Research Paper E-ISSN No : 2455-295X | Volume : 2 | Issue : 4 | April 2016

DISCRIMINANT ANALYSIS OF DIABETES PATIENTS DATA

Prof. dr. Hana M. Al-Aukyli [1] | Nabaa Mohammed Al-shamary [1]

Faculty of Computer Sciences and Mathematics, University of Kufa.

ABSTRACT

This paper consisted of an application of Discriminant Analysis (DA) which is a multivariate analysis of variance (MANOVA) where the independent variables are the predictors and the dependent variable are the groups. DA is useful for situations like diabetes patients data to determine and predict that the patient like to have diabetes type 1 or type 2. A sample of (42) patients were chosen from Al-Sader hospital in Najaf city in Iraq with (17)personal characteristics of the patients which are representing the independent variables or predictor variables were used as liner combinations to provide the best discrimination between the groups. SPSS package were used to achieve the calculation of the Analysis in four steps .The most important results is the use of simple Discriminant method on diabetes data to classify the patients into two groups type1 and type2.

1-INTRODUCTION:

This paper is an application of Biostatistics method -Discriminant analysis on Diabetes mellitus. A sample of size (42) diabetes were chosen from "Al-Najaf Center for Endocrine" in Iraq , randomly. Linear Discriminant analysis(DA) was used to Discriminate two groups of diabetes patients with variables representing information and symptoms of the disease .Thought out the DA the factors influencing Diabetes used a stepwise regression analysis.

The linear Discriminant function, used in this paper assuming the population is multivariate normal distribution with equal covariances.

The main contribution of this paper is the use of simple Discriminant method on diabetes data to classify the patients into two groups type1 and type2 diabetes patients .

Aim of the research:

Discriminant Analysis for diabetes data were used to determine memberships

2-Theoretical Background:

Discriminant analysis (D A)[7]:

DA is one of the important statistical methods in classification with statistical analysis of multiple variables that are interested in differentiating between two or more groups which are similar in most characteristics or variables.

D A was first introduced by R. A. Fisher in 1936 to classify sample into two groups with equal covariance[6], It was started the idea of using the Discriminant analysis for multivariate population . Also discussed by Gillbert in 1969 the effect of different covariances and variances matrices to study of influencing factors in the disease of the nervous system with the children under siege by 1999 hama[9].

Discriminant Function (D F[7]):

Discriminant function is a multivariate analysis of variance(MANOVA) reversed "in which the independent variables are a set of variables and the dependent variables are predicted.

The number of D F computed is one less than the number of groups in the dependent variable.

$$y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_pX_p \quad (1)$$

Where **y** is the discriminant function

is discriminant coeffient or weight αi for the variable

is a constant α_0

I is the number of independent variables

X is respondent's score for that variable

Assumption of discriminant analysis:

A required assumption for the discriminant analysis are:

1-AS a rule the sample size (n) of the smallest group should exceed the number of independent variables i.e $n_1+n_2 -2 \geq p$.

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

3-Homogeneity of variance / covariance: Variance among group variable are the same as cross levels of predators. Can be test with Box's M statistic. The liner Discriminant analysis can be used when covariance equal, and that quadratic discriminant analysis may be used when covariance are not equal.

4- Outliers: DA is highly sensitive to the inclusion of outliers. Run a test for univariate and multivariate outliers for each group, and transform or eliminate them.

There are several step in finding the (D F):

Step 1: (Test of significance)[8]

A-Test of hypotheses about mean vector (unknown covariance matrix)[8]:

In multivariate we use (T^2 Hotelling). If we have two group, assume (i=1,2) As a random variable of x_i where normal distribution $x_i \sim N(M_i,\Sigma)$, then to test the hypothesis

$\mathbf{H_0: M_1=M_2}$

$\mathbf{H_1: M_1 \neq M_2}$

The estimation of the covariance matrix:

$$S = \frac{1}{n_1+n_2-2}[(n_1-1)S_1{}^2 + (n_2-1)S_2{}^2 \quad (2)$$

$$T^2 = \frac{n_1 n_2}{n_1+n_2} * D^2 \quad (3)$$

Where $D^2=(\bar{X}_1-\bar{X}_2)'S^{-1}(\bar{X}_1-\bar{X}_2)$ (4)

$$F = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p} * T^2 \quad (5)$$

Degree of freedom (p, n_1+n_2-p-1)

P is the number of independent variable

If $\mathbf{F_{cal} > F_{tab}}$ accept $\mathbf{H_1}$

If There are more than two group. We should be used wilks' lambda , The hypotheses:

$H_0: M_1=M_2=M_3=\cdots=M_K$

$H_1: M_1 \neq M_2 \neq M_3 \neq \cdots \neq M_K$

Wilks' Lambda : $\Lambda = \frac{|W|}{|T|}$ (6)

Where

T:common variation Matrix and contrast overall collections

W: variation and co-variation Matrix within groups

Where the range of $\Lambda \in [0,1]$. When v_H=1 or 2 or when p=1 or 2 wilk's Λ transforms to exact F-statistic.The transformations from Λ to F for these special case are given in Table1. The hypothesis is rejected when the transformed value of Λexceeds.

Table1:

Parameters p, v_H	Statistics having F-distribution	Degree of freedom
Any p, v_H = 1	$\frac{1-\Lambda}{\Lambda} * \frac{v_E - p + 1}{p}$	$p, v_E - p + 1$
Any p, v_H=2	$\frac{1-\sqrt{\Lambda}}{\Lambda} * \frac{v_E - p + 1}{2p}$	$2p, v_E - p + 1$
P=1 any v_H	$\frac{1-\Lambda}{\Lambda} * \frac{v_E}{v_H}$	v_H, v_E
P=2 any v_H	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} * \frac{v_E - 1}{v_H}$	$v_H, v_E - 1$

For value of p,v_H other in Table, an approximate F-statistic is given by

$$F = 1 - \frac{\Lambda^{1/t}}{\Lambda^{1/t}} * \frac{df_2}{df_1} \quad (7)$$

With df_1 and df_2 are degrees of freedom, where

$df_1 = pv_H$

df_2=w-0.5(pv_H-2), w = v_E+v_H-0.5 (p+ v_H^{-1} and t = $\sqrt{\frac{p^2 v_H^2 - 4}{p^2 + v_H^2 - 5}}$

Where v_E = degree of freedom for hypothesis

v_H =degree of freedom for error Number of independent

p= variable(dimension)

B-(Multivariate TEST Of Equality of covariance matrices):
For g multivariate population, the hypothesis of equality of covariance matrices is $= \Sigma_2$ *V.S* $H_1 : \Sigma_1 \neq \Sigma_2$ $H_0 : \Sigma_1$

Box's Test for Equality of Covariance matrices

$$u = \left[\Sigma_i \frac{1}{n_i - 1} \frac{1}{\Sigma_i (n_i - 1)}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right]$$

$C = (u-1)\{[\Sigma_i (n_i-1)] \ln|S_{Poold}| - \Sigma_i (n_i-1) \ln|S_i|\}$ **(9)**

where p is the number of variables and k is the number of groups has an approximate χ^2-distribution with

v= g*0.5 p(p+1)-0.5 p(p+1)

=0.5 p(p+1)(g-1)

At significance level

α, reject H_0 if **$C > \chi^2$ P(P+1)(g-1)/2(α).**

Box's χ^2 Appoximation works well if each n_i exceeds 20 and if p and g do not exceed 5. In situations where these conditions do not hold, Box ([4],[5]) has provided a more precise F approximation to sampling distribution of M.

Step2: (Interpretation)[8]
D.F. interpretation by means of standardized coefficients and the structure matrix which are the correlations between the variable in the model and the discriminant functions .

This will give the Canonical Correlation Analysis and canonical roots. Then the highest discriminant classification can be known for each patient .

Step3: (Classification using the discriminant function)[2]
There are us e several method to classification. We will used (Midpoint), Where are consider of the best method because it will be reduce the misclassification ratio less than

(10) $\bar{y}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_1$

(11) $\bar{y}_2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_2$

(12) *If* **y>c** *the new sample is* $C = \frac{\bar{y}_1 + \bar{y}_2}{2}$

belong to first population.

And

If y<c the new sample is belong to second population.

Where y is the discriminant function of sample

Diabetes mellitus[1][3]:
Diabetes mellitus is a group of metabolic diseases characterized by elevated blood glucose levels(hyperglycemia) resulting from defects in insulin secretion, insulin action or both. Insulin is a hormone manufactured by the beta cells of the pancreas, which is required to utilize glucose from digested food as an energy source. Chronic hyperglycemia is associated with microvascular and macrovascular complications that can lead to visual impairment, blindness, kidney disease, nerve damage, amputations, heart disease, and stroke. In 1997 an estimated 4.5% of the US population had diabetes.

There are two types of diabetes. In type 1 diabetes a person's body does not make enough insulin to help move glucose into the cells for energy. In type 2 diabetes a person's body does not use insulin effectively and over time will not make enough insulin. Type 1 diabetes typically happens to people under the age of 30 and cannot be prevented. Type 2 diabetes can be prevented.

Results:
Discriminant analysis is useful for situations in which you want to build a predictive model of group membership for "Diabetes mellitus" and use SPSS V(20)

The data collected from " Al-Najaf Center for Diabetes and Endocrine "for (42) patient, (15) patient of them have type 1and (27) patient type2. The data was obtained from a questionnaire filled in by patients.

The dependent variable is type of disease(type1 and type2) and predictor variables are shown in Table 2. DA using SPSS (Statistical package for social sciences) software .The variables in the questionnaire where coded as in table 2:

Table 2: The coding of the independent variables

Variable	Coding
Age in years	Reported by the patient's
Gender	1 : male 2 :female
Weight (in KG)	Reported by the patient's Weight
Highets (in CM)	Reported by the patient's Length
Income	1: Non Enough 2: Enough to some extent 3: Enough
Place of Living	1: Ruler 2: Urban
Marital status	1: Unmarried 2: Separated 3: Widowed 4 :Married
Genetics	1 :from father side 2 :from mother side 3: both father and mother side
Drug Type	1: injection needles 2: oral 3: both injection and oral
Sugar before breakfast	Reported by the patient's Sugar before breakfast
Extreme thirst	1 :No 2: To some extent 3: Yes

Poly urea	1 :No 2: To some extent 3: Yes
Hunger	1 :No 2: To some extent 3: Yes
Blurred vision	1 :No 2: To some extent 3: Yes
Weight loss	1 :No 2: To some extent 3: Yes
Tired	1 :No 2: To some extent 3: Yes

SPSS Application:
DataView→Analysis→Classify→Discriminant Analysis→Use stepwise method.

- By clicking on button Grouping Variable to input dependent variable and click Define Range to inter 1 for Minimum, 2 for Maximum.
- By clicking on button Statistics to Selection:Mean, Box's M, Univariate ANOVs and Unstandardized.
- By clicking on button Method to select wilk's lambda. - By clicking on button Classification to select All groups equal and Summary table.

Steps of Finding DF:

Group Statistics

Type	Mean	Std. Deviation	Valid N (listwise)		
			Unweighted	Weighted	
Type 1	Age	27.13333	14.759339	15	15.000
	Sex	1.40000	.507093	15	15.000
	Weight	58.73333	20.693224	15	15.000
	Length	158.46667	13.968672	15	15.000
	Income	2.06667	.961150	15	15.000
	Place of Living	1.80000	.414039	15	15.000
	Marital status	2.20000	1.521278	15	15.000
	Gantics	.80000	.861892	15	15.000
	Diagnosis	2.60000	.828079	15	15.000
	Type drug	1.13333	.516398	15	15.000
	Sager before breakfast	180.33333	91.991200	15	15.000
	Extreme thirst	2.66667	.723747	15	15.000
	Poly Urea	2.66667	.723747	15	15.000
	Hunger	2.60000	.736788	15	15.000
	Blurred vision	2.20000	1.014185	15	15.000
	Weight Loss	2.33333	.975900	15	15.000
	Tired	2.53333	.833809	15	15.000
Type 2	Age	53.55556	10.024329	27	27.000
	Sex	1.55556	.506370	27	27.000
	Weight	78.74074	27.087597	27	27.000
	Length	157.66667	34.592129	27	27.000
	Income	2.48148	.700020	27	27.000
	Place of Living	1.85185	.362014	27	27.000
	Marital status	3.66667	.832050	27	27.000
	Gantics	.85185	.988538	27	27.000
	Diagnosis	2.77778	.640513	27	27.000
	Type drug	1.85185	.863967	27	27.000
	Sager before breakfast	221.44444	70.519683	27	27.000
	Extreme thirst	2.22222	.974022	27	27.000
	Poly Urea	2.44444	.891556	27	27.000
	Hunger	2.66667	.733799	27	27.000
	Blurred vision	2.55556	.800641	27	27.000
	Weight Loss	2.18519	.962250	27	27.000
	Tired	2.55556	.800641	27	27.000

Step 1: Test of significant
A-Tests Significant of the hypothesis:

$H_0: M_1 = M_2$

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.189a	100.0	100.0	.737

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.457	30.944	1	.000

Thus we see that there are one function in the analysis to explains 100% of the variance that has wilk's lambda 0.457, a chi-square is 30.944 and significant is 0.0001.Then accept H_1 since p<0.05 discriminant function has the ability to classily.

B-Test of equality of covariance matrices:
The hypothesis of these test is
$= \Sigma_2 V.S\ H_1 : \Sigma_1 \neq \Sigma_2\ H_0 : \Sigma_1$

Test Results

Box's M		2.876
F	Approx.	2.798
	df1	1
	df2	3746.706
	Sig.	.094

Accept H_0 since p>0.05. About this results is linear discriminant function.

C-Test significant of ANOVA:

	Wilks' Lambda	F	df1	df2	Sig.
Age	.457	47.556	1	40	.000
Sex	.978	.909	1	40	.346
Weight	.867	6.158	1	40	.017
Length	1.000	.007	1	40	.932
Income	.939	2.585	1	40	.116
Living	.996	.179	1	40	.675
Marital status	.708	16.463	1	40	.000
Gantic	.999	.029	1	40	.866
Diagnosis	.985	.602	1	40	.443
Type drug	.823	8.605	1	40	.006
Sager before breakfast	.938	2.631	1	40	.113
Extreme thirst	.944	2.381	1	40	.131
Poly Urea	.983	.680	1	40	.414
Hunger	.998	.079	1	40	.780
Blurred vision	.962	1.570	1	40	.218
Weight Loss	.994	.226	1	40	.637
Tired	1.000	.007	1	40	.933

A bout this table '"Test of Equality of Group Means", It is shown that the independent variables(Age, Weight , Marital status and Type of the drug) are significant According to F-test at ($p \leq 0.05$). The age variable is the most significant.

Step2:Interprtation
A-The use of the stepwise statistics find out which variable that enter in discriminant analysis

Variables in the Analysis

Step	Tolerance	F to Remove	
1	Age	1.000	47.556

In conclusion, one variable (Age) inter the discriminant function.

B-Standardized Canonical Discriminant Function Coefficients:

	Function
	1
Age	1.000

It is clear that the only variable "Age" is important variable entered within the analysis.

C-Correlations:
Structure matrix is the pooled within groups correlation between discriminating variables and discriminant functions.

Structure Matrix

	Function
	1
Age	1.000
Mariatal status[a]	.340
Blurred vision[a]	.203
Tired[a]	-.179-
Sager before breakfast[a]	-.149-
Living[a]	.141
Sex[a]	.095
Hunger[a]	.079
Poly Urea[a]	-.075-
Length[a]	.073
Extreme thirst[a]	-.067-
Type drug[a]	.063
Income[a]	-.056-
Diagnosis[a]	-.053-
Weight Loss[a]	-.051-
Gantica[a]	.024

D-A further way of interpreting discriminant analysis results is to describe each group in terms of its profile, using means of the predictor variables. These are the group means -1.428 for type 1 while type 2 produce a mean of 0.793. So these table is means of groups.

Functions at Group Centroids

Type	Function
	1
Type 1	-1.428-
Type 2	.793

Unstandardized canonical discriminant functions evaluated at group means

So, the discriminant function is:

y=0.084-3.708Age

Step 3: Classification Results

Y			Predicted Group Membership		Total
			Type1	Type2	
Original	Count	Type1	13	2	15
		Type2	3	24	27
	%	Type1	86.7	13.3	100.0
		Type2	11.1	88.9	100.0

The ratio of persons classified within the original groups is 88.1%. where the correct sample in original group is 37 of 42.

Apparent error rate of type 1: 2/13

And apparent error rate of type2: 3/24

Conclusion:
By Applying the SPSS package V20

for Discriminant Analysis, the following results are concluded:

1-Through the tests :The data applied were resulted a normal distribution with an equal covariance .

2-The Application of forward Regression Analysis the only extracted variable is the age of the Patient.

3-There is no conformity classification with original data for" Al-Najaf Center for Diabetes and Endocrine" patients at age less than 30 years were classified in type1. Where asby the Discriminant function classified the patients at age 32 years only other wise the discriminant function is classified in type2.

REFERENCES:

1. American Diabetes Association (ADA): Type-2 diabetes in children and adolescents, Diabetes Care, 2008, p 381-389.
2. Anderson, T.W. "An Introduction to Multivariate Statistical Analysis (3rd ed.). New York: John Wiley, 2003.
3. Atkinson MA. And Eisenbarth GS, Type 1 diabetes, new prospective on disease pathogenesis and treatment, lancent,2001. P 358.
4. Box, G.E. P., "Problems in the Analysis of Growth and Wear Curves" Biometrics, 6(1950), 362-389.
5. Box, G. E. P.," A General Distribution Theory for a Class of Likelihood Criteria." Biometrika, 36 (1949), 317-346
6. Fisher,R.A."The Statistical Utilization of Multiple Measurements." Annals of Eugenics,8(1938),376-386.
7. Richard A. Johnson and Dean W. Wichern"Applied Multivariate Statistical Analysis" Discrimination and Classifiction,11(2007),575-670.
8. Rencher Alvin C. ,"Method of Multivariate Analysis. "John Wiley &Sons(2002),117-295.