in testing the equality of the mean vectors we use the test statistics

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)'S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) \qquad (6)$$

Which is distributed as $T_p^2, n_1 + n_2 - 2$ when H_0 is true. We reject H_0 if $T^2 \geq T^2_{\alpha, n_1 + n_2 - 2}$.

2.5.2 Wilks Likelihood Ratio Test

If $y_{ij}, i = 1,2,...,g$, $j = 1,2,...,n$ are independently observed from $N_p(\mu_i, \Sigma)$, then the likelihood ratio test statistics for $H_0 : \mu_1 = \mu_2 = ... = \mu_g$ can be expressed as

$$\Lambda = \frac{|E|}{|E+H|} \qquad (7)$$

Where $H = n\sum_{i=1}^{g}(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})'$, $E = \sum_{I=1}^{g}\sum_{j=1}^{n}(\bar{y}_{ij} - \bar{y}_i)(\bar{y}_{ij} - \bar{y}_i)'$

The test statistics is distributed as the Wilk's Λ-distribution. We reject H_0 if $\Lambda \le \Lambda_{\alpha, p, \upsilon_H, \upsilon_E}$. p, υ_H and υ_E are the dimensions and degrees of freedom for hypothesis and error respectively.

2.5.3 Box's M-Test

For a one way MANOVA with g groups $(g \ge 2)$ the assumption of equality of covariance matrices can be stated as a hypothesis to be tested: $H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_g$ Versus H_1: at least two Σ_i's are unequal. Define

$$W_i = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)' \text{ and } M = \frac{|S_1|^{\upsilon_1/2}|S_2|^{\upsilon_2/2}...|S_g|^{\upsilon g/2}}{|S_{pl}|\sum_i \upsilon_i/2} \quad (8)$$

where $\upsilon_i = n_i - 1$, $S_i = W_i/\upsilon_i$ is the unbiased sample covariance matrix and $S_{pl} = \frac{\sum_{i=1}^{g}\upsilon_i S_i}{\sum_{i=1}^{g}\upsilon_i} = \frac{E}{\upsilon_E}$.

The statistics

$$u = -2(1 - c_1)\ln M \qquad (9)$$

Has an approximated χ^2 distribution with $\frac{1}{2}(k-1)p(p+1)$ degrees of freedom where

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

, we reject H_0 if $u > \chi^2_\alpha$

2.6 Error Rate Estimation

The performance of any classification procedure is based on the error rates or misclassification probabilities.

2.6.1Cross Validation

Let n_{1M}^{CV} and n_{2M}^{CV} denote the number of left out observations misclassified in group 1 and 2 respectively and it's given by $CV = \frac{n_{1M}^{CV} + n_{2M}^{CV}}{n_1 + n_2}$ (10)

3.0 Results and Discussion

First and foremost, Box M test of equality of covariance matrix of the two groups (Type 1 and Type 2) diabetic patients were tested. Table 1 shows the test of homogeneity of the two covariance matrices for Type 1 and Type 2 diabetic patients groups. From the table the log determinant values for the two groups as well as the pooled within groups were obtained. The three log determinant values as observed from Table 1 are almost the same indicating that the covariance matrices of the two diabetic patient groups are equal. The *P-value* of 0.350 which is greater than the significant level (*α=0.05*) indicates that the two covariance matrices are equal (i.e. $H_0 : \Sigma_1 = \Sigma_2$) and hence the data do not differ significantly from multivariate normal distribution.

3.1 Test of Equality of the Two Mean Vectors

Hotelling T^2 was then used to test for the equality of the mean vectors for Type 2 and Type 1 diabetic patients. The hypothesis tested for equality of the two mean vectors were:

$H_0 : \mu_1 = \mu_2$ Vs. $H_0 : \mu_1 \neq \mu_2$. And the mean vectors for Type 1 and type 2are

$$\mu_1 = \begin{array}{r} Age \\ Wt \\ Ht \\ BPS \\ DPS \\ FBS \\ BMI \end{array} \begin{pmatrix} 29.98 \\ 67.22 \\ 1.64 \\ 130.51 \\ 80.02 \\ 11.11 \\ 25.06 \end{pmatrix} \text{and } \mu_2 = \begin{array}{r} Age \\ Wt \\ Ht \\ BPS \\ DBS \\ FBS \\ BMI \end{array} \begin{pmatrix} 57.22 \\ 67.61 \\ 1.64 \\ 134.90 \\ 81.47 \\ 8.93 \\ 25.13 \end{pmatrix} \text{ respectively.}$$

From Table 2 of test for equality of the two mean vectors, the *p-value*=0.000 is less than significant level (*α=0.05*), hence we reject the null hypothesis (H_0) and conclude that the two mean vectors of the diabetes patient groups are not equal.

Hence the pooled within group covariance matrix as well as the bivariate correlation coefficients are computed and shown in Table 3. As already indicated, the pooled within group covariance matrix satisfies one of the assumptions of Linear discriminant function and the bivariate correlation coefficients detects potential problems with multicollinearity. From the correlation table, it is clear that, none of the bivariate correlations between two of the measured variables were even closer to 0.80. This means that, multicollinearity was not observed among any two of the seven independent variables.

Next was to derive the canonical discriminant function for providing maximum separation between types 1 and 2 diabetic patients based on the identified seven independent variables. The eigenvalues table (i.e. Table 4) shows the eigenvalues of the discriminant function as well as the canonical correlation for the discriminant function. The larger the eigenvalue, the more amount of variance shared in the linear combination of variables. Since only one function is involved, the function then explains majority of variance in the relationship. An eigenvalue of 0.414 and the percentage variance of 100 percent for the function indicates that, the derived discriminant function explains 100 percent variation in the relationship. This therefore reveals the importance of the discriminant function in the provision of maximum separation between the groups. Also since only one discriminant function was involved, the cumulative percentage of the variance was recorded as 100 percent. From the same table the Canonical correlation value was observed to be 0.541 and it explains an above average relationship between the discriminant scores and the levels of the dependent variable.

Wilks lambda

Wilks' Lambda is the ratio of within-groups sums of squares to the total sums of squares. This is the proportion of the total variance in the discriminant scores not explained by differences among groups. A small lambda indicates that group means appear to differ. The Wilks lambda value of 0.707 from Table 5 indicates that not all of the independent variables contribute significance in the function. The table also provides a Chi-Square statistic to test the significance of Wilk's Lambda.

As evident from the table, the *Wilks lambda* of 0.707, the *chi-square statistics* of 212.812 with *p-value* of 0.000 is less than the significant level (α) of 0.05 and hence, the derived discriminant function explains the group membership well, thus, the group means appear to differ.

Table 6 summarises the output of the standardised canonical discriminant function coefficient and the structure matrix. The standardised canonical discriminant function coefficient was used to rank the importance of each of the seven independent variables. From Table 6, the standardised canonical discriminant function coefficient for the *age*and *BMI* variables was observed to be 0.979 and 0.307 respectively.

This means that, the group separation depends mostly on the *age* and *BMI*of the patients. In other words a patient is being diagnosed as being type 1 or type 2 diabetic status based on their *age* and measured *BMI*. Other variables that the group separation depended on were the patient's height, systolic Blood pressure and the diastolic blood pressure. From Table 6, the canonical structure matrix revealed the correlations between each variable in the model and the discriminant function. It is expected that, a variable with correlation of 0.3 or more is considered to be very important. Similarly the age of patients was observed to be a major determining factor in classifying a patient as either type 1 or type 2 since a strong correlation of 0.989 between the ages and the function was observed. Also a weak positive correlation between the *BPS*, *Ht*, *DBP*, *Wt*, *BMI* and the function was observed.

The canonical discriminant function coefficients/ Fishers LinearDiscriminant Function obtained from the study was:

$$D_{12} = -7.176 + 0.078 Age - 0.024 Wt + 1.746 Ht + 0.004 BPS - 0.001 BPD - 0.025 FBS + 0.051 BMI \quad (11)$$

Hence based on the derived Fishers Discriminant Function, the classification rule for the two diabetes patient groups were obtained and was used to compute the discriminant scores for classifying the original observations into their respective groups. The classification rules obtained are:

$$D_1 = -1128.14 + 0.25 Age - 12.39 Wt + 1351 Ht + 0.384 BPS + 0.08 DBP + 0.24 FBS + 32.18 BMI \quad (12)$$

$$D_2 = -1141.91 + 0.42 Age - 12.45 Wt + 135535 Ht + 0.39 BPS + 0.08 DBP + 0.18 FBS + 32.29 BMI \quad (13)$$

From Table 7, the computed discriminant scores for Type 1 (D_1) and Type 2 (D $_2$) diabetes groups were able to correctly classify 85.3 percent of the original observations into their respective groups. Also based on the error rates obtained by the cross-validation method as evident in Table 7, 84.8 percent of the cross validated grouped cases were correctly classified. The results shows a clear indication that, the derived FLDF as well as the classification rules provided maximum separation between the two main diabetes group patients (i.e. either a type 1 or type 2 diabetic patient).

4.0 Conclusion

This study focused on deriving a discriminant function based on some identified variables in providing maximum separation between two groups of diabetes patients at Komfo Anokye teaching Hospital. Fishers Linear Discriminant Function based on the seven measured variables as well as the corresponding classification rule were developed. From the study 85.3 percent of the original observations were correctly classified whilst 84.8 percent of cross-validated observation were correctly classified. Also the classification of the patients into their respective diagnosed diabetes status depended hugely on the patient's*age* as well as their *BMI* and to some small extent their *FBS*. The derived linear discriminant function provided maximum separation and the classification rule obtained will be use to classify future diabetic patients with similar identified variables as whether the person will belong to a type 1 or type 2 diabetes status group.

5.0 References

Bhattacharyya, S. (1998). Predicting hospitalisation of patients with diabetes mellitus: an application of the Bayesian discriminant analysis. Pharmacoeconomics 13, 519–29.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7, 179–188.

Hyodo, M. & Kubokawa, T. (2014). A variable selection criterion for linear discriminant rule and its optimality in high dimensional and large sample data. Journal of Multivariate Analysis 123, 364–379. 270

Johnson, R. A. & Wichern, D. (2007). Applied Multivariate Statistical Analysis. NJ: Pearson Education Inc.

Krzanowski, W. J. (1977). The performance of fisher's linear discriminant function under non-optimal conditions. Technometrics 19, 191–200.

Lachenbruch, P. A., Sneeringer, C. & Revo, L. T. (1973). Robustness of the linear and quadratic discriminant functions to certain types of non-normality. Comm. Statist 1, 39–56. 275

Maharaj, E. A. & Alonso, A. (2014). Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals. Computational Statistics and Data Analysis 70, 67–87.

Olosunde, A. & Soyinka, A. (2013). Discrimination and classification of poultry feeds data. International Journal of Mathematical Research 2, 37–44.

6.0 Tables

Table 1: Box's M Test of Equality of Covariance Matrices

Log Determinants			**Test Results**	
Groups (*DM Types*)	*Rank*	*Log determinants*	Box's M	199.5
Type 2	7	22.107	F Approx.	6.814
Type 1	7	21.971	df1	28
Pooled within groups	7	22.553	df2	35338.49
			P-value	0.35

Table 2: TestorEqualityofMeanVectors

HotellingT^2	*P- value*	*Significantlevel*(α)
255.7643	0.000	0.05

Table 3:Pooled within Group Matrices

		Age	Wt	Ht	BPS	DBP	FBS	BMI
Covariance	Age	159.185	11.193	.056	4.026	-.220	-6.350	1.656
	Wt	11.193	238.951	.421	52.837	36.374	-4.579	75.701
	Ht	.056	.421	.009	.014	-.100	-.008	-.162
	BPS	4.026	52.837	.014	518.792	164.189	12.026	17.114
	DBP	-.220	36.374	-.100	164.189	1009.707	-.818	16.262
	FBS	-6.350	-4.579	-.008	12.026	-.818	23.287	-1.350
	BMI	1.656	75.701	-.162	17.114	16.262	-1.350	36.499
Correlation	Age	1.000	.057	.047	.014	-.001	-.104	.022
	Wt	.057	1.000	.288	.150	.074	-.061	.811
	Ht	.047	.288	1.000	.007	-.033	-.017	-.284
	BPS	.014	.150	.007	1.000	.227	.109	.124
	DBP	-.001	.074	-.033	.227	1.000	-.005	.085
	FBS	-.104	-.061	-.017	.109	-.005	1.000	-.046
	BMI	.022	.811	-.284	.124	.085	-.046	1.000

Table 4: EigenValues

Function(s)	*Eigenvalue*	*% of Variance*	*Cumulative %*	*Canonical correlation*
1	0.414	100.0	100.0	0.541

Table 5: Wilk's Lambda

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*df*	*P-value*	*Significance level(α)*
1	.707	212.812	7	.000	0.05

Table 6: Table ofStandardizedCanonical Discriminant FunctionCoefficients andStructure Matrix

Standardise canonical discriminant function coefficient		*Structure Matrix*	
Variables	Function	Variables	Function
Age	.979	Age	.986
Weight	-.364	FBS	-.206
Height	.165	SBP	.088
SBP	.102	Height	.022
DBP	.004	DBP	.021
FBS	-.121	Weight	.012
BMI	.307	BMI	.005

Table 7: Classification Results

		DM TYPES	*Predicted Group Membership*		Total
			Type 1	Type 2	
Original	Count	Type 1	57	2	59
		Type 2	89	472	561
	%	Type 1	96.6	3.4	100.0
		Type 2	15.9	84.1	100.0
Cross-validated	Count	Type 1	57	2	59
		Type 2	92	469	561
	%	Type 1	96.6	3.4	100.0
		Type 2	16.4	83.6	100.0

DM=Diabetes Mellitus