An - Najah Univ. J. Res. (Humanities). Vol. 28(6), 2014

A Comparative Study between Linear Discriminant Analysis and Multinomial Logistic Regression

دراسة مقارنة بين التحليل التمييزي الخطي والانحدار اللوجستي المتعدد

Abdalla El-habil & Majed El-Jazzar

عبد الله الهبيل، وماجد الجزار

Head of Department of Applied Statistics, Faculty of Economics & Administrative Sciences, Al-Azhar University, Gaza - Palestine.

E-mail: abdalla20022002@yahoo.com

Received: (28/1/2013). Accepted: (23/12/2013)

Abstract

This paper aimed to compare between the two different methods of classification: linear discriminant analysis (LDA) and multinomial logistic regression (MLR) using the overall classification accuracy, investigating their quality of prediction in terms of sensitivity and specificity, and examining area under the ROC curve (AUC) in order to make the choice between the two methods easier, and to understand how the two models behave under different data and group characteristics. Model performance had been assessed from two special cases of the k-fold partitioning technique, the 'leave-one-out' and 'hold out' procedures. The performance evaluation for the two methods was carried out using real data and also by simulation. Results show that logistic regression slightly exceeds linear discriminant analysis in the correct classification rate, but when taking into account sensitivity, specificity and AUC, the differences in the AUC were negligible. By simulation, we examined the impact of changes regarding the sample size, distance between group means, categorization, and correlation matrices between the predictors on the performance of each method. Results indicate that the variation in sample size, values of Euclidean distance, different number of categories have similar impact on the result for the two methods, and both methods LDA and MLR show a significant improvement in classification accuracy in the absence of multicollinearity among the explanatory variables.

Key words: Hold- out validation - cross-validation – confusion Matrix - sensitivity – specificity - overall classification accuracy.

ملخص

هدفت هذه الدراسة إلى إجراء مقارنة بين أسلوبين من أساليب التصنيف و التنبؤ، وهي التحليل التمييزي الخطي وأسلوب الانحدار اللوجستي المتعدد وذلك لفهم كيفية عمل كلا النموذجين في التصنيف والتنبؤ تحت تأثير الخصائص والصفات المختلفة للبيانات. تم في هذه الدراسة تقييم كلا الأسلوبين من خلال استخدام مجموعة بيانات حقيقية حيث كان المعيار المستخدم للمقارنة بين هذين النموذجين هو دقة التصنيف التي تم حسابها بطريقتين مختلفتين والمساحة تحت المنحنى AUC لتحليل الـ ROC، كما تم توليد بيانات باستخدام برامج الحاسوب بحيث تحقق هذه البيانات الافتراضات الأساسية لنموذج التحليل التمييزي الخطي في أنها تتبع التوزيع الطبيعي المتعدد وتتساوى فيها مصفوفة التباينات المشتركة، وذلك بهدف مقارنة قدرة كل من النموذجين على التصنيف والتنبؤ تحت تأثير الاختلاف في حجم البيانات وعدد فئات المتغير التابع والمسافة بين متوسطات المجموعات التي تحتاج إلى تصنيف والارتباط الداخلي بين المتغيرات المستقلة. عند تطبيق كلا النموذجين على البيانات الحقيقية، فقد وجد أن النتائج كانت متشابهة لكلا النموذجين من حيث المعاملات التي تم تقديرها والتي يمكن استخدامها للتنبؤ بالإصابة أو عدم الإصابة بمرض السكري، ورغم أن دقة التصنيف لأسلوب الانحدار اللوجستي كانت أعلى بقليل من دقة التصنيف لأسلوب التحليل التمييزي، إلا أنه عند أخذ معيار sensitivity و specificity والمساحة تحت المنحنى AUC لتحليل الـROC فقد وجد أن الفرق بين كلا النموذجين في التصنيف كان ضئيلاً. وفي حالة البيانات المولدة فقد أكدت النتائج أن تأثير الاختلاف في حجم البيانات والاختلاف في عدد فئات المتغير التابع والمسافة بين متوسطات المجموعات كان لها نفس الأثر على أداء كل من النموذجين، وكلا النموذجين كان أداؤهما في التصنيف أفضل في ظل عدم وجود ارتباط داخلي بين المتغيرات المستقلة.

1. Introduction

Two of the most widely used statistical methods for analyzing categorical outcome variables are linear discriminant analysis and logistic regression. This paper will fundamentally evaluate the efficacy of linear discriminant analysis (LDA) and multinomial logistic regression (MLR) in terms of multi-group classification problems. Even though the two techniques often reveal the same patterns in a set of data, and both are appropriate for the development of linear classification models, the two methods differ in their basic idea, they do so in different ways and require different assumptions. Linear discriminant analysis makes more assumptions about the underlying data, while MLR makes no assumptions on the distribution of the explanatory data. Hence, it is assumed that logistic regression is the more flexible and more robust method in case of violations of these assumptions, so setting some guidelines for proper choice between the two methods is required.

There are various studies in literature that have been performed to compare and evaluate the performance of linear discriminant analysis and multinomial logistic regression: (Wim Van den Noortgate & Paul De Boeck, 2005, p. 443) present logistic mixed models that can be used to model uniform differential item functioning (DIF), treating the item effects and their interaction with groups (DIF) as random. They showed that the logistic mixed models approach is not only a comprehensive and economical way to detect these different kinds of DIF; it also encourages exploring explanations of DIF by including group or item covariates in the model. (Tianshu Pan, 2008) presents the multilevel logistic regression (MMLR) models to detect DIF, which are likely to detect DIF when the responses of an examinee are not locally independent. He compared the uses of the three MMLR models, three modified versions of kamata's Hierarchical Generalized Linear Model (HGLM) and the standard logistic regression model as DIF detection methods. His study results show that MMLR can be used for DIF detection. It also found that the heterogeneous variances of the two groups influence power and type I error rates of these methods, and the HGLM DIF models are unsuitable to identify DIF. (Yovhane L. Metcalfe, 2012) examined the factors on which students veterans with disabilities differed from their student veteran peers without reported disabilities by using univariate tests of significance, a logistic regression, and a discriminant function analysis. Univariate tests of significance revealed that students with disabilities had a significantly lower mean GPA, where more often male, tended to favor certain academic majors over others, more often enrolled in bachelor's degree versus associate and certificate programs, and had a lower risk of attrition based on their index of risk. Major, degree program type, and risk index proved to be the most significant predictors of disability status in logistic regression and discriminant function analysis. (Pohar, Blas, & Turk, 2004, p. 143) considered the problem of choosing between the LDA and MLR by using several simulated datasets, they concluded that linear discriminant analysis is a more appropriate method when the explanatory variables are normally distributed, and logistic regression overcomes discriminant analysis only when the number of categories is small and the results of LDA and MLR are close whenever the normality assumptions are not too badly violated. (Press & Wilson 1978, p. 699) compared logistic regression and parametric discriminant analysis in terms of the proportion of correct classifications and concluded that logistic regression with maximum likelihood estimators is preferable to parametric discriminant analysis in cases for which the variables do not have multivariate normal distributions, but if the populations are normal with identical covariance matrices, discriminant analysis estimators are preferred to logistic regression estimators for the discriminant analysis problem. (Hossain, Wright, & Peterson 2002, p. 400) compared the performance of multinomial logistic regression (MLR) and discriminant analysis (DA) models to predict arrival time at the hospital; the goal was to determine the best statistical method for prediction of arrival intervals for patients with acute myocardial infarction symptoms. The correct classifications were 62.6% by MLR, 62.4% by DA using proportional prior probabilities, and 48.1% using equal prior probabilities of the groups. (Montgomery, White, & Martin 1987, p. 495) concluded that logistic regression is preferable to discriminant analysis particularly when the assumptions of normality and equal variance are not met. (Kiang 2003, p. 441) stated that the logistic model is superior to DA in all cases, especially when the normality, linearity, and identical covariance assumptions do not hold and only the normality assumption has an impact on DA.

Hence we will try to see how these two models behave under different data and group characteristics and if there is a significant difference between them.

The paper is organized as follows: Section 2 recalls the theoretical concepts, section 3 data, section 4 simulations and section 5 conclusion and recommendations.

2. Theoretical concepts

Linear Discriminant Analysis

Discriminant analysis is a statistical technique that allows one to understand the differences of objects between two or more groups with respect to several variables simultaneously. It is the first multivariate statistical classification method used for decades by researchers and practitioners in developing classification models (Hamid, 2010). Discriminant analysis is used in situations where the clusters are known a priori. The aim of discriminant analysis is to classify an observation, or several observations, into these known groups. The exploratory multivariate procedure of determining variables and a reduced set of functions called discriminant analysis. In general, discriminant analysis concerns with the development of a rule for allocating objects into one of some distinct groups. Then, the constructed classification rule will be used to determine a group of some future objects see (Timm, 2002).

Multinomial Logistic Regression

Logistic regression (LR) is statistical modeling method for categorical data has expanded from its origins in biomedical research to fields such as business and finance, engineering, marketing, economics, and health policy (Meyers, Gamst, Guarino, 2006). Logistic regression has found two broad applications in applied research: classification (predicting group membership) and profiling (differentiating between two groups based on certain factors) (Tansey, White, & Long, 1996, p. 339). The assumptions required for statistical tests in logistic regression are far less restrictive than those for ordinary least squares (OLS) regression (Dinesh, 2009, p. 15). There is no formal requirement for multivariate normality, homoscedasticity, or linearity of the independent variables within each category of the dependent variable (Allison, 1999). Also there is an important difference between logistic regression model and the linear regression model concerning the nature of the relationship between the outcome and independent variable (Hosmer & Lemeshow, 2000).

Performance Criteria

The objective of building a classification rule is to correctly classify as many future units as possible. There are several criteria available to evaluate a set of classification rules. The simplest and the most frequently used criterion for comparison between two methods is error rate or misclassification rate (Harrell, 1997). A simple estimate of the error rate can be obtained by trying out the classification procedure on the same data set that has been used to compute the classification functions. This is called the substitution or re-substitution method (Rencher, 2002).

The Hold-Out Method

Another evaluation procedure is to split the total sample into a training sample and a validation sample. The training sample is used to construct the classification function and the validation sample is used to evaluate it. The error rate is determined by the proportion misclassified in the validation sample. This method overcomes the bias problem by not using the same data to both build and judge the classification function. There are problems associated with holdout method (Hussein, 2010, p. 45). First, this method requires a large sample size, but in applications large sample is not always available. Second, the classification rule that is validated is not the one that would actually be used. Third, there is a problem associated with the appropriate relative size of the training sample to the testing sample.

The Cross-Validation Method

In most real applications, only a limited amount of data is available, which leads to the idea of splitting the data: Part of data (the training sample) is used for training the algorithm, and the remaining data (the validation sample) are used for evaluating the performance of the algorithm. The validation sample can play the role of new data (Arlot & Alain, 2010, p. 40). In this method, $\binom{n}{k}$ classifiers are designed. Each classifier is designed by choosing k of the n observations as a training set, and its error rate is estimated using the remaining ($n-k$) observations. This process is repeated for all distinct choices of k patterns and the average of the error rates is computed. The average of the error rate of each subset is therefore an estimate of the error rate of the classifier (Kotsiantis, 2007, p. 159). This type of validation is, of course, more expensive computationally, but useful when the most accurate estimate of a classifier's error rate is required. A popular choice for the value of k is, yielding the well-known leave-one-out method.

Leave-one-out cross-validation

As the name suggests, leave-one-out cross-validation (LOOCV) involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sample. The actual classification rule for future observations would be based on all N observations. As before the average error is computed and used to evaluate the model. The evaluation given by leave-one-out cross validation error (LOOCV) is good, but at first view it seems very expensive to compute because of the large number of times the training process is repeated.

Evaluation of a Classification Model

For the evaluation of two methods, sensitivity, specificity can be also measured in the same dataset. Sensitivity of a binary classification test with respect to some class is a measure of how well this test identifies a condition and expresses the probability of a case being classified in that class, meaning the proportion of true positives of all positive cases in the population. Specificity, on the other hand, expresses the proportion of the true negative classified cases of a binary classification test of all the negative cases in the population. If the number of classes is two, for example, Table 2.1 shows the predicted classification and true classification. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), are the four different possible outcomes of classification prediction for a two-class case with classes “1” and “0”.

The sensitivity is calculated by = $\frac{\mathbf{TP}}{\mathbf{TP + FN}}$, and the specificity is $\frac{\mathbf{TN}}{\mathbf{TN + FP}}$

Table (2.1): True class and predicted class.

Actual Group	Predicted Group	
	1	**2**
1	**TP**	**FP**
2	**FN**	**TN**

Both the sensitivity and specificity are usually given in percentages. A decision method is considered good if it simultaneously has a high sensitivity and a high specificity, so there is a trade-off between sensitivity and specificity (Tom Fawcett, 2004).

3. Data

By using the R software statistical package program, linear discriminant analysis and logistic regression methods are compared using the diabetes data as a case study. These data were collected by the National Institute of Diabetes and Digestive and Kidney Diseases (web1). A population of women who were at least 21 years old, of Pima Indian Heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization (WHO) criteria. The classification task consists of predicting whether a patient would test positive for diabetes. The class labels of the Pima data are 1 for diabetes and 0 otherwise. There are 8 predictor variables for 768 patients, all females, at least 21 years of age, and among the 768 patients, 268 tested positive for diabetes (class 1) according to WHO criteria. For details the frequency and percentage distributions of the groups are presented in table 3.1

Table (3.1): The Class Distribution of the Diabetes Data

class name	class size	class distribution
Positive	268	34.9 %
Negative	500	65.1%

From the table 3.1, it is seen that the data set, actually consisted of two sub-groups with n_1= 268 (34.9%) cases in the first group, and n_2=500 (65.1%) in the second group. The data set has 768 cases, all with the following numeric attributes and they are: “pregnant” number of times pregnant, “glucose”plasma glucose concentration, “pressure” diastolic blood pressure, “triceps” triceps skin fold thickness, “insulin”two hour serum insulin, “mass” body mass index, “pedigree” diabetes pedigree function, and “age” age in years. The Class variable (9) is treated as 0 (false), 1 (true – tested positive for diabetes). A brief statistical analysis is given in table 3.2.

Table (3.2): Brief Statistical Analysis of Diabetes Data.

Attribute	Mean	Std. Deviation	Min/Max
pregnant	3.85	3.370	0/17
plasma glucose	120.89	31.973	0/199
pressure	69.11	19.356	0/122
triceps skin fold thickness	20.54	15.952	0/99

… Continue table (3.2)

Attribute	Mean	Std. Deviation	Min/Max
insulin	79.80	115.244	0/846
body mass index	31.993	7.884	0/67.1
diabetes pedigree	.47188	.331	.078/2.42
age in years	33.24	11.760	21/81

Linear Discriminate Analysis

The main assumptions of LDA are tested here. Shapiro test statistic for the multivariate normality of Diabetes data had a value of 0.9056 with p-value < 0.001. Since the p value is less than 0.05 (the level of significance for the test), we conclude that the data are not multivariate normally distributed. Also by Box's M test the assumption of equal covariance for equality of the group covariance matrices was tested. In this case, Box's M statistic had a value of 229.559 with a p-value < 0.001, so there is sufficient evidence that we reject the hypothesis that the groups' covariance matrices are equal.

Significance tests and strength of relationship statistics for each discriminant function for the diabetes grouping are presented in the Wilks' Lambda Table From table 3.3 we can conclude that the corresponding function explains the group membership well.

Table (3.3): Wilks' Lambda Table.

Wilks' Lambda	Chi-square	df	Sig.
0.697	275.336	8	< 0.001

Table 3.4, the standardized canonical discriminant function coefficients, which measure the relative importance of the selected variables, the larger absolute value of the coefficient corresponds to greater discriminating ability, and mean that the groups differ a lot on that variable, indicate that the independent variable “plasma glucose” was the most powerful discriminating variable, followed by “body mass index” and “number of times

pregnant", while "triceps skin fold thickness", "insulin" and "age" were less successful as predictors.

Table (3.4): Standardized Canonical Discriminant Function Coefficients.

Variables	Function1
Pregnant	0.309
plasma glucose	0.764
Pressure	-0.205
triceps skin fold thickness	0.011
Insulin	-0.094
body mass index	0.455
diabetes pedigree	0.219
age in years	0.137

To compare these two groups, two classification functions were used to assign cases into each group (table 3.5). For each observation, two classification scores were computed for each function. The cases were assigned to the group whose function obtained the higher score.

Table (3.5): Linear Discriminate Function Coefficients.

Variables	LDA functions	
	tested negative	tested positive
pregnant	-0.056	0.074
plasma glucose	0.116	0.153
pressure	0.093	0.078
triceps skin fold thickness	0.005	0.006
insulin	-0.010	-0.011
body mass index	0.442	0.525
diabetes pedigree	2.807	3.735
age in years	0.164	0.181
(Constant)	-19.435	-27.926

Logistic Regression analysis

Logistic regression analysis was performed on the diabetes data set. The presence of a relationship between the dependent variable and combination of independent variables is based on the statistical significance of the final model chi-square.

Table (3.6): Model Fitting Information.

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	Df	Sig.
Intercept Only	993.484	270.039	8	.000
Final	723.445			

According to the results shown in table 3.6, it is seen that -2 log likelihood value of basic model only with intercept term was 993.484, this value decreased into 723.445 with the independent variables appearance in the model. In this analysis, the probability of the model chi-square (270.039) was 0.00, less than the level of significance (0.05). The null hypothesis that there was no difference between the model without independent variables and the model with independent variables was rejected. The existence of a relationship between the independent variables and the dependent variable was supported.

Table (3.7): Pseudo $\mathbf{R}^{\square}$ measurements.

Measurements	$\mathbf{R}^{\square}$ values
Cox and Snell	0.296
Nagelkerke	0.408
McFadden	0.272

Nagelkerke R^2 value is the modified form of Cox & Snell coefficient. According to the results shown in table 3.7, it is seen that dependent variables define 29.6% of the variance in independent variables according to Cox & Snell R^2 value, 40.8% according to Nagelkerke R^2 value, and 27.2% according to

McFadden value. The likelihood ratio test evaluates the overall relationship between an independent variable and the dependent variable. Statistics of likelihood ratio tests obtained from MLR are presented in table 3.8.

Table (3.8): Likelihood Ratio Tests.

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	Df	Sig.
Intercept	934.653	211.207	1	.000
preg	738.678	15.233	1	.000
plas	838.372	114.927	1	.000
pres	729.993	6.548	1	.011
skin	723.453	.008	1	.929
insu	725.187	1.742	1	.187
mass	764.225	40.779	1	.000
pedi	733.785	10.340	1	.001
age	725.968	2.522	1	.112

According to the results shown in table 3.8, it is seen that there is a statistically significant relationship between the independent variables (preg, plas, pres, mass, and pedi) and the dependent variable, so it play significant role in the cause of diabetes. Other variables (skin, insu, and age) contributions into the model are not significant which implies that these factors don’t play significant role in the cause of diabetes.

Table (3.9): Results of Fitting the Logistic Regression Model to the Diabetes Data.

Variable	Coeff.	Std. Error	Wald	df	Sig.	Exp(B)
Intercept	-8.405	0.717	137.546	1	0.000	
preg	0.123	0.032	14.747	1	0.000	1.131
plas	0.035	0.004	89.897	1	0.000	1.036
pres	-0.013	0.005	6.454	1	0.011	0.987
skin	0.001	0.007	0.008	1	0.929	1.001
insu	-0.001	0.001	1.749	1	0.186	0.999
mass	0.090	0.015	35.347	1	0.000	1.094
pedi	0.945	0.299	9.983	1	0.002	2.573
age	0.015	0.009	2.537	1	0.111	1.015

Estimates for the parameters obtained through the maximum likelihood estimation method for the final model are shown in table 3.9. Odds ratios were close to or greater than 1 for most of the variables.

Our equation can be written:

$$\mathit{logit}\,P = -8.405 + 0.123\,\mathrm{preg} + 0.035\,\mathrm{plas} + -0.013\square\mathrm{res} + 0.090\,\mathrm{mass}$$

$+\ 0.945\ \mathrm{pedi}$

Classification results of the Diabetes data set

In order to compute the correct classification rate for the Diabetes data we will use the ‘leave-one-out’ method, table 3.10 and table 3.11 illustrate the confusion matrix of LDA and MLR classification methods respectively, assigned to the dependent variable associated with the Diabetes dataset to classify the class membership of women that diabetics using ' leave-one-out’ method. The class labels of the data are 1 for diabetes and 0 otherwise.

Table (3.10): Confusion Matrix of Diabetes Data LDA Classification using leave-one-out.

Actual Population	Predicted		Sample Size
	0	1	
0	442	58	500
1	115	153	268

Table (3.11): Confusion Matrix of Diabetes Data MLR Classification using leave-one-out.

Actual Population	Predicted		Sample Size
	0	1	
0	445	55	500
1	112	156	268

For the ‘hold out’ procedures, we used 70 percent of the sample as training data and the remaining 30 percent as the validation data, the process of holdout is repeated 20 times, and the average of those 20 correct classification rate is then taken to estimate the true classification rate of LDA and MLR for Diabetes data, table 3.12 gives the overall correct classification results of each method. Our next step is to perform paired t-test, to test the significance of the performance difference between LDA and MLR, over the Diabetes dataset.

Table (3.12): Overall Accuracy for Diabetes Classification.

Method	LOOCV%	Hold Out %
LDA	77.47	77.49
MLR	78.26	78.40

To evaluate the performance of LDA and MLR for the Diabetes data, we employ the confusion matrix, which shows the actual versus predicted group membership for the two groups**.** From table 3.10, we can see that 442 of 500 women from the first group were correctly classified, and 153 of 268 women from the second group were correctly classified, so LDA succeeded to classify 595 instances from original 768 instances correctly, and the overall correct classification rate was 77.47 percent (table 3.12).

For the MLR performance, table 3.11 illustrates the confusion matrix of MLR, it can be seen that MLR succeeded to classify 601 instances from original 768 instances correctly, and the overall correct classification rate was 78.26 percent (table 3.12). According to the results shown in table 3.12, it can be seen that the classification accuracy of MLR using "leave one out" cross validation and holdout validation, was slightly better than classification accuracy of LDA. However, t-test results indicate that there are significant difference between the two models performances, and MLR performed better than LDA in the Diabetes data set. Given these classification rates, it seems that MLR is more appropriate to classify the class membership of women that diabetics. This mainly, due to violating the assumption of multivariate normality for the Diabetes data and there were no homogeneous covariance matrices.

Table (3.13): Sensitivity and specificity and AUC of logistic regression and discriminant analysis models.

Linear discriminant analysis			Logistic regression		
Sensitivity (%)	**Specificity (%)**	**AUC (%)**	**Sensitivity (%)**	**Specificity (%)**	**AUC (%)**
79.35	72.51	83.50	79.89	73.93	83.90

From table 3.13, we can see that both logistic regression and discriminant analyses gave approximately similar results.

The overall classification rate for both was good, and either can be helpful in classifying the class membership of women that diabetic. Logistic regression slightly exceeds discriminant function in the correct classification rate but when taking into account sensitivity, specificity and AUC the differences in the AUC were negligible.

4. Simulation

To compare and evaluate the performance of LDA and MLR in terms of the group and data characteristics, 7 simulated datasets with different numbers (n) of individuals are used to compare the performance of linear discriminant analysis and multinomial logistic regression. The samples are simulated from normal distributions with the same covariance matrix and different mean vectors, which are divided equally into 2 classes. These simulations are based on an R function mvrnorm for simulating from a multivariate normal distribution from R package MASS. This simulation experiment generated data with sample sizes of 50, 100, 200, 400, 500, 800 and 1000. We set the simulations where the explanatory variables are normally distributed with the same covariance matrix and different mean vectors to observe the impact of changes related to sample size. Table 4.1 shows the classification performance of LDA and MLR versus the sample size illustrates the classification performance obtained for each of the two methods.

Table (4.1): Simulation results for the effect of sample size.

Sample Sizes	LDA %	MLR %
50	82.60	82.00
100	86.00	86.00
200	83.50	83.10
400	87.30	87.90
500	85.60	85.80
800	84.60	84.40
1000	85.60	85.50

According to the simulation results for the effect of sample size shown in table 4.1, the variation in sample size has similar effect on the two methods and as the sample size increases the classification accuracy increases and the sample size significantly affects the performance of LDA and MLR, but the performance of

LDA is slightly better than MLR. However, for the large sample sizes the results of the two methods become very close and the differences between the two methods become negligible.

We can see that the effect of sample size for the performance of multinomial logistic regression could become the largest when the sample size is 400 and growing rapidly when increasing the sample size from 50 to 400, so the increase of the sample size has a significant impact but when increasing the sample size of 400, the classification accuracy rates decrease incrementally as the sample size increases and the variation in sample sizes has little impact on classification accuracy rates.

To examine the behavior and the efficiency of LDA and MLR when the differences between group means has different values, the samples are simulated from normal distributions with the same covariance matrix but the Euclidean distance between each pair of sample mean is differ from a sample to another, the numbers of observations of each group are fixed with $n_1 = n_2$=200. Table 4.2 shows the classification performance of LDA and MLR versus the distance between each pair of sample mean.

Table (4.2): Simulation results for the effect of Distance between Group Means.

D	LDA %	MLR %
0.50	60.00	60.00
1.00	69.50	69.00
1.50	74.50	74.00
2.00	80.50	81.00
3.00	94.50	94.00
4.00	98.00	99.00

According to the simulation results for the effect of distance between group means shown in table 4.2, it can be seen that the variation in values of Euclidean distance has similar effect on the two methods and as the values increases the classification accuracy increases and the distance between group means significantly affects the performance of LDA and MLR, but for low values of Euclidean distance the performance of LDA is slightly better than MLR, but as this distance increases and it takes values above 3, MLR performs better.

Five simulated samples with different numbers of categories for the discrete dependent variable are used to compare the performance of linear discriminant analysis and multinomial logistic regression to assess each method in terms of number of categories. The samples are simulated from normal distributions with the same covariance matrix and different mean vectors, which are categorized into a certain number of categories and divided equally for every category to have the same number of observations.

Table 4.3 shows the classification performance of LDA and MLR versus the numbers of categories.

Table (4.3): Simulation results for the effect of the number of categories.

No. Of Categories	LDA %	MLR %
2	86.00	86.80
3	79.30	79.60
4	81.40	81.30
5	81.70	81.10
10	74.60	74.20

According to the simulation results for the effect of the number of categories shown in table 4.3, it can be seen that as the number of categories increases, the classification accuracy decreased significantly for the performance of LDA and MLR, and different number of categories have a similar impact on the result for the two methods. But although the number of categories have a similar impact on both methods, the performance of MLR was slightly better when the number of categories was small or less than 4, but with the case of the number of categories above 4, the performance of LDA was better than the performance of MLR. To compare and evaluate the performance of LDA and MLR in terms of the presence of multicollinearity and examine the effect of correlation between explanatory variables on the performance of LDA and MLR, 6 simulated samples with two predictors have a correlation coefficients of 0.25, 0.50, 0.65, 0.75, 0.83 and 90 were used for this purpose. The samples are simulated from normal distributions with the same covariance matrix and different mean vectors, which has a sample size of 200 for each sample.

Table (4.4): Simulation Results for the Effect of the Proportion of Correlation between Explanatory Variables.

Correlation	LDA %	MLR %
0.25	86.10	86.00
0.50	85.00	85.50
0.65	85.00	84.00
0.75	83.50	84.00
0.83	81.00	83.50
0.90	83.50	83.50

According to the simulation results for the effect of correlation between explanatory variables shown in table 4.4, it can be seen that the performance of LDA and MLR differ in maximum 2%, the classification accuracy rates decrease incrementally as the values of correlation coefficient increases so the value of correlation coefficient affect the LDA and MLR performance and both methods LDA and MLR show a significant improvement in classification accuracy in the absence of multicollinearity among the explanatory variables.

5. Conclusion and recommendations

In this paper we have compared between the two different methods of classification: linear discriminant analysis (LDA) and multinomial logistic regression (MLR) using the overall classification accuracy, investigating their quality of prediction in terms of sensitivity and specificity. and examining area under the ROC curve (AUC). The performance evaluation was carried out using real data; we also conducted a simulation study to examine the group and data characteristics that may affect the performance of LDA and MLR. We concluded that the performance of both logistic regression and linear discriminant analysis on the diabetes data gave similar results. The overall classification rate for both was good, and either can be helpful in classifying the class membership of women that diabetic. Logistic regression slightly exceeds linear discriminant analysis in the correct classification rate and MLR performed better than LDA in the Diabetes data set, this due to violating the assumption of multivariate normality for the diabetes data, and there were no homogeneous covariance matrices. But when taking into account sensitivity, specificity and AUC, the differences in the AUC were negligible. Also, by the conducted simulation study, we concluded that the variation in sample size has similar effect on the two methods, and as the sample size increases the classification accuracy increases, and the sample size significantly affects the performance of LDA and MLR, and the performance of LDA is slightly better than MLR. In the case of the differences between group means has different values, we can say the variation in values of Euclidean distance has similar effect on the two methods, and as the values increases the classification accuracy increases and the distance between group means significantly affects the performance of LDA and MLR, but for low values of Euclidean distance the performance of LDA is slightly better than MLR, but as this distance increases and when it takes values above 3, MLR performs better. For the number of categories, we concluded that, as the number of categories increases, the classification accuracy decreased significantly for the performance of LDA and MLR, and different number of categories had a similar impact on the result for the two methods, and the performance of MLR was slightly better when the number of categories was small or less than 4, but with the case of the number of categories above 4, the performance of LDA was better than the performance of MLR. Also in terms of the presence of multicollinearity, we examined the effect of correlation between explanatory variables on the performance of LDA and MLR, and we concluded that the value of correlation coefficient affects the LDA and MLR performance, and the both methods LDA and MLR show a significant improvement in classification accuracy in the absence of multicollinearity among the explanatory variables.

According to the conclusion reported above we may recommend that it is important to examine the group and data characteristics that may affect the performance of LDA method, since real-world data are usually contaminated, the MLR model seems suitable to be used for classification problem in cases for which the variables do not have multivariate normal distributions, nor equal variance within each group, for linear discriminant analysis and multinomial logistic regression, a large sample size is required in order to achieve its maximum prediction accuracy, and the linear discriminant seems suitable to be used for classification when the dependent variable has more than four groups/categories, but with the case of the number of categories less than 4, the performance of MLR is better than the performance of LDA.

References

- Allison, Paul D. (1999). *Logistic regression using the SAS system: theory and application*. Cary, N.C.: SAS institute Inc., USA.
- Arlot Sylvain & Celisse Alain (2010). *A survey of cross-validation procedures for model selection*. Statistics surveys: Vol. 4, pp. 40–79.
- Dinesh R. Pai (2009). *Determining the efficacy of mathematical programming approaches for multi-group classification*. Rutgers University electronic theses and dissertations, pp. 15.

- Hamid & Hashibah (2010). *A new approach for classifying large number of mixed Variables*. World academy of science, Engineering and Technology.
- Harrell, F.E. (1997). *Translating probability models into clinical decisions*. Lecture notes.
- Hosmer, David & W. Lemeshow, Stanley (2000). *Applied logistic regression*. 2nd edition, New York: Wiley.
- Hossain M, Wright S, & Petersen L. (2002). *A Comparing performance of multinomial logistic regression and discriminant analysis for monitoring access to care for acute myocardial infarction*. Journal of Clinical Epidemiology, Volume: 55, Issue: 4, Pages: 400-406.
- Hussein, Mohamed (2010). *The linear classification functions versus the logistic discriminant function in a three population situation with a mixture of continuous and categorical variables*. J. Edict. Psychol., 22:45–55.
- Kiang, M. (2003). *A comparative assessment of classification methods*. Decision Support Systems, 35, 441- 454.
- Kotsiantis S. B. (2007). *Supervised machine learning: a review of classification techniques. Department of computer science and technology*, University of Peloponnese, Greece, Artificial intelligence review, 26(3):159-190, Springer.
- Meyers, L., G. Gamst, & A. Guarino (2006). *Applied multivariate research: design and interpretation*. Newbury Park, CA: Sage publications, London.
- Montgomery Me, White Me, & Martin Sw (1987). *A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows*. Canadian Journal of Veterinary Research, 51(4):495–498.

- Pohar M., Blas M., & Turk S. (2004). *Comparison of logistic regression and linear discriminant analysis: a simulation study*. Metodoloski Zvezki, vol. 1, no. 1, pp. 143–161.
- Press, S. J. & Wilson, S. (1978). *Choosing between logistic regression and discriminant analysis*. Journal of the American Statistical Association, 73, 699-705.
- Rencher Alvin C. (2002), *Methods of multivariate analysis*, 2nd edition, Brigham Young University, John Wiley & sons, Inc. publication.
- Tansey, R., M. White, & R. Long (1996). *A comparison of log linear modeling and logistic in management research*. Journal of management, 22: 339-358.
- Tianshu Pan (2008). *Using the multivariate multilevel logistic regression model to detect DIF: A comparison with HGLM and logistic regression DIF detection methods*. a PhD dissertation submitted to Michigan State University.
- Timm Neil H. (2002). *Applied multivariate analysis*. Springer.
- Tom Fawcett (2004). *ROC Graphs*: notes and practical considerations for researchers.
- Wim Van den Noortgata and Paul De Boeck (2005). *Assessing and explaining differential item functioning logistic mixed models*. Journal of Education and Behavioral Statistics, Vol. 30, No. 4, pp.443-464.
- Yovhane L. Metcalfe (2012). *A logistic regression and discriminant function analysis of enrollment characteristics of student veterans with and without disabilities*. a PhD dissertation submitted to Virginia Commonwealth University.

Web (1): http://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes