**Logistic Regression and Discriminate Analysis to Identify the Risk Factors of Diabetes**

**Abstract**

Many medical studies point out that there is a close relationship between the diagnostic aspects of a disease and some statisticalanalysis like Logistic Regression Analysis (LRA) and Linear Discriminant Analysis (LDA) both of them are two widely used multivariate statistical methods for data analysis they applied in order to predict the probability of a specific categorical outcome based upon several explanatory variables (predictors).In this study both analyses were discussed and implemented on 250 individuals collected from Erbil diabetic Center (Layla Qasim). The data contained (8) variables, one variable is the diabetes disease (its presence and absence) which this variable is called (dependent variable), 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:

[Hypertension, family history, body mass index (BMI) or (Obesity rate), diet (nutrition), high lipid in blood, physical activity and age]. The study aims the comparison between the LDA and LR is based on several measures of predictive accuracy to choose the best statistical model for identify risk factors of diabetes by each method. The study contains two parts, Theoretical aspects and Practical aspects.

The results proved that the used logistic regression model contains two steps [(step1) and(step 2) ].In step1 [this step is used to compute and enter only the constant], the classification 90.8 % cases correctly classified, predicted odds [Exp (β)] = 9.870 and Wald statistic is computed(109.466) with p-value.= 0.000, indicating that the constant does make a significant contribution to the model the contribution of each predictor illustrate which variables are not in the equation, it indicates to the variable [X7 (age)] is not in the equation because it is significant (p-value=0.000< 0.05 )

,In step2 tests the contribution of all the variables entered the Omnibus tests ,Hosmer-Lemeshow (H-L) test leading to the fact that the model is quite a good and reclassification shows the accuracy of the model classification the overall correct percentage was 95.6% which reflects the model’s Over all explanatory strength . By Discriminant Analysis the reclassification table shows the accuracy decision prediction %84 were correctly classified into group recommended for direct, in the Standardized Canonical Discriminant Function. The area under ROC curve reached 89.7%, and (Asymptotic p-value = 0.000)ROC curve was used to compare prediction powers of the models .

To specify a model with the highest prediction the results illustrated that Logistic Regression Analysis was better than Linear Discriminant Analysis .So Logistic regression has emerged as a robust alternative to discriminant analysis.

**Keywords**: logistic regression, Maximum Likelihood (Likelihood Ratio Test) , (Hosmer and Lemeshow test) and (R Square Nagelkerke,)R Cox & Snell R-Square , Omnibus Tests, Discriminant analysis , Canonical Correlation**,** Wilk’s Lambda, ROC curve

**1. Introduction**

T

he 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 or 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

**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 projects that diabetes will be the seventh leading cause of death in 2030.

 In this study, the two classification methods namely linear discriminant analysis (LDA) and logistic regression analysis (LRA) are used, both of them are two widely used multivariate statistical methods for data analysis with categorical outcome variables. Both methods 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 extant 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 a [categorical](https://en.wikipedia.org/wiki/Categorical_variable) [dependent](https://en.wikipedia.org/wiki/Dependent_variable) [variable](https://en.wikipedia.org/wiki/Variable_(mathematics)#Applied_statistics) by one or more [continuous](https://en.wikipedia.org/wiki/Continuous_variable) or [binary](https://en.wikipedia.org/wiki/Binary_variable) [independent](https://en.wikipedia.org/wiki/Independent_variable) variables (called predictor variables). Both methods can construct linear classification model which creates a linear boundary between two groups. In other words, LDA and LRA assist in profiling the characteristics of the subjects being studied and then assign them to the most suitable group. 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, LR 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 LR is much longer than the time taken by LDA, making it a less desirable alternative to LDA.

This study puts its focus on the problem of choosing between the two methods, and set some guidelines for proper choice.

**2.Objectives of the study**

The study aims the comparison between the LDA and LR is based on several measures of predictive accuracy to choose the best statistical model for identify risk factors of diabetes by each method.

**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,Cronbach’s alpha, Logistic Regression (LR) and linear discriminant analysis (LDA ).

2. Part II: The Practical aspects study for Explaining the statistical file of the logistic regression (LR) and discriminate model (LDA) and focusing on the characteristics and how to estimate the parameters.

**4. Theoretical aspects**

**4.1What is Cronbach’s alpha?**

Cronbach’s alpha is a measure used to assess the reliability, or internal consistency, of a set of scale or test items. In other words, the reliability of any given measurement refers to the extent to which it is a consistent measure of a concept, and Cronbach’s alpha is one way of measuring the strength of that consistency. Cronbach’s alpha is computed by correlating the score for each scale item with the total score for each observation (usually individual survey respondents or test takers), and then comparing that to the variance for all individual item scores:

………… (1)

|  |  |
| --- | --- |
| where: | refers to the number of scale items |
|  | refers to the variance associated with item i |
|  | refers to the variance associated with the observed total scores |

Cronbach’s alpha is thus a function of the number of items in a test, the average covariance between pairs of items, and the variance of the total score. The resulting αα coefficient of reliability ranges from 0 to 1 in providing this overall assessment of a measure’s reliability. If all of the scale items are entirely independent from one another (i.e., are not correlated or share no covariance), then αα = 0; and, if all of the items have high covariance's, then αα will approach 1 as the number of items in the scale approaches infinity. In other words, the higher the αα coefficient, the more the items have shared covariance and probably measure the same underlying concept.

**4.2 Binary Logistic Regression**

In [statistics](https://en.wikipedia.org/wiki/Statistics), logistic regression, is a model where the [dependent variable](https://en.wikipedia.org/wiki/Dependent_and_independent_variables)  is [categorical](https://en.wikipedia.org/wiki/Categorical_variable). This article covers the case of a [binary dependent variable](https://en.wikipedia.org/wiki/Binary_variable) that is, where the output can take only two values, "0" and "1", which represent outcomes such as healthy/sick. Cases where the dependent variable has more than two outcome categories may be analyzed in [multinomial logistic regression](https://en.wikipedia.org/wiki/Multinomial_logistic_regression), or, if the multiple categories are [ordered](https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_type), in [ordinal logistic regression](https://en.wikipedia.org/wiki/Ordinal_logistic_regression). In the terminology of [economics](https://en.wikipedia.org/wiki/Economics), logistic regression is an example of a [qualitative response/discrete choice model](https://en.wikipedia.org/wiki/Discrete_choice). Logistic regression was developed by statistician [David Cox](https://en.wikipedia.org/wiki/David_Cox_(statistician)) 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, i.e. 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.

**4.2.1 The Model of Binary Logistic Regression**

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 Glaucoma. Similarly, we will define 1-P(x) to be the probability that Y =0, which is absence of Glaucoma. These probabilities are written in the following form:

………… (2)

………… (3)

The log distribution (or logistic transformation of (g) is also called the logit of p or logit (g) 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:

………… (4)

Whereas can only range from 0 to 1, scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero).

Formula (2) below shows the relationship between the usual regression equation , which is a straight line formula, and the logistic regression equation. The form of the logistic regression equation is thus rewritten as:

………… (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 coefficients.

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:

………… (6)

Where:

= the probability that a case is in a particular category,

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

= the constant of the equation

= the coefficient of the predictor variables

**4.2.2 Estimate the Parameters of Logistic Regression Model**

estimated using Maximum Likelihood Estimation method (MLE) , 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 β0 and β1 In logistic regression the method Maximum likelihood will provide values of β0 and β1 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 (β0 and βb1). "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:

………… (7)

For estimation, we will work with the log-likelihood

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

………… (8)

**4.2.3 For analyzing by**

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

**4.2.3.1 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**

Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the [Wald statistic](https://en.wikipedia.org/wiki/Wald_test). 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.

………… (9)

For large n, with 1 degree of freedom. In R, you will see

reported **.**

**3-Score test**

If the MLE equals the hypothesized value, , then would maximize the likelihood and 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 , and with 1 degree of freedom.

**4.2.3.2 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

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](http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/glossary/index31e8.html?selectedLetter=C#chisquare) to see if there is a significant difference between the Log-likelihoods (specifically the [-2LLs](http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/glossary/index1695.html?selectedLetter=D#deviance-2ll)) 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

So the model tested can be defined by:

### The omnibus test relates to the hypotheses

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**

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).

………… (10)

For large n, with degrees of freedom equal to the number of parameters being estimated.

For the binary outcome discussed above, if the hypothesis is

Vs: , then

and the LRT statistic is

………… (11)

**Where**

**3-Test by two R2**

**a)Cox and Snell R2CS or []**

In this model in instead of the coefficient of determination by () ()

………… (12)

………… (13)

: Maximum likelihood function in the case the model being fitted

: Maximum likelihood function in the case of the   null model

**b)**

Can be calculated as

………… (14)

L0: Maximum likelihood function in the case of the   null model

n: Sample Size

**4-Hosmer and Lemeshow**

The [Hosmer–Lemeshow test](https://en.wikipedia.org/wiki/Hosmer%E2%80%93Lemeshow_test" \o "Hosmer–Lemeshow test) uses a test statistic that asymptotically follows a[{\displaystyle \chi ^{2}} distribution](https://en.wikipedia.org/wiki/Chi-squared_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.

: There is no significant difference between observed values and expected values.

: 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.

………… (15)

Where; -

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

:Is the number of observation in the ith group.

Is the number of event outcomes in the ith group?

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

**5-Odds and** **Odds Ratio (OR)**

### Definition of the odds

The odds of the dependent variable equaling a case (given some linear combination {\displaystyle x}() of the predictors) is equivalent to the exponential function of the linear regression expression. This illustrates how the [logit](https://en.wikipedia.org/wiki/Logit" \o "Logit) serves as a link function between the probability and the linear regression expression. Given that the logit ranges between negative and 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

and

Recall the logistic model: is the probability of disease for a given value of , and

………… (15)

Then for (),

………… (16)

For (),

………… (17)

odds of =

odds of ={\displaystyle {\text{odds}}=e^{\beta \_{0}+\beta \_{1}x}.}

**Odds Ratio (OR)**

The odds ratio is a measure of association for 2×2 contingency table

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

|  |  |  |
| --- | --- | --- |
|  | exposed | |
|  | 1 | 0 |
| 1 | a | b |
| 0 | c | d |

Where:

………… (18)

………… (19)

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

………… (20)

**6-The standard error**

S.E.: Standard error is a statistical term that measures the accuracy with which a sample represents a population. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error.  
The standard error is an important indicator of how precise an estimate of the population parameter the sample statistic is. Taken together with such measures as effect size, p-value and sample size, the effect size can be a useful tool to the researcher who seeks to understand the accuracy of statistics calculated on random samples

In final analysis it indicate to the variables in prediction equationWald test for each variables ,Odds Ratio (OR) and standard error

**4. 3 Second type of Analysis** : **Linear Discriminant Analysis (LDA) also known as**

**4. 3.1 Discriminant Analysis (DA) or Two-group discriminant analysis.**

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. 3.2 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. 3.3 Mathematical formula of linear discriminate analysis**

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 groups.

The prediction equation may be defined as

………… (21)

Where:

D : discriminate function

: are the discriminant coefficient or weight for that variable

: Respondent’s score for that variable

: a constant

: The number of predictor variables

**a) Development of discriminant functions( Group Statistics**)

**1-The** **Canonical Correlation**

 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). A CanonicalVariate is the weighted sum of the variables in the analysis and it is the best measure of variable importance is the correlation between each variable (𝑋𝑖) and a discriminant function ( , 𝑟𝑥𝑖𝑧𝑗 . It is claimed that these correlations are more informative than standardized coefficients with respect to the joint contribution of the variables to the discriminant functions.

**2-Eigen Values**

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**

**1-Wilk's Lambda ( Λ )**

In **discriminant analysis**, Wilk’s lambda tests how well each level of [independent variable](http://www.statisticshowto.com/independent-variable-definition/) contributes to the model.

Wilk's lambda is the ratio of the within-group sum of squares to the total sum of squares. At each step, the variable that is included in the function is the one with the smallest Wilk's lambda (Λ) after the effect of variables already in the discriminant function is removed.

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

………… (22)

Where is the sum of squares within groups, is the total sum of squares, and 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), and 1 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

………… (23)

Where and 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 () and () 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 regression.

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

………… (24)

**i=1,2,3,…,n**

The standardized variablesare scale free, and the standardized

coefficients,

**Ranking importance of the Variables.**

**2-Canonical Discriminant Function Coefficients**

The correlations are often referred to as loadings or structure coefficients and are routinely provided in many major programs.contains 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**

**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 = ………… (26)

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

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**

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,

, ………… (27)

Expanding about (the sample mean) in a Taylor series, we get

………… (28)

Where 'denotes a remainder containing terms of order . Neglecting , this may be interpreted as the linear function, where

………… (29) , ………… (30)

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

………… (31)

………… (32)

as the reverse Taylor series approximation. The results are easily generalized when 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**

**5.1Data and method of analysis**

The data collected in Erbil (Layla Qasim) diabetic center for 250 patients. Seven Risk Factors for diabetes disease are high blood pressure (hypertension), family history, body mass index, diet, high blood lipid, daily activity and age.

**Dependent (response) variable**

Y: dependent variable is coded in binary response:

Y = 1 if the person has diabetes disease (presence of disease)

Y = 0 if the person has not diabetes disease (absence of disease)

**Explanatory (Independent) Variables** Risk Factors for diabetes disease

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

**Table 1:** Risk Factors for diabetes disease

|  |  |  |  |
| --- | --- | --- | --- |
| **Codes (Categories)** | **Explanatory(Independent) Variables [Risk Factors]** | | |
| (Yes=1,No=0) | high blood pressure (hypertension | | X1 |
| (Yes=1,No=0) | family history | | X2 |
| (Yes=1,No=0) | | body mass index or (Obesity rate) =(Weight/height)2 | X3 |
| (Yes=1,No=0) | diet (nutrition) | | X4 |
| (Yes=1,No=0) | high lipid in blood | | X5 |
| (Yes=1,No=0) | physical activity | | X6 |
| the age from 24 -69 years | age | | X7 |

**5.2 Results**

The results of the study are summarized as follows:

**5.2.1 Reliability Scale: ALL VARIABLES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table (2) Reliability statistics** | | | | |
|  | | | |
| Cronbach's Alpha | N of Items | Image result |
| 0.854 | 7 |

**5.2.2 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 -- even if you’ve specified only a single “block” of variables, as in this case.

Table (3) 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, diet, high blood lipid , physical activity and age] are entered into the equation. Logistic regression compares this model with a model including all the predictors to determine whether the latter model is more appropriate. The table suggests that if we knew nothing about our variables and guessed that a person would be sick with eye glaucoma, we would be correct 90.8% because the predicted result indicated to the overall correct percentage was 90.8% Percentage Correct, which reflects the model’s overall explanatory strength .

**Table 3:** Shows the classification table. Using the obtained Y function observations which are classified as follows, using a prior probability of 0.50

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Observed | | | | | Predicted | | | |  | | Y | | Percentage Correct | |  | | 0 .00 | 1.00 | | Step 1 | Y |  | 0.00 | not sick | 0.00 | 23 | 0.00 | |  | |  | 1.00 | sick | 0.00 | 227 | 100.0 | |  | | Overall Percentage | | | |  |  | 90.8% | |  | |  | | | | | | | | |

B-Regression weights and a test of the H0: b = 0 for the variables in the equation (only the constant for Step1) is provided.

Table (4)**:** 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 odds of having diabetes disease is 9.870. Since 227 of the sampled persons have diabetes disease and 23 are not sick, 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 4: Variables in the Equation** | | | | | | | | |
| Step 1 |  | | β | S.E. | Wald | df | Sig. | Exp(β) |
|  | 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 (5) this table illustrate which variables are not in the equation, it indicates to the variable [X7 (age)] is not in the equation because it is significant (p-value=0.000< 0.05 )and the rest of the risk factors seem to be not important at this step, but the overall significance P-value of 0.000 assures that the logistic model will represent the data very much.

**Table 5: Variables not in the Equation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Step 1 | Variables | | Score | df | p-value |
| X1 | hypertension | 0.019 | 1 | 0.891 |
| X2 | family history | 0.513 | 1 | 0.100 |
| X3 | (BMI) (Obesity rate) | 0.569 | 1 | 0.101 |
| X4 | diet (nutrition) | 0.087 | 1 | 0.768 |
| X5 | high lipid in blood | 1.370 | 1 | 0.166 |
| X6 | physical activity | 0.530 | 1 | 0.467 |
| X7 | 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-**Tests contain [ Chi-squares tests Step, Block and Model] tell us that the

model was improved

Table (6)Omnibus Tests of model coefficients presents the model chi square value of (77.535), with( 7) degrees of freedom, with p-value = 0.000.

**Table 6: Omnibus Tests of Model Coefficients**

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

Table (6) represents Omnibus tests of model coefficients the model

Step chi square value = 77.535, with 7 degrees of freedom, and a p- value = 0.000 which means whether the effect of the variable that was entered in the final step significantly differs from zero.

The block chi-square, 77.535, tests whether every of the variables included in this block have effects that differ from zero.

The model chi-square, 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.

**B-**Two R² values are presented to estimate the fit of the model to the data -- both are transformations of the-2log likelihood values:

table (7) provides the value of (Nagelkerke’s R2 [R-squared gets all of the attention when it comes to determining 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 186.036a  and from the table of Cox & Snell R2 (R-Square) indicating that 66.7% of the variation in the independent variable is explained by the logistic model.

**Table 7:** Model Summary

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

**C- The data fit the model**

**Hosmer and Lemeshow Test**

Table (8) 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 7 degree of freedom is 0.810 as (p-value =0.810 >.05) .  which means that it is not statistically significant and therefore our model is quite good( A non significant chi-square indicates that the data fit the model well).

**Table 8:** 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. The % correct reclassification also improved, by adjusting the cutoff value from the default of(0 .500).  In (Table 9)classification table including the constant term and the rest of the predictors 65.2% were correctly classified for the absence of diabetes disease98.7% were correctly for its presence65.2% from all person direct entry were correctly classified he overall correct percentage was 95.6% which reflects the model’s over all explanatory strength .  **Table 9: Classification 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. The cut value is .500 | | | | | | |

**E-** Interpreting the model: test of the contribution of each parameter

Table 10 is about the variables that are included in the logistic regression equation. This is illustrated in the following equation:

Wald statistic states that the risk factor[ high blood pressure (hypertension) , family history, body mass index(BMI) (Obesity rate) , diet(nutrition) , high lipid in blood, physical activity] are statistically significant factors.

**Table10 : Variables in the Equation**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | β | S.E. | Wald | df | p-value | Exp(β) |
| Step 2 | X1 | hypertension | 0.918 | 0.229 | 16.06 | 1 | 0.002 | 2.505 |
| X2 | family history | 1.821 | 0.643 | 8.063 | 1 | 0.005 | 6. 178 |
| X3 | (BMI) (Obesity rate) | 1.732 | 0.593 | 8.542 | 1 | 0.003 | 5.652 |
| X4 | diet (nutrition) | 0.710 | 0.135 | 26.757 | 1 | 0.000 | 2.033 |
| X5 | high lipid in blood | 1.476 | 0.667 | 4.971 | 1 | 0.026 | 4.375 |
| X6 | physical activity | 1.073 | 0.173 | 38.468 | 1 | 0.000 | 2.924 |
| X7 | age | -0.021 | 0.024 | 0.765 | 1 | 0.251 | 0.979 |
| Constant | | 3.626 | 1.295 | 7.84 | 1 | 0.005 | 37.562 |

Variable(s) entered on step 1: X1, X2, X3, X4, X5, X6, X7 . The last column, 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 is tested by the Wald Chi-Square—testing the null that the β coefficient = 0 (the alternate hypothesis is that it does not β coefficient = 0).

Tests of individual parameters shown on the "variables in the equation table", which Wald test (W=( b / sb.)2, where b is β estimation and sb is its (S.E.) standard error estimation ) that is testing whether any individual parameter equals zero .since the Wald test referred to next is biased under certain situations. When parameters are tested separately, by controlling other parameters.

We see that the effects of (risk factors) variables(X1, X2, X3, X4, X5, X6) are statistically significant, but the effect of(X7) age is not statistically significant.

The value Exp(β) is odds ratio Exp(β) only for(X7) age 0.979 less than 1, implying that the probability the risk of age is less than getting other risks factors this emphasize that but (X7) age is notin the pridiction Equation and

According to the variable (x1) which represents the hypertension, value of Exp(β) is odds ratio equals to (2.505) indicates the probability of the risk hypertension by (2.505) higher.

According to the variable (x2) which represents the family history value equals to (6. 178). which means that every increase in the family history of patient by one, the percentage of risk increases by (0.061).

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

(X2= family history= 6. 178, X3=(BMI) (Obesity rate)=5.652 , X5= high lipid in blood =4.375 , X6= physical activity =2.924 , X1= hypertension =2.505, X4= diet (nutrition)=2.033)

More so, the “β” values are the logistics coefficients that can be used to create a predictive equation.

**Logistic function= Y= 3.626+0.918 X1+1.821 X2+1.732 X3+0.710 X4 +1.476 X5+ 1.073 X6.**

**Logistic function=** Y= 3.626 +0.918 high blood pressure (hypertension) +1.821 family history +1.732 body mass index or (Obesity rate) +0.710diet(nutrition) +1.476high lipid in blood+ 1.073 physical activity.

**5.3 Discriminant Analysis**

Discriminant Analysis contains the following analyses

**a) Development of discriminant functions**

(table11)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 (no sick) as ‘0’ and presence of disease( sick) ‘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. Since in the present research we have used two discrimination variables, one function has been calculated. The function gives the projection of the data that best discriminant between the groups

**Table11:** **illustrate the development of discriminant functions( Group Statistics**)

|  |  |  |  |
| --- | --- | --- | --- |
| Y | | List wise | |
| Un weighted | Weighted |
| 0.00  (No Sick) | X1 | 23 | 23.000 |
| X2 | 23 | 23.000 |
| X3 | 23 | 23.000 |
| X4 | 23 | 23.000 |
| X5 | 23 | 23.000 |
| X6 | 23 | 23.000 |
| X7 | 23 | 23.000 |
| 1.00  Sick | X1 | 227 | 227.000 |
| X2 | 227 | 227.000 |
| X3 | 227 | 227.000 |
| X4 | 227 | 227.000 |
| X5 | 227 | 227.000 |
| X6 | 227 | 227.000 |
| X7 | 227 | 227.000 |
| Total | X1 | 250 | 250.000 |
| X2 | 250 | 250.000 |
| X3 | 250 | 250.000 |
| X4 | 250 | 250.000 |
| X5 | 250 | 250.000 |
| X6 | 250 | 250.000 |
| X7 | 250 | 250.000 |

In table (12) The maximum number of discriminant functions produced is the number of groups minus 1, since only two groups used here, namely ‘sick’ and ‘no sick’, so only one function is displayed. A larger Eigen value explains a strong function 0.373 abut here the value of Eigen value indicates to weak function. The canonical relation is a correlation between the discriminant scores and the levels of these dependent variables. The higher the correlations value, the better the function that discriminates the values. 1 is considered as perfect. Here, we have the correlation of (0.521) measures the strength of relationship is low. With only one function it provides an index of overall model fit which is interpreted as being the proportion of variance explained (R2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table (12):represent Eigenvalues 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 (13) refer to tests of equality of group means by testing hypothesis regarding discriminating power of the variables Wilk’s Lambda

Wilk’s Lambdais the standard statistic that is used to denote the statistical significance of the discriminatory power of the current model. Its value will range from ( 1.0) (no discriminatory power) to (0.0) (perfect discriminatory power).

H**0** : There is no significant discriminating power in the variables.

H**1** : There may be a significant discriminating power in the variables

The value of Wilks' Lambda is 0.728 with Chi-square(77.487 ) p-value (0.000) it mean H**0** is rejected

**Table (13)** **Tests of Equality of Group Means**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test of Function 1 | Wilks' Lambda | Chi-square | df | p-value |
| 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 tableis 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 (14) has the highest discriminating power due to the highest discriminant coefficient of 0.883 followed by body mass index(BMI) (Obesity rate) this indicates that X1 the body mass index(BMI) (Obesity rate) has a first predictor of risk factors **,** X2 family history has discriminant coefficient by 0.458….ect

**Table 14: Standardized Canonical Discriminant Function Coefficients**

|  |  |  |
| --- | --- | --- |
| Variables | | **Standardized Canonical Discriminant**  **Function Coefficients** |
| X1 | hypertension | 0.146 |
| X2 | family history | 0.458 |
| X3 | (BMI) (Obesity rate) | 0.883 |
| X4 | diet (nutrition) | 0.077 |
| X5 | high lipid in blood | 0.307 |
| X6 | physical activity | -0.081 |
| X7 | age | -0.113 |
| Constant |  | 8.820 |

**Ranking importance of the Variables.**

Theranking in (Table 15) Based on the coefficients in (Table 14) for the relative important predictor variables summarized as follow:

**Table 15: Ranking of the Variables**

|  |  |  |
| --- | --- | --- |
| **Ranking of the Variable** | | **Predictor Variable** |
| X3 | body mass index | 0.883 |
| X2 | family history | 0.458 |
| X5 | high lipid in blood | 0.307 |
| X1 | high blood pressure (hypertension) | 0.146 |
| X7 | age | -0.113 |
| X6 | physical activity | -0.081 |
| X4 | diet (nutrition) | 0.077 |

In (Table 16) Canonical Discriminant Function Coefficients represent the Coefficients of final Canonical Discriminant Function

|  |  |  |
| --- | --- | --- |
| **Table 16:Canonical Discriminant Function Coefficients** | | |
| Variables |  | Function 1 |
| X1 | hypertension | 0.319 |
| X2 | family history | 1.021 |
| X3 | (BMI) (Obesity rate) | 0.276 |
| X4 | diet (nutrition) | 0.163 |
| X5 | high lipid in blood | 0.647 |
| X6 | physical activity | -0.162 |
| X7 | 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 classification**

**Table 17:** **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.

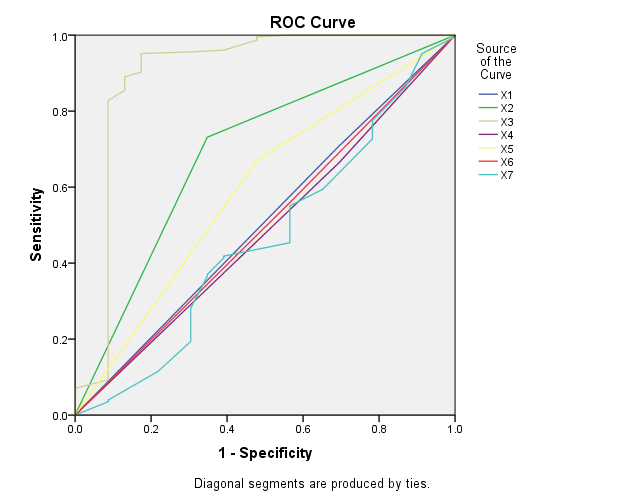
Table (Table 17) indicates that the LRA model can correctly classify the first category

No sick with accuracy of step0 (65.2 %) compared (13.6 %) for LDA model .

At the second category sick, 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 18: Classification Result of Two Statistical Methods LRA and** LDA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | **Percentage** | |
| **Observed** | | | **Predicted** Y | |
| **LRA** | LDA |
| **Y** | **0.00** | **No Sick** | 65.2% | 13.6% |
| **1.00** | **Sick** | 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 19)The test result variable(s): X1, X2, X3, X4, X5, X6, X7 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 19: Area under the Curve of all variables**

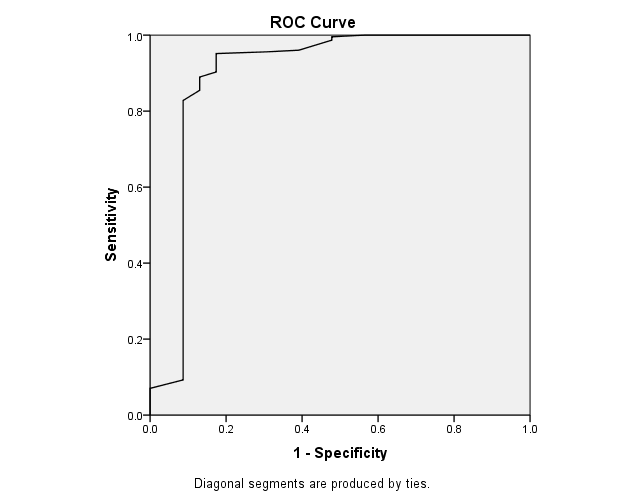
|  |  |  |
| --- | --- | --- |
| Test Result Variable(s) | | Area |
| X1 | hypertension | 0.507 |
| X2 | family history | 0.692 |
| X3 | (BMI) (Obesity rate) | 0.897 |
| X4 | diet (nutrition) | 0.485 |
| X5 | high lipid in blood | 0.596 |
| X6 | physical activity | 0.497 |
| X7 | age | 0.463 |

**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 age0.463

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 18)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 20: 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**

The study contained two parts, theoretical part and application part in which data was used from experimental group of 250 diabetic data collected in Erbil diabetic center (Layla Qasim) it contained (8) variables, where the dependent variable was diabetes disease (presence and absence) of disease and (7 ) predictor( independent) variables are taken in the model, like[ Hypertension, family history, body mass index(BMI), diet, high blood lipid and physical activity, age].Before starting the analysis the data was tested for Reliability by Cronbach’s alpha which it is a way of measuring the strength of that consistency and scale of all variables illustrate Reliability statisticsit resultwas0.854. they are two of the most widely used statistical methods for analyzing categorical outcome variables .

The effectiveness of the logistic model was shown to be supported by significance tests and descriptive and inferential goodness-of-fit indices, and predicted probabilities. LDA was conducted to predict whether in analyzing by the Logistic Regression Analysis contains two Steps [(Step 1) and(Step 2) ]

In Step 1 [this step is to compute and enter just the constant ] ,the classification 90.8 % cases correctly classified, predicted odds [Exp (β)] = 9.870 and Wald statistic is computed(109.466) with rejected P-value.= 0.000, indicating that the constant does make a significant contribution to the model, The contribution of each predictor illustrate which variables are not in the equation, it indicates to the variable [X7 (age)] is not in the equation because it is significant (p-value=.000< 0.05 ).In step 2 tests the contribution of all the variables entered the Omnibus Tests of model coefficients presents the model chi square value of (77.535), with( 7) degrees of freedom, with p-value = 0.000.Two R² values are presented to estimate the fit of the model to the data (Nagelkerke’s R2 is 0.781, indicating a strong relationship between the predictors and the prediction, and Cox & Snell R2 (R-Square) indicating that 66.7% of the variation in the independent variable is explained by the logistic model. illustrates Hosmer- Lemeshow (H-L) test. 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. According to value of odds ratio we can ranking the risk factors on diabetes as follows : (X2= family history= 6. 178, X3=(BMI) (Obesity rate)=5.652 , X5= high lipid in blood =4.375 , X6= physical activity =2.924 , X1= hypertension =2.505, X4= diet (nutrition)=2.033)

By Linear Discriminant Analysis(LDA) the Canonical Correlation(0.521) ,Wilk’s Lambda**(**0.728**)** is statistical significance of the discriminatory power of the current model with [p-value](http://www.statisticshowto.com/p-value/) =0.000**,** the classification table shows the accuracy decision prediction( step 1) 85%, By(LRA) and (LDA) ranking [body mass index, family history, high lipid in blood, high blood pressure (hypertension), age, physical activity and diet (nutrition).

Over all explanatory strength with Logistic Regression Analysis (LRD) by reclassification table shows the accuracy of the model classification table including the constant term and the rest of the predictors the overall correct percentage was 95.6% which reflects the model’s Over all explanatory strength and with Linear Discriminant Analysis(LDA) the overall correct percentage was 84% .The prediction powers of the models by the area under ROC curve represent as follow [BMI (Obesity rate)**,** family history**,** high lipid in blood and hypertension]. Area under ROC curve reached 89.7%, and (Asymptotic p-value = 0.000) ROC curve was used to compare prediction powers of the models. So to specify a model with the highest prediction the results illustrated that Logistic Regression Analysis (LRA) was better than Linear Discriminant Analysis (LDA) .

**7. Recommendations**

According to the conclusion reported above we may recommend the followings

1- Conducting extensive studies on expanding the sample size further to cover more and more aspects of the diabetes diseases.

4- We recommend to use also Multinomial Logistic Regression methods in classifying and

prediction technique in other fields**.**

**References**

[1] Agresti, A. A.. An introduction to categorical data analysis. New York : John Wiley & Sons, 1996, pp. 156-178.

[2] Agresti, A.. Categorical data analysis, 2nd edition. New York : John Wiley & Sons, 2000, pp. 52-143.

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

[4] 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. (Retrieved on 6/11/2013)

[5]Dr.P.Venkatesh (2012) Discriminant Analysis: An OverviewScientist, Division of Agricultural Economics, IARI

[6] El-Habi, A., & El-Jazzar (2014). A Comparative Study Between Linear Discriminant Analysis and Multinomial

[7] 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.

[8] Hauck, W.W. and A. Donner. "Wald’s Test as applied to hypotheses in logit analysis". Journal of the American Statistical Association, vol. 72, pp. 851-853, 1977.

[9] Hosmer, D. W. and Lemeshow, S. Applied logistic regression, 2nd edition. New York : John Wiley & Sons, 2000, pp. 79-125.

[10] Maja Pohar1, Mateja Blas2, and Sandra Turk3 Wuensch, Karl L. (2014)BinaryLogisticRegression with SPSS. Retrieved from http://www.care.ecu.edu/psyc/wunesch/MV/Multerg/logistic\_SPSS

[11] Metodološki zvezki, Vol. 1, No. 1, 2004, 143-161Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study

[12] Marzia Cescon(2013 )"Modeling and Prediction in Diabetes Physiology" PhD ThesisISRN LUTFD2/TFRT--1099—SE ISSN 0280–5316 Department of Automatic Control Lund University Box 118 SE-221 00 LUND Sweden Cf All rights reserved.

Printed in Sweden by MediaTryck.

[13] Nagelkerke, N.J.D. "A note on the general definition of the coefficient of determination". Journal of Biometrika, vol. 78, pp. 691-692, 1991.

[14] Parastoo RAHIMLOO, Ahmad JAFARIAN,(2016) Prediction of Diabetes by Using Artificial Neural Network,Logistic Regression Statistical Model and Combination of **Them,**Department of Mathematics, Urmia Branch, Islamic Azad University, Urmia, Iran Bulletin de la Société Royale des Sciences de Liège, Vol. 85, p. 1148 – 1164 ,1148

[15] Pickup, J.C., Williams G., (2003) Textbook of diabetes, Blackwell Science, Oxford,

[16]Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio et al. Lancet. 2010; ) [Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Emerging Risk Factors Collaboration](http://www.ncbi.nlm.nih.gov/pubmed/20609967) 26;375:2215-2222

[17]S. James Press; Sandra Wilson, Choosing Between Logistic Regression and Discriminant Analysis , Journal of the American Statistical Association, Vol. 73, No. 364. (Dec., 1978), pp. 699-705.

[18]Sernyak, Michael J., et al.( 2014) "Association of diabetes mellitus with use of atypical

neuroleptics in the treatment of schizophrenia", American Journal of Psychiatry,.

[19] Thirugnanam, Mythili, et al.( 2012) "Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach."Procedia Engineering, Vol.38, pp.1709-118,.

[20] T. Ramayah1\*, Noor Hazlina Ahmad1, Hasliza Abdul Halim1, Siti Rohaida Mohamed Zainal and May-Chiun Lo (2010) Discriminant analysis: An illustrated example

[21]Wen-Chih Wu1,2,,Mary E. Lacy,1,2 Gregory A. Wellenius,1Mercedes R. Carnethon,3 Eric B. Loucks, April P. Carson, Xi Luo, Catarina I. Kiefe,Annie Gjelsvik, Erica P. Gunderson, Charles B. Eaton,1 and 8Care 2016;Racial Differences in the Performance of Existing Risk Prediction Models for Incident Type 2 Diabetes: The CARDIAStudy

39:285–291 | DOI: 10.2337/dc15-0509

[22] Yusuff, N. M., Ngah, U. K., & Yahaya, A. S. (2012). Breast Cancer Analysis Using Logistic Regression. IJRRAS,10(1), 14-22.

**ثوختة**

زؤر لة تويَذينةوة ثزيشكيةكان ئاماذة ئةدةن بة بووني ثةيوةندييةكي بةهيز لةنيَوان دةست نيشان كردني نةخؤشي وهةندىَ شيكاري ئاماري وةك شيكردنةوةى لؤجستي Logistic Regression Analysis وشيكردنةوةي جودا كةريLinear Discriminant Analysis (LDA). كةهةردوكيان رِيَطاي شيكردنةوةي فرة طؤرِاون (multivariate statistical methods) بة شيَوةيةكي بةرفراوان بةكار ئةهيَنريَن لةشيكردنةوةي ي بؤ ثيَش بيني كردن ( predict)

لةم تويَذينةوةيةدابؤ هةردووشيكردنةوةكة ئةنجامدرالةسةركؤمةلَةداتايةكي تايبةت بةنةخؤشى شةكرة كةذمارةيان( 250 )كةسةكة لةناوةندي(لةيلا قاسم)ي نةخؤشي شةكرةي هةوليَر وةرطيراوة ثيَكهاتووة لة(8 ) طؤرِاوو(1) يةكيَكيان طؤرِاوي ثشت بةستوو Dependent variable) ) كة (بوون) و(نةبووني) نةخؤشي شةكرة دةنويَنىَ وة(7) طؤرِاوةكةي تركةطؤرِاوي ثيَشبينيكراو(( predictorيان طؤرِاوي سةربةخؤ( independent) دةنويَنن كةبريتين لة هؤكاري مةترسيدار بؤتوش بووني نةخؤشي شةكرة وةك (تةوذمي خويَن , بؤماوةي خيَزاني , بارستايي لةش(ريَذةي قةلَةوي) , جؤري خؤراك (ضةور وبىَ ضةور) , رِيذةي ضةوري لةخويَندا , ضالاكي رؤذانة(وةرزش) وة تةمةن)

ئامانجي ئةم تويََذينةوةيةدياريكردني باشترين مؤديلي ئاماري بؤ ثيَش بينيكردني هؤكارةترسناكةكان لةسةر نةخؤشي شةكرة بةرِيَطاي هةريةك لة ريَطاي Logistic Regression Analysis (LRA))وة Linear Discriminant Analysis (LDA) كؤمةلَةتاقيكردنةوةيةكي تايبةت بةهةردوو رِيَطاكةئةنجامدرا. ئةم تويََذينةوةيةلة دووبةش ثيَكهاتووة بةشي تيؤري

وبةشي كرداري لةسةرةتاي شيكردنةوةكاندا (reliability)َكردن بة داتاكةئايا طونجاوة بؤ شيكردنةوةي ئاماري بةرِيكاي (Cronbach's Alpha) تاقيكرايةوة لةئةنجامدا ريَذةي (0.854%) دةرضوو كة ئةنجاميَكي باشة ماناي داتا كة باوةرِثيَكراوة بؤ شيكردنةوة .

ثاشان هةردوورِيَطاي Logistic Regression Analysis (LRA) وة Linear Discriminant Analysis (LDA) بةكارهيَنرا بةكؤمةلَةتاقيكردنةوةيةكي تا يبةت بةهةريةك لة دوو رِيَطاكةلة ئةنجامدا دةركةوت Logistic Regression Analysis (LRA) كة طؤرِاوي تةمةن كاريكةري خراثي نية بةلَام (6)طؤرِاوةكةي تركاريطةري خراثيان هةية. وةهةروةهابؤLogistic Regression Analysis (LRA) بةثيَي كؤتا classification table نرخي برِياري رِيَكي وثيَكي مؤديَلي ثيَشبيني كراو accuracy decision prediction) ) دةكاتة95.6% . وةبؤ Discriminant Analysis (LDA) Linear بةثيَي كؤتا classification table نرخي برِياري رِيَكي وثيَكي مؤديَلي ثيَشبيني كراو accuracy decision prediction دةكاتة84% بةثيَي ئةم دوونرخةجياوازيةكي بةرضاو دياري دةكريَت.

كةواتةبةثيَي ئةنجامي هةموو جؤرةتاقيكردنةوةكاني هةردووجؤرة شيكردنةوةكة Logistic Regression Analysis (LRA) وة Discriminant Analysis (LDA) Linear ئةتوانين بلَيَين كةبؤ باشترين ثيَشبيني كردن شيكردنةوة بةرِيَطاي Logistic Regression Analysis (LRA) باشترة لة شيكردنةوة بةرِيَطاي Discriminant Analysis (LDA) Linear بؤثيَشبيني كردني هؤكارةترسناكةكان بؤ تووش بوون بة نةخؤ شى شةكرة وةك لةرِيَطاي Logistic Regression Analysis (LRA) دةركةوت هؤكارةترسناكةكان بريتين : 1-بؤماوةي خيَزاني=, 6. 178 2- بةرزي رِيَذةي قةلةوي =5.652 و3-بةرزي رِيَذةي ضةوري لةخويَندا=4.375و 4- وةرزش = 2.924وبةرزي تةوذمي خويَن=2.505و خواردن=2.033)

وة بةثيَي رِووبةري ذيَر ضةماوةي (ROC Curve) روبةرةكان بةثيَي طةورةيي رِووبةرةكة (1-بةرزي رِيَذةي قةلةوي= 0.897 و 2-بؤماوةي خيَزاني=0.692 و 3-بةرزي رِيَذةي ضةوري لةخويَندا=0.596 و4-بةرزي تةوذمي خويَن =0.507 و 5- وةرزش=0.497و 6- خواردن=0.485 ثاشان تةمةن كة دةكةويَتة ذيَرهيَلَي جياكةرةوةي( ROC Curve) كة طرنطييةكي ئةوتؤي نية.