Using Logistic Regression Analysis and Linear Discriminant Analysis to identify the risk factors of Diabetes

(PP 248 - 268)

ID No. 2115

https://doi.org/10.21271/zjhs.22.6.17

Nazeera Sedeeq kareem Barznji
College: Administration and Economic/Statistics
nazeera.barznji@gmail.com

Received: 10/03/2018
Accepted: 16/08/2018
Published:27/12/2018

Abstract

Many medical studies point out that there is a close relationship between the diagnostic aspects of a disease and some statistical analysis like Logistic Regression Analysis (LRA) and Linear Discriminant Analysis (LDA), both of them are two widely used multivariate statistical methods for data analysis, they are used in order to prediction. In this paper both analyses were discussed and implemented on data with sample size 250 Diabetes patients it collected from Erbil Layla Qassem Center for Diabetes. The data contained (8) variables, one of them is dependent variable that represents the presence or absence of Diabetes, and the other 7 variables are predictors (Independent variables), they are taken in the model in which they represent risk factors of diabetes disease like: [High Blood Pressure (Hypertension), Family History, Body Mass Index (BMI)-Obesity, Diet (Nutrition), High Lipid in Blood, Physical Activity and Age].

The paper aims to the comparison between Logistic Regression Analysis and Linear Discriminant Analysis based on several measures of predictive accuracy to choose the best statistical model for identifying the risk factors of diabetes. This paper contains two parts, Theoretical aspects and Practical aspects .The results of every test was done with both analyses (Logistic Regression Analysis and Linear Discriminant Analysis),reflects to a high ratio of prediction of Logistic Regression Analysis and the result of area under the ROC Curve of all variables, which is used to compare prediction powers of the models, emphasized on that the Logistic Regression Analysis has the best prediction of risk factors of diabetes and it has the appropriate model so Logistic Regression Analysis has emerged as a robust alternative to Linear Discriminant Analysis. By logistic regression the ranking of risk factors on diabetes is as follows>

1-Family History 2-(BMI)(Obesity rate),3-HighLipid in Blood,4-Physical Activity ,5-Hypertension, 6-Diet (Nutrition)=2.033, but (age) it is not represents the risk factor

Keywords: logistic regression, Maximum Likelihood.

1. Introduction

The incidence of diabetes has doubled in the last ten years in the world wide . about 200 million people are infected and about six percent increase in the annual prevalence of diabetes in the world with the proportion in Kurdistan is 10.2% of the population are infected with this disease is ranked 30th in the world.

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin o

when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. There are different types of diabetes that usually are distinguished at diagnosis [11] [18] [19] [22]

Type 1 diabetes (it is known as an insulin-dependent, juvenile or childhood-onset) is characterized by deficient insulin production and requires daily administration of insulin. The cause of type 1 diabetes is not known and it is not preventable with current knowledge.

Type 2 diabetes (Formerly called non-insulin-dependent, or adult-onset) results from the body's ineffective use of insulin. Type 2 diabetes comprises the majority of people with diabetes around the world, and is largely the result of excess body weight and physical inactivity.

According to the World Health Organization (WHO) "Global report on diabetes" The number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014.The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014 .Diabetes prevalence has been rising more rapidly in middle- and low-income countries. In 2015, an estimated 1.6 million deaths were directly caused by diabetes.WHO (World Health Organization) projects that diabetes will be the seventh leading cause of death in 2030 . In this paper, the two classification methods namely linear discriminant analysis (LDA) and logistic regression analysis (LRA) both of them are widely used in multivariate statistical methods for data analysis with categorical outcome variables. they can construct linear classification model which creates a linear boundary between two groups. Logistic Regression is a linear regression in terms of the relationship between the (dependent Variable) and a set of independent variables or explanatory variables and Linear Discriminant analysis is a statistical analysis usually is useful in determining whether a set of variables is effective in predicting category membership to predict both methods are found to be different in their basic assumptions. LDA makes a few assumptions such as the explanatory or predictor variables must be normally distributed. LRA appears as a robust alternative to LDA as it does not need any underlying assumption made on the distribution of the data. Hence, LRA has always been suggested as the first choice to carry out data classification especially for a situation where the data is not normally distributed. Nevertheless, computing time of LRA is much longer than the time taken by LDA, making it a less desirable alternative to LDA.

2. Objectives of the study

1-Comparison between the LRA and LDA is based on several measure of predictive accuracy to choose the best statistical model.

2- Identify risk factors of diabetes by each method (LRA and LDA)

3.Methods and Materials:

The research methodology will be divided into two parts:

1. Part I : The Theoretical aspects of the methods used in the analysis, Logistic Regression (LRA) and Linear Discriminant Analysis (LDA).
2. Part II: The Practical aspects study for Explaining the logistic regression (LR) and discriminate model (LDA) and focusing on the characteristics and how to estimate the parameters.

4. Theoretical aspects

4.1 Binary Logistic Regression

In statistics, Logistic Regression Analysis ,is a model where the dependent variable is categorical . This article covers the case of a binary dependent variable that is, where the output can take only two values, "0" and "1", which represent outcomes such as(presence /absences) of Diabetic . Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor .Binary logistic regression has other application of combining the independent variables to estimate the probability that a particular event will occur a subject will be a member of one of the groups defined by the dichotomous dependent variable. The variety or value produced by binary logistic regression is a probability value between 0 and 1. If the probability for group

membership in the modeled category is above some cut point (usually 0.5), the subject is predicted to be a member of the modeled group. If the probability is below the cut point, the subject is predicted to be a member of the other group.

4.1.1 The Model of Binary Logistic Regression [1] [12] [15]

The Logistic Regression model indirectly models the response variable based on probabilities associated with the values of the dependent variable Y. We will use P(x) to represent the probability that Y =1, which is the presence of Diabetic. Similarly, we will define 1-P(x) to be the probability that Y =0, which is absence of Diabetic. These probabilities are written in the following form:

$$P(x) = P(Y = 1|X_1, X_2, X_3, \cdots, X_n) \quad \ldots\ldots\ldots\ldots (2)$$

$$1 - P(x) = P(Y = 0|X_1, X_2, X_3, \cdots, X_n) \quad \ldots\ldots\ldots\ldots (3)$$

The log distribution (or logistic transformation of (p) is also called the logit of g or logit (g(x))which is the log (to base e) of the odds ratio or likelihood ratio that the dependent variable is 1. In symbols it is defined as:

$$Logit(g(x)) = \log_e \left[P(x)/\left(1 - P(x)\right)\right] = \ln\left[P(x) / (1 - P(x))\right] \quad \ldots\ldots\ldots\ldots (4)$$

Where the range of $P(x)$ from 0 to 1, but $Logit(g(x))$ scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero).

Formula below shows the relationship between the usual regression equation $Y = \beta_0 + \beta X$, which is a straight line formula, and the logistic regression equation. The form of the logistic regression equation is thus rewritten as:

$$Logit\, g(x) = \log_e \left[P(x)/\left(1 - P(x)\right)\right] = \ln\left[P(x)/\left(1 - P(x)\right)\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n x_n \quad \ldots\ldots\ldots\ldots (5)$$

This looks just like a linear regression and although logistic regression always finds a 'best fitting' equation, just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficient

A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression is different from those used in linear regression. P can be calculated with the following formula:

$$\ln\left[P(x)/\left(1 - P(x)\right)\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n x_n$$

$$\frac{P(x)}{\left(1 - P(x)\right)} = \left(e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n x_n}\right)$$

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n)}} \quad \ldots\ldots\ldots\ldots (6)$$

Where:

$P(x)$ = the probability that a case is in a particular category,

e = the base of natural logarithms (approx. 2.72),

β_0 = the constant of the equation

$\beta_1, \beta_2, \beta_3, \cdots, \beta_n$ the coefficient of the predictor variables

4.1.2 Estimate the Parameters of Logistic Regression Model [1] [2]

By Maximum Likelihood Estimation method (MLE) will be Estimate the Parameters, the mathematical formula for the Likelihood function in binary data, is given as the following to fit a set of data in order to estimate the parameters β_0 and β_1 In logistic regression the method Maximum likelihood will provide values of β_0 and β_1 which maximize the probability of obtaining the data set. It requires iterative computing and is easily done with most computer software .We use the likelihood function to estimate the probability of observing the data, given the unknown parameters (β_0 and β_1). "Likelihood" is a probability, specifically the

probability that the observed values of the dependent variable may be predicted from the observed values of the independent variables. Like any probability, the likelihood varies from 0 to 1. Practically, it is easier to work with the logarithm of the likelihood function. This function is known as the log-likelihood, and will be used for inference testing when comparing several models. The log likelihood varies from 0 to minus infinity (it is negative because the natural log of any number less than 1 is negative).The log likelihood is defined as:

$$L = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i} \quad \ldots\ldots\ldots\ldots (7)$$

For estimation, we will work with the log-likelihood

$$\log(L) = \sum_{i=1}^{n} y_i \log p + \left(\left(n - \sum_{i=1}^{n} y_i\right)\log(1-p)\right)$$

$$U(P) = \frac{\partial L}{\partial P} = \sum_{i=1}^{n} y_i / p - \left(\left(n - \sum_{i=1}^{n} y_i\right)/(1-p)\right)$$

and is referred to as the score function. To calculate the MLE of p, we set the score function, $U(p)$ equal to 0 and solve for p. In this case, we get an MLE of p is

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \ldots\ldots\ldots\ldots (8)$$

4.1.3 For analyzing by

1- Logistic Regression Analysis it needs at least two steps [1] the process is inherently stepwise for forming and testing nested hierarchical models. It needs following tests analysis

4.1.3.1 In Step one: only the constant for is provided (the constant only included).

1-The classification table tells % cases correctly classified by the model. -The first step is to compute and enter just the constant. it indicate to the variables not in the prediction equation by Score test

2-Wald test [5] [6] [7]

Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution

$$W_i = \frac{\beta_j^2}{SE_{\beta_i}^2} \quad \ldots\ldots\ldots\ldots (9)$$

For large n, $W_i \sim \chi_1^2$ with 1 degree of freedom

$$\sqrt{W_i} \sim N(0,1)$$

Wald test, where β parameters estimation and (S.E.) is standard error estimation

Wald Chi-Square testing the null that the β coefficient = 0 (the alternate hypothesis is that it does not β coefficient = 0).

3-Score test [1] [24]

If the MLE equals the hypothesized value, p_0, then p_0 would maximize the likelihood and $U(p_0) = 0$ The score statistic measures how far from zero the score function is when evaluated at the null hypothesis. The test statistic for the binary outcome example is

$S = U(p_0)^2/I(p_0)$, and $S \sim \chi_1^2$ with 1 degree of freedom

4.1.3.2 In step two needs the following tests or analysis: this step tests the contribution of all the variables entered

1-Omnibus Tests in Logistic Regression [10]

The *Omnibus Tests of Model Coefficients* is used to check that the new model (with explanatory variables included) is an improvement over the baseline model. It uses chi-square

tests to see if there is a significant difference between the Log-likelihoods (specifically the -2LLs) of the baseline model and the new model. If the new model has a significantly reduced -2LL compared to the baseline then it suggests that the new model is explaining more of the variance in the outcome and is an improvement

$$P(x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_nx_n}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_nx_n}}$$

$$P(x) = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_nx_n)}}$$

So the model tested can be defined by:

$$g(x) = \ln\left[P(x)/\left(1-P(x)\right)\right] = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_nx_n$$

The omnibus test relates to the hypotheses

$$\mathbf{H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_n}$$

H_1: at least one pair $\beta_j \neq \beta_j'$

The omnibus test, among the other parts of the logistic regression procedure, is a likelihood-ratio test based on the maximum likelihood method .So logistic regression uses the maximum likelihood procedure to estimate the coefficients that maximize the likelihood of the regression coefficients given the predictors and criterion

2- Likelihood ratio tests [10] **[16]**

The likelihood ratio test for overall significance of the beta's coefficients for the independent variables in the model is used.

under the null hypothesis that the beta's coefficients for the covariates in the model are equal to zero.

The likelihood ratio test (LRT) statistic is the ratio of the likelihood at the hypothesized parameter values to the likelihood of the data at the MLE(s).

$$LR = -2\ln\left[\frac{\text{Likelihood without the variable}}{\text{Likelihood with the variable}}\right] = -2\ln\left(\frac{\text{L at } H_0}{\text{L at MLE(s)}}\right) = -2lH_0 + 2l(\text{MLE}) \quad \ldots\ldots\ldots\ldots (10)$$

For large n, $LR \sim \chi^2$ with degrees of freedom equal to the number of parameters being estimated.

For the binary outcome discussed above, if the hypothesis is

$H_0: p = p_0$ Vs: $H_A: p \neq p_0$, then

$$l(H_0) = \sum_{i=1}^{n} y_i \log p_0 + \left(n - \sum_{i=1}^{n} y_i\right)\log(1-p_0)$$

$$l(\text{MLE}) = \sum_{i=1}^{n} y_i \log \hat{p} + \left(\left(n - \sum_{i=1}^{n} y_i\right)\log(1-\hat{p})\right)$$

and the LRT statistic is

$LRT = -2[\sum_{i=1}^{n} y_i \log(p_0/\hat{p}) + (n - \sum_{i=1}^{n} y_i)\log\{(1-p_0)/(1-\hat{p})\}]$ ..... (11)

Where $LR \sim \chi_1^2$

3-Test by R^2

a)Cox and Snell R^2_{CS} or [R^2 Cox & snell] [3]

In this model in instead of the coefficient of determination by (R^2Cox & snell) ($0 \leq R^2$Cox & snell < 1)

R^2Cox & snell =

$$= 1 - \left[\frac{\text{Maximum likelihood function in the case the model being fitted}}{\text{Maximum likelihood function in the case of the the null model}}\right]^{2/n} \quad \ldots\ldots\ldots\ldots (12)$$

$$= 1 - \left[\frac{L_M}{L_0}\right]^{2/n} = 1 - e^{\left(2\frac{ln(L_M)-ln(L_0)}{n}\right)} \quad \ldots\ldots\ldots\ldots (13)$$

L_M: Maximum likelihood function in the case the model being fitted
L_0: Maximum likelihood function in the case of the null model

b) R^2 Nagelkerke or R^2_N [14] [15]

Can be calculated as

$$R^2 Nagelkerke = \frac{R^2 Cox\ \&\ snell}{1-L_0^{(2/n)}} \quad \ldots\ldots\ldots\ldots (14)$$

L_0: Maximum likelihood function in the case of the null model
n: Sample Size

4-Hosmer and Lemeshow [8] [9] [12]

The Hosmer–Lemeshow test uses a test statistic that asymptotically follows a distribution to assess whether or not the observed event rates match expected event rates in subgroups of the model population. This test is considered to be obsolete by some statisticians because of its dependence on arbitrary binning of predicted probabilities and relative low power.

It developed a goodness-of-fit test for logistic regression models with binary responses. They proposed grouping based on the value of the estimated probabilities. This test is obtained by calculating the Pearson chi-square statistic from the 2×2 table of observed and expected frequencies, where g is the number of groups. The statistic is written this test is used to determine whether well-fitting model or not, through the difference between observed values and expected values and it model is distributed chi- square distribution., and used to test the following hypothesis.

$\mathbf{H_0}$: There is no significant difference between observed values and expected values.
$\mathbf{H_1}$: There is a significant difference between observed values and expected values.

Then if (Hosmer & Lemeshow) static is > (0.05, then the model representatives will be a good model.

$$HL\ \chi^2_{k-2} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{N_i \pi_i (1-\pi_i)} \quad \ldots\ldots\ldots\ldots (15)$$

Where; -

HL :is referred to as the Hosmer-Lemeshow test statistic, , which is approximately distributed as a chi-square with $k-2$ degrees of freedom

N_i : Is the number of observation in the ith group.

O_i: Is the number of event outcomes in the ith group?

π_i: Is the average estimated probability of an event outcome for the ith group?

5-Odds and Odds Ratio (OR) [13] [17] **[21]**

Definition of the odds

The odds of the dependent variable equaling a case (given some linear combination (x) of the predictors) is equivalent to the exponential function of the linear regression expression. This illustrates how the logit serves as a link function between the probability and the linear regression expression. Given that the logit ranges between$(-\infty)$ negative and $(+\infty)$positive infinity, it provides an adequate criterion upon which to conduct linear regression and the logit is easily converted back into the odds

Consider first the case of a single binary predictor, where

$$x = \begin{cases} 1 & \text{if exposed to factor} \\ 0 & \text{if not exposed to factor} \end{cases}$$

and

$$y = \begin{cases} 1 & \text{if develops disease} \\ 0 & \text{if dose not develop disease} \end{cases}$$

Recall the logistic model: $g(x)$ is the probability of disease for a given value of x, and

$$\text{Logit } g(x) = \log_e \left[P(x)/(1 - P(x))\right] = \ln \left[P(x)/(1 - P(x))\right] = \beta_0 + \beta_1 x$$

$$\text{Logit } g(x) = \beta_0 + \beta_1 x$$

Then for $x = 0$ (not exposed),

$$\text{Logit } g(x) = \text{Logit } g(0) == \beta_0 + \beta_1(0) = \beta_0 \quad \ldots\ldots\ldots\ldots \quad (16)$$

For $x = 1$ (exposed),

$$\text{Logit } g(x) = \text{Logit } g(1) == \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \quad \ldots\ldots\ldots\ldots \quad (17)$$

$$\text{Logit } g(x) = \beta_0 + \beta_1 X_1$$

$$\text{Odds} = P(x)/(1 - P(x))$$

odds of not exposed to factor = $P(0)/(1 - P(0))$

odds of exposed to factor = $P(1)/(1 - P(1))\ P(1)/(1 - P(1))$

Odds Ratio (OR)

The odds ratio is a measure of association for(2 X 2) contingency table

Results can be summarized in a simple(2 X 2) contingency table as

	exposed	
disease	1	0
1	a	b
0	c	d

Where:

$$\textbf{Odds Ratio} = \widehat{OR} = \frac{ad}{bc} = \frac{\text{odds of presence of disease}}{\text{odds of absenceof disease}} = \frac{P(1)/(1-P(1))}{P(0)/(1-P(0))} \quad \ldots\ldots\ldots\ldots \quad (18)$$

$$\beta_1 = \text{Logit } g(1) - \text{Logit } g(0) = \log_e \left[\frac{P(1)}{(1-P(1))}\right] - \log_e \left[\frac{P(0)}{(1-P(0))}\right] =$$

$$\log_e \frac{P(1)/(1-P(1))}{P(0)/(1-P(0))} = \log_e (OR) \quad \ldots\ldots\ldots\ldots \quad (19)$$

The regression coefficient in the population model is the log(OR), hence the OR is obtained by exponentiation β_1

$$e^{\beta_1} = e^{\log_e (OR)} = OR \quad \ldots\ldots\ldots\ldots \quad (20)$$

Or Exp(β) (taking the β value by calculating the inverse natural log of(β) indicates odds ratio: the probability of an event occurring, divided by the probability of the event not occurring. An Exp (β) value over 1.0 signifies that the independent variable increases the odds of the

dependent variable occurring. An Exp (β) under 1.0 signifies that the independent variable decreases the odds of the dependent variable occurring, depending on the decoding that mentioned on the variables details before. A negative β coefficient will result in an Exp(β) less than 1.0, and a positive β coefficient will result in an Exp(B) greater than 1.0. The statistical significance of each β

4. 2 Second type of Analysis :

Linear Discriminant Analysis (LDA) also known as Discriminant Analysis (DA) or Two-group discriminant analysis. [4]

is a technique for analyzing data when the simplest type of LDA is two group LDA which the dependent variable has two groups. In this case, a linear discriminant function (LDF) that passes through the means of the two groups can be used to discriminate subjects between the two groups . Two-group LDA is a linear combination of the two or more independent variables that discriminate best between a priori defined groups. Discrimination is achieved by setting weights for each independent variable to maximize the between-group variance to the within group variance

4. 2.1The Assumptions of Liner Discriminant Analysis

The analysis is quite sensitive to outliers and the size of the smallest group must be larger than the number of predictor variables.

1-Multivariate normality: Independent variables are normal for each level of the grouping variable.

2-Homogeneity of variance/covariance (Homoscedasticity): Variances among group variables are the same across levels of predictors.

3-Multicollinearity: Predictive power can decrease with an increased correlation between predictor variables.

4-Independence: Participants are assumed to be randomly sampled, and a participant's score on one variable is assumed to be independent of scores on that variable for all other participants.

4. 2.2 Objective of Liner Discriminant Analysis.

The main objectives for performing discriminant analysis are:

1- To identify the variables that best discriminate between groups using the most parsimonious way (i.e. to determine most influential predictors).

2- To use the identified variables or factors to develop a good classification function that is linear combination of the predictor variables and would be reliable in classification cases.

4- To assess the relative importance of the independent variables in classifying the dependent variable.

4. 2.3 Mathematical formula of linear discriminate analysis[4] [12] [13] [20]

This is a statistical method of classifying members of a population into one of two (or more) groups. The analysis entails the postulation and estimation of one or more Discriminant functions .In Discriminant analysis, we try to develop a model that will help us predict the values of a dependent variables on the bases of a set independent variables. The dependent variable in Discriminant Analysis is qualitative and appears in the form of success or failure, male or female, repay or default. The Discriminant Analysis attempts to derive a linear combination of these characteristics that best Discriminate between the group.

The prediction equation may be defined as

$$D = a + b_1x_1 + b_2x_2 + \cdots b_nx_n = a + \sum_{i=1}^{n} b_ix_i \quad \ldots\ldots\ldots\ldots \quad (21)$$

Where:

D : discriminate score

$b_1, b_2, b_3, \ldots, b_n$: is the discriminant coefficient or weight for that variable

x_n : predictor or independent variable

a : a constant

i: The number of predictor variables

a) Development of Discriminant functions (Group Statistics)

1-The Canonical Correlation[12]

is a multivariate analysis of **correlation**. **Canonical** is the statistical term for analyzing latent variables (which are not directly observed) that represent multiple variables (which are directly observed

2-Eigen Values[12][4]

The Eigen values are related to the canonical correlations and describe how best discriminating ability the functions possess. The % of variances is the discriminating ability of the 2 groups. Since there is only one function, 100% of the variance is accounted by this function. The cumulative % of the variance gives the current and preceding cumulative total of the variance. As mentioned above, as there is only one function in the present research wehave 100% of the cumulative variance.

b- Examination of whether significant differences exist among the groups, in terms of the predictor variables [4] [12] [13]

1-Wilk's Lambda (Λ)

In **Discriminant Analysis**, Wilk's lambda tests how well each level of independent variable contributes to the model

Wilk's lambda (Λ) after the effect of variables already in the discriminant function. Since the Wilk's lambda can be approximated by the F-ratio, Wilk's lambda (Λ) is equal to entering the variable that has the highest partial F-ratio. Wilk's lambda **(Λ)** is thus given by

$$\Lambda = \frac{SS_w}{SS_t} = \frac{SS_w}{(SS_w + SS_b)} \quad \ldots\ldots\ldots\ldots \quad (22)$$

Where SS_w is the sum of squares within groups, SS_t is the total sum of squares, and SS_b is the sum of squares between groups.

The scale ranges of Wilk's Lambda varies from 0 to 1, with 0 meaning group means differ (thus the variable highly differentiates the groups means total discrimination) (perfect discriminatory power), and 1(no discriminatory power) meaning all group means are the same means no discrimination. The assessment of the Wilk's lambda is done by converting to F- ratio with the transformation

$$F = \left[\left(\frac{1-\Lambda}{\Lambda}\right)\right]\left[\frac{(n_1+n_2-p-1)}{p}\right] \quad \ldots\ldots\ldots\ldots \quad (23)$$

Where n_1 and n_2 are the number of cases in group one and two respectively, p is number of variables for which the statistic is computed and Λ is the Wilk's lambda of the distribution .F-ratio follows an F-distribution with $(p-1)$ and (n_1+n_2-p-1) degrees of freedom.

Lambda tests the significance of each discriminant function in discriminate analysis specifically, the significance of the eigenvalue for a given function. minimizing Wilk's lambda is an indication that the within-group sum of squares is minimized and the between-group sum of squares is maximized.

c- Determination of which predictor variables contribute to most of the intergroup differences (Checking for relative importance of each independent variable)

1-The Standardized Canonical Discriminant Function Coefficients table

Displays coefficients which indicate relative importance of each variable in the model. These coefficients have similar sense with beta coefficients of multiple regressions.

These coefficients (b) are used to create the discriminant function (equation). It operates just like a regression equation. The discriminant function coefficients b or standardized form beta both indicate the partial contribution of each variable to the discriminate function controlling for all other variables in the equation. They can be used to assess each independent variables unique contribution to the discriminate function and therefore provide information on the

relative importance of each variable. If there are any dummy variables, as in regression, individual beta weights cannot be used and dummy variables must be assessed as a group through hierarchical LDA running the analysis, first without the dummy variables then with them. The difference in squared canonical correlation indicates the explanatory effect of the set of dummy variables.

To offset differing scales among the variables, the Discriminant Function Coefficients can be standardized using the equation 24

$$Z_1 = b_1^* \frac{Y_{1i_1} - \bar{Y}_{11}}{S_1} + b_2^* \frac{Y_{1i_2} - \bar{Y}_{12}}{S_2} + b_3^* \frac{Y_{1i_3} - \bar{Y}_{13}}{S_3} + \cdots + b_K^* \frac{Y_{1iK} - \bar{Y}_{1K}}{S_K} \quad \ldots\ldots\ldots\ldots \quad (24) i=1,2,3,\ldots,n$$

The standardized variables $\frac{Y_{1ir} - \bar{Y}_{1r}}{S_r}$ are scale free, and the standardized coefficients b_K^*

Ranking importance of the Variables.

2-Canonical Discriminant Function Coefficients[4]

The correlations are often referred to as loadings or structure coefficients and are routinely provided in many major programs contain the unstandardized coefficients for the dicsriminant model, similar to B coefficient in the multiple regression.

There is a close correspondence between interpreting discriminant functions and determining the contribution of each variable. In interpretation, the signs of the coefficients are taken into account; in ascertaining the contribution, the signs are ignored, and the coefficients are ranked in absolute value.

d- Evaluation of the accuracy of classification [20]

Classification table

Finally, there is the classification phase. The classification table, also called a confusion table, is simply a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications

Classification		Expected		Total
		P	N	
Observed	P	PP (TP)	PN (FN)	P
	N	NP (FP)	NN (TN)	P'
	Total	Q	Q'	1

As we see from the above table , we have the results of classification :

PP : actually positive and classified as positive.

PN : actually positive, but classified as negative.

NP: actually negative, but classified as positive.

NN: actually negative, and classified as negative

Ratio of correct classification (Hit Ratio) :

Is defined as the probability value of the correct classification, or efficiency ratio, such that if Efficiency (EF) calculated from

the equation : EF = TP + TN ..... (25)

Then Ratio of correct classification is obtained as follow:

Hit Ratio $= \frac{EF}{Total} = \frac{TP+TN}{P+P'} = \frac{TP+TN}{Q+Q'}$ ..... (26)

4. 3 ROC Curve [Receiver Operating Characteristic Curve] [4]

Using ROC curve for the classification accuracy, it is found that the area under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminate between those subjects who experience the response of interest versus those who do not. plotting sensitivity versus (1 – specificity) over all possible cut-points.

Sensitivity and specificity as well as other measures of classification performance computed from a 2 × 2 table depend on the single cut point used to classify a test result as positive. A better and more complete description of classification accuracy is the area under the (ROC) curve. The area under the ROC curve, which ranges from 0.5 to 1.0 provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not.

4. 5 Discriminant Function and Logistic Regression[16]

Discriminant function estimators have often been used in logistic regression, in both theory. When, such estimators were compared empirically with maximum likelihood estimators for logistic regression problems, however, they were found to be generally inferior, although not always by substantial. We will show why we prefer alternatives to discriminant function estimators for the logistic regression problem, as well as for the non normal discriminant analysis problem. It has been common practice to use discriminant function estimators as starting values in iterative maximum likelihood estimation and in exploratory data analysis, for the purpose of fitting logistic regression models. Other starting and exploratory estimators that have been suggested include "reverse Taylor series approximations," and "conditional estimators. "Conditional estimators" are obtained by maximizing the conditional likelihood (conditional on the explanatory variables). "Reverse Taylor series approximations" arise from the logistic cdf"

$$f(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad , \ \beta_1 \neq 0 \quad -\infty < x < \infty \ldots\ldots\ldots\ldots \quad (27)$$

Expanding about $x = Z$ (the sample mean) in a Taylor series, we get

$$f(x) = \left\{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} - \frac{\beta_1 \bar{x} e^{-(\beta_0+\beta_1 x)}}{\left[1+e^{-(\beta_0+\beta_1 x)}\right]^2}\right\} + \left\{\frac{\beta_1 \bar{x} e^{-(\beta_0+\beta_1 x)}}{\left[1+e^{-(\beta_0+\beta_1 x)}\right]^2}\right\}^2 x + R(x) \quad \ldots\ldots\ldots\ldots \quad (28)$$

Where $R(x)$ 'denotes a remainder containing terms of order $0(x - \bar{x})^2$. Neglecting (x), this may be interpreted as the linear function $A + Bx$, where

$$A = \left\{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} - B\bar{x}\right\} \ldots\ldots\ldots\ldots \quad (29) \quad , \quad B = \frac{\beta_1 e^{-(\beta_0+\beta_1 x)}}{\left[1+e^{-(\beta_0+\beta_1 x)}\right]^2} \ldots\ldots\ldots\ldots \quad (30)$$

Solving these equations for β_0 and β_1 (in reverse from the usual direction), we find

$$\beta_1 = B \,/\, [(A + B\bar{x})(1 - A - B\bar{x})] \quad \ldots\ldots\ldots\ldots \quad (31)$$

$$\beta_0 = -\beta_1 \bar{x} - log\left(\frac{1}{A + B\bar{x}} - 1\right) \quad \ldots\ldots\ldots\ldots \quad (32)$$

as the reverse Taylor series approximation. The results are easily generalized when $x, \beta_1 \ and \ B$ are vectors. We prefer the reverse Taylor series estimators to the discriminant function estimators since the former are appropriate regardless of the underlying distribution of explanatory variables, while the latter are really appropriate and justifiable only under

(a) Multivariate normality of the explanatory variables (a difficult assumption to satisfy in practice) (b) Complete equality of all of the underlying covariance matrices. (Transformations to induce multivariate normality will not typically induce equality of covariance matrices).

5. Practical aspects

5.1 Data and method of analysis

The data collected in Erbil (Layla Qasim) diabetic center for 250 individuals illustrated by two types of variable as follow:

a- Dependent (response) variable

Y: dependent variable is coded in binary response:

$$Y = \begin{cases} 1 \text{ Presence of diabetes} \\ 0 \text{ Absence of diabetes} \end{cases}$$

b-Independent (Explanatory) Variables Risk Factors for diabetes

The categories of the risk factors together with their description are demonstrated in table (1).

Table 1: Risk Factors for diabetes

	Independent (Explanatory) Variables [Risk Factors]	Codes (Categories)
X_1	High Blood Pressure (Hypertension)	(Yes=1,No=0)
X_2	Family History	(Yes=1,No=0)
X_3	Body Mass Index or (Obesity rate) =(Weight/height)^2	(Yes=1,No=0)
X_4	Diet (Nutrition)	(Yes=1,No=0)
X_5	High Lipid in Blood	(Yes=1,No=0)
X_6	Physical Activity	(Yes=1,No=0)
X_7	Age	the age from 24 -69 years

5.2 Results of statistical Analysis

The results of the study are summarized as follows

5.2.1 Logistic Regression Analysis

We start analyzing by the logistic method and it contains two Steps [Step 1 and Step 2]

1- Step 1

The process is inherently stepwise for forming and testing nested hierarchical models.

a-The classification table tells % cases correctly classified by the model.The first step is to compute and enter just the constant

Table (2) presents the results of the Binary Logistic Regression with the constant only included before any coefficients [risk factors: High Blood Pressure (Hypertension), Family History, Body Mass Index or (Obesity rate), Diet (Nutrition), High Lipid in Blood, Physical Activity and Age] are entered into the equation. Logistic Regression Analysis compares this model with a model including all the predictors to determine whether the latter model is more appropriate. The predicted result indicated to 90.8%, which reflects the model's overall explanatory strength.

Table 2: Shows the classification table

Observed				Predicted		
				Y		Percentage Correct
				0 .00	1.00	
Step 1	Y	0.00	Absences of Diabetic	0.00	23	0.00
		1.00	presence of Diabetic	0.00	227	100.0
	Overall Percentage					90.8%

b- Illustrating the variables in the equation (only the constant for Step1) is provided.

Table (3)**:** illustrates the variables in the equation, which is the constant term It can be realized that the intercept-only model has ln (odds) β = 2.289 with predicted odds [Exp (β)] = 9.870. That is, the predicted odd of having diabetes is 9.870. Since 227 of the sample of

persons have diabetes disease and 23 are have not diabetes, our observed odds are [227/23 = 9.870] Wald statistic is computed and since it is 109.466, with the null hypothesis is rejected Sig.= 0.000 , indicating that the constant does make a significant contribution to the model.

Table 3: Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1	Constant	2.289	0.219	109.466	1	.000	9.870

C-The contribution of each predictor were it added alone into the equation on the next step
Table (4) this table illustrate which variables are not in the equation, it indicates to the variable [X_7 (age)] is not in the equation because it is significant (p-value=0.000<0.01)and the rest of the risk factors seem to be not important at this step, but.
Overall it is statistically significant by (p-value=0.000<0.01) and therefore our model is quite good

Table 4: Variables not in the Equation

	Variables		Score	df	p-value
Step 1	X_1	Hypertension	0.019	1	0.891
	X_2	Family History	0.513	1	0.100
	X_3	(BMI) (Obesity rate)	0.569	1	0.101
	X_4	Diet (Nutrition)	0.087	1	0.768
	X_5	High Lipid in Blood	1.370	1	0.166
	X_6	Physical Activity	0.530	1	0.467
	X_7	Age	14.513	1	0.000
Overall Statistics			52.569	1	0.000

2- Step 2
This Step tests the contribution of all the variables entered
a- Omnibus Tests of model coefficients present Tests contain [Chi-squares testsof (Step, Block and Model)] tell us that the model was improved
Table (5) Show that:
Chi square value of Step is 77.535, by p- value = 0.000 which means whether the effect of the variable that was entered in the final step significantly differs from zero.
Chi square value of block is 77.535, tests whether every of the variables included in this block have effects that differ from zero.
Chi square value of model is 77.535 tells you whether any of the seven Independent Variables has significant effects. Overall it is statistically significant and therefore our model is quite good.

Table 5: Omnibus Tests of Model Coefficients

		Chi-square	df	p-value
Step 2	Step	77.535	7	0.000
	Block	77.535	7	0.000
	Model	77.535	7	0.000

.
b- Both R^2[Cox & Snell R Square and Nagelkerke R Square] values are presented to estimate the fit of the model to the data -- both are transformations of the-2log likelihood values:
table (6) provides the value of (Nagelkerke's R^2 gets all of the attention how well a linear model fits the data. is 0.781, indicating a strong relationship between the predictors and the

prediction. Under Model Summary the value of the -2 Log Likelihood statistic is 186.036[a] and from the table of Cox & Snell R^2 (R-Square) indicating that 66.7% of the variation in the independent variable is explained by the logistic model.

Table 6: Model Summary

Step 2	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
	186.036[a]	0.667	0 .781

c- The data fit the model by (Hosmer and Lemeshow) Test

Table (7) illustrates Hosmer-Lemeshow (H-L) test. The statistic under consideration has a significance of 0.810 which means that it is not statistically significant and therefore leading to the fact that the model is quite a good fit.

the value of the Hosmer- Lemeshow test of the goodness -of-fit statistic for the full model suggests the model is a good fit to the data Chi-square = 4.199 and the corresponding p-value from the chi-square distribution with (p-value =0.810 *>01* is not statistically significant for(a non significant chi-square indicates that the data fit the model well it emphasize that model is quite good which indicate that the risk factor in the model really cause diabetes .

Table 7: Hosmer and Lemeshow Test

Step 2	Chisquare	df	p-value
	4.199	7	0.810

d- The reclassification table shows the accuracy of the model

In (Table 8) reclassification table including the constant term and the rest of the predictors 98.7% were correctly for the presence of diabetes and correctly classified the overall correct percentage was 95.6% which reflects the model's over all explanatory strength it refer to the best prediction of the risk factors of diabetes.

Table 8: Reclassification Table

Observed			Predicted		
			Y		Percentage Correctly
			0.00	1.00	
Step 2	Y	0.00	15	8	65.2
		1.00	3	224	98.7
	Overall Percentage				95.6

a-After the reclassification: Interpreting the Logistic Regression Equation(model) by testing of the contribution of each parameter

Table 9 is about the variables that are included in the logistic regression equation(model).

Values of (Wald statistic and Exp(β) is odds ratio) are greater than(1) refer to the risk factors are[X_1= High Blood Pressure (Hypertension) , X_2=Family History, X_3= Body Mass Index(BMI) (Obesity Rate) , X_4=Diet(Nutrition) , X_5=High Lipid in Blood, X_6= Physical Activity] are statistically significant factors.The value for only (X_7= age) less than 1 and it is not statistically significant., implying that the probability the risk of age is less than getting other risk factors this emphasize that (X_7) age is not in the prediction Equation.

Table9 : Variables in the equation

			β	S.E.	Wald	df	p-value	Exp(β)
	X_1	Hypertension	0.918	0.229	16.06	1	0.002	2.505
	X_2	Family History	1.821	0.643	8.063	1	0.005	6. 178

Step 2	X_3	(BMI) (Obesity rate)	1.732	0.593	8.542	1	0.003	5.652
	X_4	Diet (Nutrition)	0.710	0.135	26.757	1	0.000	2.033
	X_5	High Lipid in Blood	1.476	0.667	4.971	1	0.026	4.375
	X_6	Physical Activity	1.073	0.173	38.468	1	0.000	2.924
	X_7	Age	-0.021	0.024	0.765	1	0.251	0.979
	Constant		3.626	1.295	7.84	1	0.005	37.562

b-Determining the Logistic Equation (Model) by the "β" values which they are the logistics equation coefficients that can be used to create a predictive equation.

According to value of odds ratio we can ranking the risk factors on diabetes as follows :

(X_2= family history= 6. 178, X_3=(BMI) (Obesity rate)=5.652 , X_5= high lipid in blood =4.375 , X_6= physical activity =2.924 , X_1= hypertension =2.505, X_4= diet (nutrition)=2.033)

Logistic Equation (Model)= $Y= 3.626+0.918 X_1+1.821 X_2+1.732 X_3+0.710 X_4 +1.476 X_5+ 1.073 X_6$.

Logistic Equation (Model)= Y= 3.626 +0.918 Hypertension+1.821 Family History +1.732 Body Mass Index or (Obesity rate) +0.710 Diet(nutrition) +1.476 High Lipid in Blood+ 1.073 Physical Activity.

5.3 Discriminant Analysis [12] [13]

Discriminant Analysis contains the following analyses

a) Development of discriminant functions

(table10)illustrate the development of discriminant **,**the group statistics gives the distribution of observations into different groups since, in the present research we have categorized into two groups absence of disease as '0' and presence of disease 1', the SPSS has grouped the data into two groups. The total numbers of 250 observations group, which represent 100% of the observations, have been grouped for the Discriminant Analysis. The function indicates the first canonical linear discriminant function. The number of function depends on the discriminating variables. The function gives the projection of the data that best discriminant between the groups

Table10: illustrate the development of discriminant functions(Group Statistics)

Y		List wise(Weighted)
0.00 (Absence of diabetes)	First group contains (X_1, X_2, X_3, X_4, X_5, X_6, X_7)	23.000
1.00 (Presence of diabetes)	Second group Contains (X_1, X_2, X_3, X_4, X_5, X_6, X_7)	227.000
Total		250.000

In table (11) The maximum number of discriminate functions produced is the number of groups minus 1, since only two groups used here, namely 'Presence of diabetes' and 'Absence of diabetes', so only one function is displayed a larger Eigen value explains a strong function but here the value of Eigen value(0.373). and the measure of the canonical correlation is (0.521)which refers to a correlation between the discriminate scores and the levels of the dependent variables is not very strong. Both values indicate to weak effect of the risk factors on the diabetes disease

Table (11):Represent Eigenvalue and Canonical Correlation

Function 1	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
	0.373	100.0	100.0	0.521

b- Examination of whether significant differences exist among the groups, in terms of the predictor variables

Table (12) shoes the value of Wilks' Lambda is 0.728 nearest to(1)(no discriminatory power) with Chi-square(77.487) p-value (0.000) it means that the factors by haven't risks on diabetes disease

Table (12) Tests of Equality of Group Means

	Wilks' Lambda	Chi-square	df	p-value
Test of Function 1	0.728	77.487	7	0.000

c- Determination of which predictor variables contribute to most of the intergroup differences (Checking for relative importance of each independent variable)

The standardized canonical discriminant function coefficients table is the interpretation of the discriminant coefficients provides an index of the importance of each predictor. The sign indicates the direction of the relationship .On comparing the standardized coefficient, it is possible to identify which independent variable is more discriminating than the other variables. The higher the discriminating powers the higher the standardized discriminant coefficient.

In table (13) only the body mass index(BMI) (Obesity rate) ($X_{1)}$ has the highest discriminate coefficient is 0.883 this indicates that it is only predictor of risk factor and the effect of X_2 family history is 0.458 ….ect

Table 13: Standardized Canonical Discriminant Function Coefficients

Variables		**Standardized Canonical Discriminant Function Coefficients**
X_1	Hypertension	0.146
X_2	Family History	0.458
X_3	(BMI) (Obesity rate)	0.883
X_4	Diet (Nutrition)	0.077
X_5	High Lipid in Blood	0.307
X_6	Physical Activity	0.081
X_7	Age	-0.113
Constant		8.820

Ranking importance of the Variables.

The ranking in (Table 14) Based on the coefficients in (Table 13) for the relative important predictor variables summarized as follow:

Table 14: Ranking of the Variables

Ranking of the Variable		**Predictor Variable**
X_3	Body Mass Index	0.883
X_2	Family History	0.458
X_5	High Lipid in Blood	0.307
X_1	(Hypertension	0.146
X_7	Age	-0.113
X_6	Physical Activity	-0.081
X_4	Diet (nutrition)	0.077

In (Table 15) Canonical Discriminant Function Coefficients represent the Coefficients of final Canonical Discriminant Function

Table 15:Canonical Discriminant Function Coefficients

Variables		Function 1
X_1	Hypertension	0.319
X_2	Family History	1.021
X_3	(BMI) (Obesity rate)	0.276
X_4	Diet (nutrition)	0.163
X_5	High Lipid in Blood	0.647
X_6	Physical Activity	-0.162
X_7	Age	-0.009
(Constant)		-8.820

Discriminant Function = -8.820+0.319high blood pressure (hypertension) +1.021family history +0.276body mass index +0.163diet +0.647high blood lipid - -0.162daily activity -0.009age

d- Evaluation of the accuracy of Prediction of **classification**

Table 16: classification table

Step1							**Step2**					
Observed		**Predicted**					**Observed**		**Predicted**			
		Y		Total					Y		Total	
		0.00	1.00						0.00	1.00		
Y	**0.00**	10.00	20.0	30.0	12%		**Y**	**0.00**	15.0	19.0	34.0	%13.6
	1.00	20.00	200.0	220	88%			**1.00**	61.0	155.0	216.0	%86.4
	Overall Percentage				85%			Overall Percentage				%84

5.4 Comparison with Result of Classification Table for Two Statistical Methods

The most frequently used criterion for comparison between the two methods is classification. (Table 18) shows the classification results of the two statistical methods (LRA and LDA) for dependent variable diabetes disease.

(Table 17) indicates that the LRA model can correctly classify the first category Absence of disease with accuracy of step0 (65.2 %) compared (13.6 %) for LDA model .

At the second presence of disease, correct classification accuracy of LRA model (98.7%) is more than the LDA model (86.4%) , As we can see from the above results, the correct classification rates for all categories by the LRA model is better than LDA model because there is (95.6%) of classified to the correct categories using LRA models, while (84%) been classified to the correct categories using LDA

Table 17: Classification Result of Two Statistical Methods LRA and LDA

		Percentage	
Observed		**Predicted** Y	
		LRA	LDA
Y	0.00	65.2%	13.6%
	1.00	98.7%	86.4%
	Overall Percentage	95.6%	84%

5.5 ROC Curve in the figure 1 the ROC curve used for the classification accuracy ,for all variables.

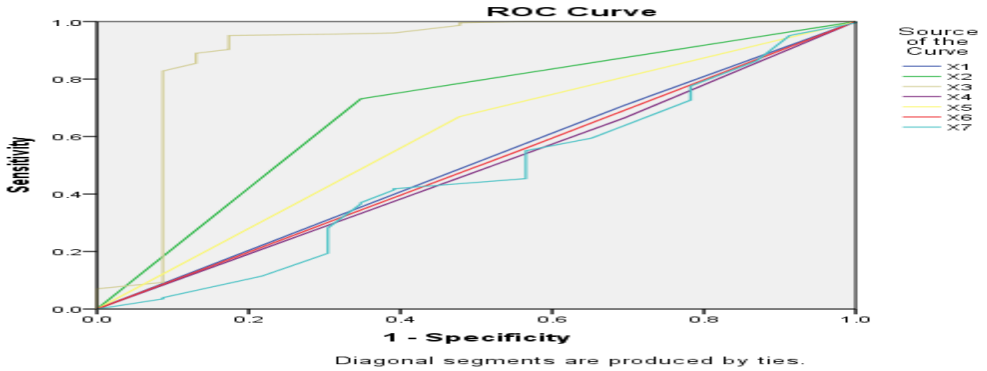


Figure (1) Area under the Curve of all variables

In (Table 18) The test result variable(s): X_1, X_2, X_3, X_4, X_5, X_6, X_7 has only one tie between the positive actual state group and the negative actual state group. the highest value is(0.897)for (BMI) (Obesity rate) it represent the highest predictor for the risk factor

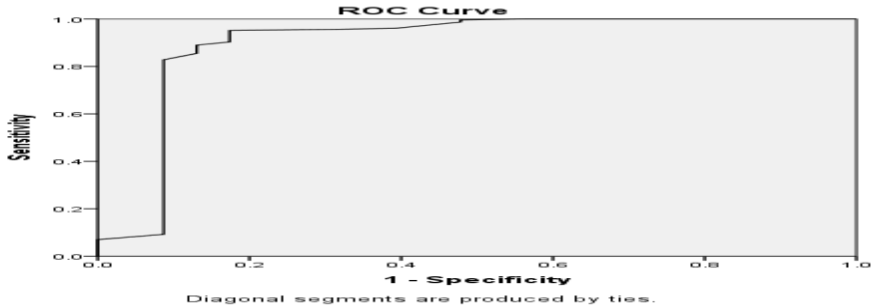
Table 18: Area under the Curve of all variables

Test Result Variable(s)		Area
X_1	Hypertension	0.507
X_2	Family History	0.692
X_3	(BMI) (Obesity rate)	0.897
X_4	Diet (Nutrition)	0.485
X_5	High Lipid in Blood	0.596
X_6	Physical Activity	0.497
X_7	Age	0.463

Ranking of the areas under curve of risk factors by Roc Curve test

(BMI) (Obesity rate)(0.897)**,** Family History (0.692**) ,** High Lipid in Blood **(**0.596**) ,** Hypertension **(**0.507**) ,** Physical Activity **(**0.497**) ,** Diet (Nutrition) 0.485,and Age 0.463

Figure ROC Curve for the highest area is 0.897 it is the result of the ROC Curve for final model



Figure(2) Area Under the Curve for the highest area

In (Table 19) the area under ROC curve reached 89.7%, and (Asymptotic Sig = 0.000) ROC curve was used to compare prediction powers of the models

Table 19: Test Result all Variable(s) Area

Area	Std. Error	Asymptotic p-value	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.897	.053	0.000	0.792	1.000

6.Conclusion

This paper consists of two parts, a theoretical part and a practical(application) part were used for two analyzes (Logistic regression Analysis) and (Linear Discriminant Analysis) for the data of size 250 individuals were taken from Erbil Layla Qassem Center for Diabetes, it consists of eight variables, one of them is a dependent variable represented the (presence and absence) of diabetes, and the other seven variables are explanatory variables that represent the risk factors of diabetes which are [Hypertension, family history, body mass index(BMI), diet, high blood lipid, physical activity, and age]. Best results have been obtained by using the Logistic regression tests, which include [Score test-for Variables not in the equation, predicted odds, Wald statistics, Omnibus tests, and Hosmer-Lemeshow (H-L) tests], by a reclassification of the accuracy decision prediction of the model, the accuracy percentage of the model was 95.6%, which reflects to a high ratio of prediction of Logistic Regression Analysis, but by Linear Discriminant Analysis of tests such as [Eigen value, Canonical Correlation, Wilks' Lambda, and Chi-square], the results of Eigen value and Canonical Correlation were minimal, but Wilks' Lamba and Chi Square were higher, Wilks' Lambda had a 0.728 which is nearest to no discriminatory power, and the Chi-square test had the same result. After a reclassification of the accuracy decision prediction of the model, the accuracy percentage of the model was 84%, which reflects to a lower ratio of prediction than Logistic Regression Analysis. The result of area under the ROC Curve of all variables, which is used to compare prediction powers of the models, was 0.897. This shows that the Logistic Regression Analysis has the best prediction of risk factors of diabetes and it has the appropriate model. The factors were determined according to the risk factors on diabetes ranking as follows : [X_2= Family History = 6.178, X_3=(BMI) (Obesity rate)=5.652,X_5= High Lipid in Blood =4.375,X_6=Physical Activity=2.924,X_1=Hypertension =2.505, X_4= Diet (Nutrition)=2.033], but by the Score test for the variables not in the equation only [X_7 (age) it is not represents the risk factor

7. Recommendations

According to the conclusion reported above we may recommend the followings:

1- Since (Family history) is the first risk factors and the age does not represent a risk factor of diabetes so the infection does not depend on age, so it is recommend that families who have diabetes in their family history to seek attention and get continuous follow-ups starting from childhood.

2- Obesity, High Lipid in Blood, and Hypertension are also risk factors of diabetes, so it is recommended to exercise and maintain a healthy nutritious diet, which are key factors to reduce the risk.

3- It is recommended to use Logistic Regression methods in every medical case to obtain the best prediction**.**

References

[1] Adeeb A. Ali AL Rahamneh , Omar M. Hawamdeh**(2017)**' The Factors Affecting Eye Patients (Cataract) In Jordan by Using the Logistic Regression Model' Modern Applied Science; Vol. 11, No. 8; 2017 ISSN 1913-1844 E-ISSN 1913-1852 Published by Canadian Center of Science and Education

[2]Ali, A., & Mohammad, A. (2010) 'The Logistic Regression Model to Breast Cancer Among Sudanese Women' Phd Thesis, Faculty of Science, Sudan University of Science and Technology

[3] Cox, D.R. and E.J. Snell, (1989) 'The Analysis of Binary Data' 2nd edition. Chapman and Hall, London, pp. 144-163.http://www.webmd.com/eye-health/glaucoma-eyes.

[4] Dr.P.Venkatesh (2012) 'Discriminant Analysis: An Overview Scientist' Division of Agricultural Economics, IARI

[5]Elizabeth Brown(2004)'Lecture 13 Estimation and hypothesis testing for logistic regression' BIOST 515 February 19, 2004 https://courses.washington.edu/b515/l13.pdf

[6] Hassan, & Ahmad, F. (2014) 'The Use of the Logistic Model to Determine the Factors Affecting the Incidence of Anemia (Anemia) In Children.' Master Thesis, Faculty of Science, Sudan University of Science and Technology.

[7] Hauck, W.W. and A. Donner. (1977)'Wald's Test as applied to hypotheses in logit analysis' Journal of the American Statistical Association, vol. 72, pp. 851-853.
[8] Hosmer, David W.; Lemeshow, Stanley (2013).' Applied Logistic Regression' New York: Wiley. ISBN 978-0-470-58247-3.
[9] Hosmer, David W, Scott Taber, and Stanley Lemeshow(1991) 'The importance of Assessing the Fit of Logistic Regression Models: A Case Study'. American Journal of Public Health , vol. 81, pp. 30-35.
[10]Karl L. Wuensch and Poteat (2016) 'Binary Logistic Regression with SPSS' published in the Journal of Social Behavior and Personality, 13, 139-150.
[11]Lazha A. Talat Shareef (2010)'Prevalence and risk factors of diabetic retinopathy among a sample of diabetic pregnant women in Erbil city'Higher Diploma ,Ophthalmology
[12] Maja Pohar, Mateja Blas, and Sandra Turk (2004) ' Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study' Metodološki zvezki, Vol. 1, No. 1, 143-161
[13]Majed F Al. (2012)' A Comparative Study Between Linear Discriminant Analysis and Multinomial Logistic Regression in Classification and Predictive Modeling ' A Thesis Submitted in Partial Fulfillment of Requirements for the Degree of M.Sc. of Applied Statistics in Al Azhar University Gaza Faculty of Economics and Administrative Science
[14] Nagelkerke, N.J.D.(1991)'A note on the general definition of the coefficient of determination'. Journal of Biometrika, vol. 78, pp. 691-692.
[15] Nidal Mohamed Mustafa Abd Elsalam(2013)) 'Binary Logistic Regression to Identify the Risk Factors of Eye Glaucoma ' Department of Statistics & Computation Faculty of Technology of Mathematical Sciences & Statistics, Al Neelain University International Journal of Sciences, Basic and Applied Research (IJSBAR) ISSN 2307-4531 (Print & Online) http://gssrr.org/index.phpjournal=JournalOf BasicAndApplied
[16]Osibanjo F. S., 2Olalude G. A. 2Akintunde M. O.and 2Ajala A. G.(2015) 'Application of Logistic Regression model to Admission Decision of foundation Program at University of Lagos'. International Journal of Mathematics and Statistics Studies Vol.3, No.4, pp.27-41, July
[17] Parastoo Rahimloo, Ahmad Jafarian (2016)' Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them'Bulletin de la Société Royale des Sciences de Liège, Vol. 85, 2016, p. 1148 - 1164
[18] Pickup, J.C., Williams G., (2003)' Textbook of diabetes'Blackwell Science, Oxford,
[19]Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S. (2010)' Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies'Emerging Risk Factors Collaboration 26; 375:2215-2222
[20]S. James Press; Sandra Wilson, (1978)' Choosing Between Logistic Regression and Discriminant Analysis ' Journal of the American Statistical Association, Vol. 73, No. 364., pp. 699-705.
[21] Szumilas Magdalena (2010) ' Explaining Odds Ratios ' T J Can Acad Child Adolesc Psychiatry.; 19(3): 227–229. MSc
[22]Warrick Junsuk Kim(2016)' Global report on Diabetes World Health Organization (WHO) 'website (http://www.who.int) or can be purchased from WHO Press, ISBN 978 92 4 156525 7 (NLM classification: WK 810)
[23] Wen-Chih Wu1,2,,Mary E. Lacy,1,2 Gregory A. Wellenius,1Mercedes R. Carnethon, Eric B. Loucks, Xi Luo (2016)'Racial Differences in the Performance of Existing Risk Prediction Models for Incident Type 2 Diabetes' The Caraia Study
39:285–291 | DOI: 10.2337/dc15-0509
[24] Zhang Biao(2006_) 'A score test under logistic regression models based on case–control data' first published: 12 October https://doi.org/10.1111/j.1467-9574.2006.00336.

پوختە

زۆر لە توێژینەوە پزیشکیەکان ئاماژە ئەدەن بە بوونی پەیوەندییەکی بەهێز لەنێوان دەست نیشان کردنی نەخۆشی وهەندێ شیکاری ئاماری وەك شیکردنەوەی لۆجستی(LRA)(Logistic Regression Analysis) وەشیکردنەوەی جودا گەری Linear Discriminant Analysis LDA کەهەردوکیان رێگای شیکردنەوەی فرە گۆڕاون (multivariate statistical methods) بە شێوەیەکی بەرفراوان بەکار ئەهێنرێن بۆ پێش بینی کردن (prediction) .لەم توێژینەوەیەدابۆ هەردووشیکردنەوەکە ئەنجامدرالەسەرکۆمەڵداتایەکی تایبەت بەنەخۆشی شەکرە کەژمارەیان (250)کەسەکە لەناوەندی(لەیلا قاسم)ی نەخۆشی شەکرەی هەولێر وەرگیراوە پێکهاتووە لە(8) گۆڕاوو(1) یەکێکیان گۆڕاوی پشت بەستوو Dependent variable)) کە (بوون) و(نەبوونی) نەخۆشی شەکرە دەنوێنێ وە(7) گۆڕاوەکەی ترکەگۆڕاوی پێشبینیکراوە predictor)) یان گۆڕاوی سەربەخۆ(independent) دەنوێنن کەبریتین لە هۆکاری مەترسیدار بۆتوش بوونی نەخۆشی شەکرە وەك (تەوژمی خوێن , بۆماوەیی خێزانی , بارستایی لەش(رێژەی قەڵەوی) , جۆری خۆراك (چەور وبێ چەور) , ڕێژەی چەوری لەخوێندا , چالاکی ڕۆژانە(وەرزش) وە تەمەن). ئەم توێَژینەوەیەپێکهاتووە لەدوو بەش یەکەمیان بەشی تیۆریە دووەمیان بەشی کرداریە وئامانجی ئەم توێَژینەوەیەبەراوردکردنی هەردووشیکردنەوەی (LRA)وە (LD A) لەسەربنچینەی چەند پێوەرێکی ووردی پێش بینی کردنە وەهەروەها بۆدیاریکردنی باشترین مۆدیلی ئاماری پێکهاتوولەهۆکارەترسناکەکان بۆ توش بوونی نەخۆشی شەکرە . بەپێی ئەنجامی هەموو جۆرەتاقیکردنەوەکانی پێش بینی کردن بەڕێگای LRA گونجاوترە.وە لەئەنجامی ڕووبەری ژێر چەماوەی (ROC Curve) کەبەکارئەهێنرێت بۆبەراوردی پێش بینی کردنی نێوان مۆدیلەکان جەخت لەسەرئەوەدەکات کەشیکردنەوەیLRA باشترین ڕێگایە بۆپێش بینی کردنی هۆکارەترسناکەکان بە توش بوونی نەخۆشی شەکرە وە بەباشترین جێگرەوەی L DAدائەنرێت وە لە ئەنجامی شیکردنەوەی (LRA) ڕیزبەندی هۆکارەترسناکەکان بە توش بوونی نەخۆشی شەکرەبریتین لەم]1- (بۆماوەیی خێزانی , 2-بارستایی لەش(رێژەی قەڵەوی) ,3-ڕێژەی چەوری لەخوێندا, چالاکی ڕۆژانە(وەرزش,(تەوژمی خوێن, جۆری خۆراك (چەور وبێ چەور)) [بەڵام تەمەن.هۆکارێکی ترسناك نیە.

ملخص

تشير العديد من الدراسات الطبية إلى وجود علاقة وثيقة بين الجوانب التشخيصية للمرض وبعض التحليلات الإحصائية مثل تحليل الانحدار اللوجستي (LRA) والتحليل التمييزي الخطي (LDA) ، وكلاهما طريقتان إحصائيتان متعددتان تستخدمان على نطاق واسع لتحليل البيانات والتنبؤ . في هذه الورقة حيث تم مناقشة وتحليل كلا التحليلين على بيانات بحجم العينة 250 مريض مصاب بداء السكري تم جمعها من مركز ليلى قاسم للمرض السكري في اربيل.هذه البيانات تحتوي على(8) متغيرات ، أحداها متغير تابع والذي يمثل وجود أو عدم وجود مرض السكري ، والمتغيرات السبعة الأخرى هي متنبئات او(متغيرات مستقلة) ، يتم أخذها في النموذج الذي تمثل فيه اخطر العوامل للاصابة بمرض السكري مثل: [ارتفاع ضغط الدم ، تاريخ العائلة مع المرض، مؤشر كتلة الجسم (BMI) - الجرعة ، والنظام الغذائي (التغذية) ، ارتفاع الدهون في الدم ,النشاط البدني والعمر]. تحتوي هذه الورقةالجانبان، الجانب النظري والجانب العملي و تهدف إلى المقارنة بين تحليل الانحدار اللوجستي والتحليل التمييزي الخطي على أساس عدة المقاييس للدقة التنبؤية لاختيار أفضل نموذج إحصائي لتحديد عوامل الخطرة للاصابة بمرض السكري.. تم إجراء اختبار نتائج كلاالتحليلين، والنتيجةتشير الى ان التنبؤبطريقة تحليل الانحدار اللوجستي ادق من طريقة التحليل التمييزي الخطي (LDA) ومن نتائج المنطقة تحت منحنى (ROC) الذي يستخدم لمقارنة قوى التنبؤ لجميع المتغيرات أكد على أن تحليل الانحدار اللوجستي لديه أفضل تنبؤ بعوامل الخطرة لمرض السكري ولديه النموذج المناسب لذلك ظهر تحليل الانحدار اللوجستي كبديل قوي لتحليل التمييز الخطي و من خلال الانحدار اللوجستي ، فإن ترتيب عوامل الخطر على مرض السكري هو على النحو التالي

1-تأثير الوراثي ,2-(زيادةمعدل السمنة) ,3-ارتفاع الدهون في الدم ,4-قلة النشاط البدني ،5- ارتفاع ضغط الدم 6- الحمية الغذائية (التغذية)، ولكن(العمر) لا يمثل عامل الخطر.