

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(3): 23-36
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 26-12-2024
Accepted: 29-01-2025

Ahmed Shamar Yadgar
Department of Statistics,
Kirkuk, Collage of
Administration and Economics,
University of Kirkuk, Iraq

Zana Najim Abdullah
Department of Statistics,
Kirkuk, Collage of
Administration and Economics,
University of Kirkuk, Iraq

Corresponding Author:
Ahmed Shamar Yadgar
Department of Statistics,
Kirkuk, Collage of
Administration and Economics,
University of Kirkuk, Iraq

Comparison of linear discriminant analysis and logistic regression for classification diabetes data in Kirkuk Governate

Ahmed Shamar Yadgar and Zana Najim Abdullah

DOI: <https://www.doi.org/10.22271/math.2025.v10.i3a.2001>

Abstract

Diabetes is a chronic metabolic disorder that has yet to be addressed. It's a collection of metabolic disorders that impair protein, fat, and carbohydrate utilisation. These disorders result from a lack of insulin synthesis or a poor tissue response. The number of these illnesses is growing internationally as the prevalence of obesity and the quantity of physical activity decreases, resulting in the development of diabetes. As a consequence, the World Health Organisation has identified obesity as a major concern. In this discipline, categorisation is the key focus of study, and an increasing number of scientists use techniques based on various approaches. This research uses logistic regression (LR), a statistical approach for interpreting complicated data combinations. Discriminant analysis is a technique of observation that categorises objects based on the degree of their dependent variable and then generates a formula that may be used to add additional information. The findings demonstrate that logistic regression of medication, blood pressure, age, gender, waist circumference, and LDL levels was used in the discriminant analysis for diabetes categorisation using a simple logistic regression model. The ROC curve and AUC value of 0.61 show that the logistic regression model can identify between diabetic and non-diabetic individuals.

Keywords: Discriminant analysis, logistic regression, diabetes data

1. Introduction

This study's main goal was to compare and evaluate the efficacy of LR and LDA, particularly in relation to the assumption of normalcy. Both LDA and LR are well-known statistical techniques that analyse data with categorical variables that are directly related to the result and create linear classification models. They have distinct core values, however. LDA usually assumes that the unknown variance-covariance matrix is the same for each category of the variable's grouping and employs predictors that are normally distributed. In general, LDA works well if these theories are true. However, if the data contain inconsistent covariance structures or do not adhere to the multivariate normality criterion, the effectiveness of LDA may be compromised. However, since LR does not rely on certain assumptions about the distribution of the data, it is regarded as a potent substitute for LDA. When handling data that is normally not dispersed in a traditional manner, LR is very helpful. Nonetheless, the computing time is often longer.

Using actual datasets, the research seeks to provide useful insights into the efficacy of different approaches, particularly with reference to the normalcy assumption. It will also compare the time required to compute LR and LDA.

2. Mythology

2.1 Linear Discriminant Analysis

To identify the characteristics that distinguish one category from another and to develop classification models that forecast whether incoming data will fall into one of the categories, linear discriminant analysis may be used. It is assumed by linear discriminant analysis that the variables used to characterise each group have a comparable covariance matrix and are

normally distributed. Two classes are distinguished via the simplest linear discriminant analysis. A linear function that goes through each class' centroid may be used to distinguish between the two classes. A more thorough discussion of linear discrimination analysis may be found in Kachigan (1991). The distribution of $X|Y$ is assumed to be a multivariate normal distribution in the conventional linear discriminant analysis paradigm, with a shared covariance matrix Σ and a mean vector μ_y . We may show that we allocate x to group 1 in the following way by using algebraic equations.:

$$P(1|x) = \frac{1}{1 + (e^{\alpha + \beta x})^{-1}}$$

The values of the coefficients of α and β are

$$\beta = (\mu_1 - \mu_0)^T \Sigma^{-1}$$

$$\alpha = -\log \frac{\pi_1}{\pi_0} + \frac{1}{2} (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 + \mu_0)$$

The likelihood of belonging to group 1 is higher than the likelihood of belonging to group 0. In actuality, the variables π_1 , π_0 , μ_1 , μ_0 , and Σ are unknown.

2.2 Logistic Regression

Finding the most accurate and useful model for the connection between a collection of independent factors (predictors or variables that have an impact on) and a dependent variable (end) is the goal of logistic regression (LR). This approach is regarded as reliable, adaptable, easy to use, and yields useful data. LR does not need the dependent variables' distribution to be normal, in contrast to Linear Discriminant Analysis (LDA). The opposite of what is often assumed is that both LDA and LR may be used when there are more than two classes. To be clear, we concentrate on the case when there are only two potential values for the outcome variable (Y). The following is a mathematical explanation of the LR model:

$$P(Y_i = 1|X_i) = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}$$

Y_i , a Bernoulli random variable, is unaffected by other variables. To find the model's parameters, the maximum likelihood technique is used.

2.3 Comparison Criteria

Two measures are used in this study to evaluate the efficacy of LR and LDA. The criteria that were used to evaluate the two approaches are described.

To evaluate the efficacy of the classification model, the first statistic is the percentage of correct classification, which is computed using the formula below:

The percentage of accurate classification

$$= \frac{\text{Number of observations being classified correctly in a particular group}}{\text{Total number of observations in a particular group}} \times 100\%$$

The classification error is extremely inaccurate and lacks any significant statistical significance, despite being the most

simple and most commonly used metric for comparing methods.

The second-best metric, the B-index, is a supplementary metric that calculates the degree to which the predictions are correct. This is beneficial because it demonstrates the value of the model in recognizing different groups and having accurate predictions regarding them. The B-index is the average of the squared differences between the estimated and actual values. A B-index of 1 (between 0 and 1) is indicative of accurate predictions. The B-index formula is as follows:

$$B = 1 - \sum_{i=1}^n \frac{(P_i - Y_i)^2}{n}$$

In this case, Y_i is the actual membership number (1 or 0), P_i is the chance of being placed in group I , and n is the size of the two populations. A B-index of 1 indicates a flawless forecast, and the value falls between 0 and 1 when two populations of equivalent size are subjected to random forecasts.

3. Application

3.1 Data Descriptions

Logistic regression and linear discriminant analysis were evaluated using the statistical software R, this was done in order to demonstrate the utilization of diabetes information for a case study. The data was obtained at Azadi Hospital in the Kirkuk Province. The population was selected from those who had been tested for diabetes according to the recommendations of the World Health Organization (WHO), and were at least 21 years old. The job of classification was to determine if the patient had diabetes or not. The World Health Organization's guidelines define 68 of the 152 individuals aged 21 years or greater with 5 predictors that were examined for diabetes (grade 1), while all other diseases were considered 0. Table 1 describes the procedure.

Table 1: The explanation of the diabetes database

S. N	Factors	Description	Categorical
1	Age	Age in years	Continuous
2	Gender	Gender of the patients	Male, Female
3	BMI	Body mass index (kg/m ²)	Continuous
4	Diastolic BP	Diastolic blood pressure (mm Hg)	Continuous
5	Systolic BP	Systolic blood pressure (mm Hg)	Continuous

3.2 Results

Table 2: Wilks' Lambda Table

Wilks' Lambda	Chi-square	Df	Sig.
0.774	281.539	6	0.000

In order to have the most effective model based on DA, we must analyze the data. Since a significant amount of 0.000 was observed in this instance, a discriminant analysis can be conducted using a progression of steps that will lead to the conclusion that the correlation function is effective in categorizing individuals as belonging to a particular population.

Wilk's lambda F is documented in Table (3). It's obvious that the stepwise strategy in this example is to decrease the value of Wilk's lambda by each iteration. Additionally, it's vital to recognize that tolerance is the portion of the variance that is not explained by the other variables in the equation.

Table 3: Factors that are examined in a stepwise manner

Variable	Tolerance	F to remove	Wilks' Lambda
Age in years	0.332	0.693	0.176
Gender of the patients'	0.443	0.558	0.161
Body mass index (kg/m ²)	0.493	0.529	0.219
Diastolic blood pressure (mm Hg)	0.533	0.431	0.121
Systolic blood pressure (mm Hg)	0.178	0.822	0.223

Variables that have a low capacity to withstand change will lead to problems with estimation and will provide limited information about the model. The F removal values are beneficial for demonstrating the effects of a component being removed from the current model (if the other components are still the same).

Table 4: The probability of the test

Model	Model fitting criteria	Likelihood ratio tests	Df	Sig.
	-2 log likelihood	Chi-squared		
Intercept only	682.430	210.089	5	0.000
Final	491.783			

The log-likelihood of the basic model (-2) with only the intercept is 682.430, according to the data in Table 4; this number drops to 491.783 when the individual variable is included. The study's chi-square probability is 0.00, which is not as significant as 0.05. There is no evidence to support the premise that the model with the independent variable is

identical to the model without it. The evidence supports the relationship between the independent and dependent variables.

Table 5: The estimation of the logistic regression model

Variable	Coeff.	SE	Wald	Df	Sig.	Exp(B)
Intercept	2.023	0.362	74.401	1	0.000	
Age	0.103	0.009	10.302	1	0.020	1.191
Gender	0.083	0.021	22.542	1	0.004	1.072
BMI	0.019	0.065	27.644	1	0.107	0.992
Diastolic BP	0.201	0.073	12.883	1	0.055	0.986
Systolic BP	0.054	0.002	5.931	1	0.036	1.108

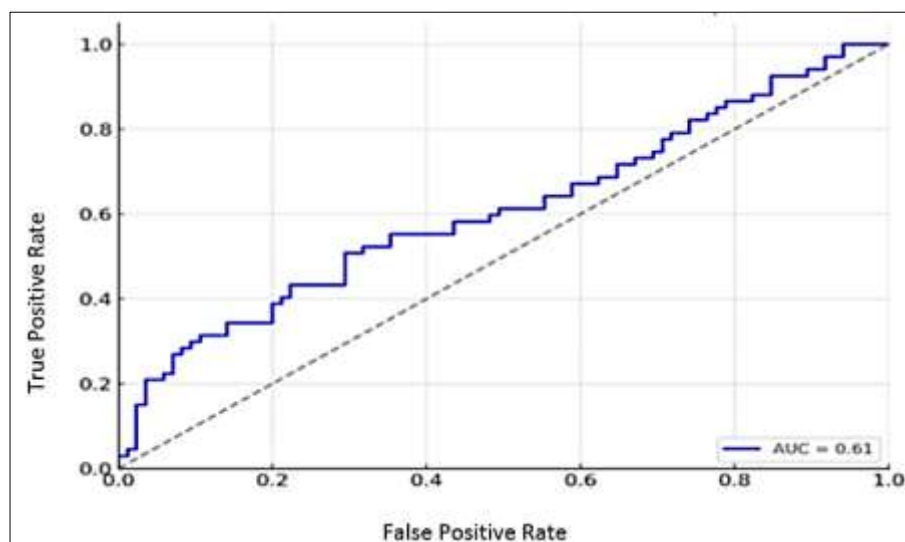
Based on the aforementioned results, Table (6) demonstrates the estimated logistic regression parameters for the final model's parameters that were obtained using the maximum likelihood method. The probability of most variables is greater than or equal to 1.

Table 6: The classification performance of (LR, LDA)

Model	Percentage of correct classification	B Index
Logistic Regression	88%	0.92
Discriminant Analysis	86%	0.83

This research discusses the classification results of the discriminant analysis model and the logistic regression model (Table 6). Additionally, the study showed that logistic regression had a higher accurate classification rate than

discrimination analysis. Furthermore, the B value of LDA was only 0.83, whereas the B value of LR was 0.92. In other words, LR produced a flawless classification and was more successful in this inquiry than LDA.

**Fig 1:** The ROC curve for the classification of diabetes.

This research discusses the classification results of the discriminant analysis model and the logistic regression model (Table 6). Additionally, the study showed that logistic regression had a higher accurate classification rate than discrimination analysis. Furthermore, the B value of LDA was only 0.83, whereas the B value of LR was 0.92. In other

words, LR produced a flawless classification and was more successful in this inquiry than LDA.

4. Conclusion

In this work, we compared the overall accuracy of two classification techniques, LR and LDA. Compared to linear

discriminant analysis, which was more successful at categorising items as either C or D, logistic regression was more successful at classifying things as either A or B. This is because both approach is impacted by a uniform covariance matrix, and the data on diabetes do not conform to the premise of a multivariate normal distribution. Furthermore, the performance of both LDA and MLR is significantly impacted by the difference between the average values of the groups, and the classification accuracy rises as the value of grows. In relation to the number of classes, we found that as the number of classes rises, the classification accuracy of LDA and LR dramatically declines. The outcomes of both approaches are similarly impacted by the difference number of categories, with LR having a little edge due to its smaller category count. If the variables are distributed in several dimensions or in a regular manner throughout the group, the LR model seems to be suitable for the classification problem. The logistic regression approach outperformed the discriminant analysis technique when our diabetes database was compared using multinomial logistic regression and linear discrimination. Given the prevalence of fabricated real data, it is important to take into account both individual and group features that might impair the LDA method's efficacy. The logistic regression model was much more successful in distinguishing between patients with and without diabetes, according to the ROC curve and AUC.

Therefore, our findings show that in order to diagnose chronic illnesses like diabetes early, it is important to explore several methods for classifying support systems. To evaluate and enhance the effectiveness of these techniques in actual clinical practice, further investigation and clinical trials are required.

5. Reference

1. Adebisi MO, Arowolo MO, Mshelia MD, Olugbara OO. A linear discriminant analysis and classification model for breast cancer diagnosis. *Applied Sciences*. 2022;12(22):11455.
2. Ali AAR, Hawamdeh OM. The factors affecting eye patients (Cataract) in Jordan by using the logistic regression model. *Modern Applied Science*. 2017, 11(8).
3. Araveeporn A. Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity. *International Journal of Mathematics and Mathematical Sciences*. 2022;2022(1):7829795.
4. El-Habil A, El-Jazzar M. A comparative study between linear discriminant analysis and multinomial logistic regression. *An-Najah University Journal for Research-B (Humanities)*. 2013;28(6):1525-1548.
5. Mahmood SH, Khider HH. Comparison of linear discriminant analysis and artificial neural networks for stroke patients classification. *University of Kirkuk Journal for Administrative and Economic Science*. 2023, 13(2).
6. Matrood DO, Al-Quraishi OAK. The use of linear discriminant analysis to determine the variables affecting cancer diseases. *University of Kirkuk Journal for Administrative and Economic Science*. 2023, 13(1).
7. Nawaiseh AK, Abbod MF. Financial statement audit utilizing Naïve Bayes networks, decision trees, linear discriminant analysis, and logistic regression. In: *The Importance of New Technologies and Entrepreneurship in Business Development: In the Context of Economic Diversity in Developing Countries: The Impact of New Technologies and Entrepreneurship on Business*

Development. Springer International Publishing. 2021:1305-1320.

8. Olufemi I, Obunadike C, Adefabi A, Abimbola D. Application of logistic regression model in prediction of early diabetes across United States. *ResearchGate*. 2023 Jan.
9. Pohar M, Blas M, Turk S. Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloski Zvezki*. 2004;1(1):143.
10. Subba B, Biswas S, Karmakar S. Intrusion detection systems using linear discriminant analysis and logistic regression. In: *2015 Annual IEEE India Conference (INDICON)*. IEEE. 2015 Dec:1-6.
11. Zhao H, Lai Z, Leung H, Zhang X. Linear discriminant analysis. In: *Feature Learning and Understanding: Algorithms and Applications*. 2020:71-85.