Individual Project report on Discriminant Analysis

By Jayachandu Bandlamudi

Project Title: ***Analyzing heart disease using patient medical records.***

Team details: Project Group -1

S.no	Name	Net ID
1	Jayachandu Bandlamudi	bandlmd2
2	Seunghyun Oh	soh45
3	Augustine Chiu	achiu9
4	Daulet Dyussekenov	dyussek2

Contents:

1) Introduction:

For the final project, our group choose to work on dataset related to heart disease from health sciences. And reason being, we know that coronary heart disease is a major cause for many deaths across the globe and the main objective of the project is to analyze several factors that influence and helps in determining the heart disease. We obtained the dataset from the 'Elements of Statistical learning' text book and it can found online at Stanford website.

[URL: http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data]

Dataset is a sample of medical records related to the Male patients from the heart-disease high-risk region of the Western Cape, South Africa. At a high level, these medical records have patient's information related to Age, Drinking habits, Smoking habits, Eating habits, Family history and Personality traits etc.

2) Data description:

Dataset has 462 observations, and each observation corresponds to individual patient record. It has 10 variables that have information related to the patient, and 8 are continuous variables, 2 variables are categorical. The 8 continuous Variables are as follows '**sbp**' (systolic blood pressure), '**tobacco**' (cumulative tobacco in kg), '**ldl**' (low density lipoprotein cholesterol), '**adiposity**', '**typea**' (Personality trait that is characterized by excessive competitiveness and aggression) which is score, '**obesity**', '**alcohol'**, '**age**'. And categorical variables include '**famhist**' (family history of heart disease) with two levels i.e.; Present/Absent, '**chd**' (coronary heart disease) with two levels i.e.; 1 – if the patient has the disease and 0 otherwise.

3) Descriptive statistics for variables:

The UNIVARIATE Procedure
Variable: sbp

Moments			
N	462	Sum Weights	462
Mean	138.32684	Sum Observations	63907
Std Deviation	20.4963172	Variance	420.099018
Skewness	1.18059063	Kurtosis	1.78164654
Uncorrected SS	9033719	Corrected SS	193665.647
Coeff Variation	14.8173104	Std Error Mean	0.95357498

Basic Statistical Measures			
Location		Variability	
Mean	138.3268	Std Deviation	20.49632
Median	134.0000	Variance	420.09902
Mode	134.0000	Range	117.00000
		Interquartile Range	24.00000

Note: The mode displayed is the smallest of 2 modes with a count of 29.

Some basic descriptive statistics for '**sbp**'(systolic blood pressure), Mean blood pressure of the patients is 138.32 and Median is 134. '**sbp**' has large range i.e.; difference between minimum and maximum for '**sbp**' is 117, also the standard deviation is 20.49 so '**sbp**' varies a lot in the sample of patients.

The UNIVARIATE Procedure
Variable: tobacco

Moments			
N	462	Sum Weights	462
Mean	3.63564935	Sum Observations	1679.67
Std Deviation	4.59302408	Variance	21.0958702
Skewness	2.07920967	Kurtosis	5.96810787
Uncorrected SS	15831.8873	Corrected SS	9725.19616
Coeff Variation	126.332978	Std Error Mean	0.21368682

Basic Statistical Measures			
Location		Variability	
Mean	3.635649	Std Deviation	4.59302
Median	2.000000	Variance	21.09587
Mode	0.000000	Range	31.20000
		Interquartile Range	5.45000

For '**tobacco**' (cumulative tobacco in kg), Mean tobacco consumption of the patients is 3.635 kg and Median is 2 kg. '**tobacco**' has large range 31.2 kg, also the standard deviation is 4.59 kg, so this variable varies a lot from the mean value in the sample of patients.

The UNIVARIATE Procedure
Variable: ldl

Moments			
N	462	Sum Weights	462
Mean	4.74032468	Sum Observations	2190.03
Std Deviation	2.07090916	Variance	4.28866475
Skewness	1.31310398	Kurtosis	2.87655294
Uncorrected SS	12358.5277	Corrected SS	1977.07445
Coeff Variation	43.6870743	Std Error Mean	0.09634741

Basic Statistical Measures			
Location		Variability	
Mean	4.740325	Std Deviation	2.07091
Median	4.340000	Variance	4.28866
Mode	3.570000	Range	14.35000
		Interquartile Range	2.52000

Note: The mode displayed is the smallest of 3 modes with a count of 5.

For '**ldl**' (low density lipoprotein cholesterol), Mean for the sample patients is 4.74 and Median is 4.34. '**ldl**' has range 14.35, also the standard deviation is 2.07.

The UNIVARIATE Procedure
Variable: adiposity

Moments			
N	462	Sum Weights	462
Mean	25.4067316	Sum Observations	11737.91
Std Deviation	7.7806986	Variance	60.5392706
Skewness	-0.2146459	Kurtosis	-0.6984386
Uncorrected SS	326130.533	Corrected SS	27908.6038
Coeff Variation	30.6245554	Std Error Mean	0.36199086

Basic Statistical Measures			
Location		Variability	
Mean	25.40673	Std Deviation	7.78070
Median	26.11500	Variance	60.53927
Mode	21.10000	Range	35.75000
		Interquartile Range	11.54000

Note: The mode displayed is the smallest of 4 modes with a count of 3.

For '**adiposity**' (fat levels), Mean for the sample patients is 25.40 and Median is 26.11. '**ldl**' has range of 35.75, also the standard deviation is 7.78

The UNIVARIATE Procedure
Variable: typea

Moments			
N	462	Sum Weights	462
Mean	53.1038961	Sum Observations	24534
Std Deviation	9.81753412	Variance	96.3839761
Skewness	-0.3464378	Kurtosis	0.47040234
Uncorrected SS	1347284	Corrected SS	44433.013
Coeff Variation	18.4874083	Std Error Mean	0.45675302

Basic Statistical Measures			
Location		Variability	
Mean	53.10390	Std Deviation	9.81753
Median	53.00000	Variance	96.38398
Mode	52.00000	Range	65.00000
		Interquartile Range	13.00000

For '**typea**'(Personality trait that is characterized by excessive competitiveness and aggression) which is a score, Mean score for the sample patients is 53.10 and Median is 53. '**typea**' has range of 65, also the standard deviation is 9.81

The UNIVARIATE Procedure
Variable: obesity

Moments			
N	462	Sum Weights	462
Mean	26.0441126	Sum Observations	12032.38
Std Deviation	4.21368023	Variance	17.7551011
Skewness	0.9052194	Kurtosis	2.25597162
Uncorrected SS	321557.761	Corrected SS	8185.10159
Coeff Variation	16.1790125	Std Error Mean	0.19603815

Basic Statistical Measures			
Location		Variability	
Mean	26.04411	Std Deviation	4.21368
Median	25.80500	Variance	17.75510
Mode	24.86000	Range	31.88000
		Interquartile Range	5.55000

Note: The mode displayed is the smallest of 2 modes with a count of 4.

For '**obesity**' variable, Mean value for the sample patients is 26.04 and Median is 25.80. '**obesity**' has range of 31.88, also the standard deviation is 4.21

The UNIVARIATE Procedure
Variable: alcohol

Moments			
N	462	Sum Weights	462
Mean	17.0443939	Sum Observations	7874.51
Std Deviation	24.4810587	Variance	599.322235
Skewness	2.31269894	Kurtosis	6.42110997
Uncorrected SS	410503.801	Corrected SS	276287.55
Coeff Variation	143.631148	Std Error Mean	1.13896193

Basic Statistical Measures			
Location		Variability	
Mean	17.04439	Std Deviation	24.48106
Median	7.51000	Variance	599.32223
Mode	0.00000	Range	147.19000
		Interquartile Range	23.46000

For '**alcohol**' variable, Mean value for the sample patients is 17.04 and Median is 7.51. '**alcohol**' has range of 147.19, also the standard deviation is 24.48 so the consumption of alcohol varies a lot in the sample.

The UNIVARIATE Procedure
Variable: age

Moments			
N	462	Sum Weights	462
Mean	42.8160173	Sum Observations	19781
Std Deviation	14.6089564	Variance	213.421608
Skewness	-0.3817343	Kurtosis	-1.016229
Uncorrected SS	945331	Corrected SS	98387.3615
Coeff Variation	34.1203067	Std Error Mean	0.67967017

Basic Statistical Measures			
Location		Variability	
Mean	42.81602	Std Deviation	14.60896
Median	45.00000	Variance	213.42161
Mode	16.00000	Range	49.00000
		Interquartile Range	24.00000

For '**age**' variable, Mean age for the sample patients is 42.81 years and Median is 45 years. '**age**' has range of 49 years, also the standard deviation is 14.60 years.

The FREQ Procedure

chd	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	302	65.37	302	65.37
1	160	34.63	462	100.00

famhist	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Absent	270	58.44	270	58.44
Present	192	41.56	462	100.00

In the data, we have two categorical variables **'chd', 'famhist'**.

From the frequency tables, 302 observations have chd=0 which represents the patients not having heart disease and 160 patients with chd=1 have the heart disease.

'famhist' variable represents if any family member of the patient has the heart disease. And we have 270 patients with their family member not having heart disease (famhist =Absent), 192 patients are with family member having heart disease (famhist=present).

4) Question to answer:

In the previous step, we did exploration all variables in the dataset. And there are two important categorical variables in the dataset **'chd' (coronary heart disease)** and **'famhist' (family history).** Respectively, these two variables have information whether the patient has heart disease or not, as well as any family member of the patient having the heart disease.

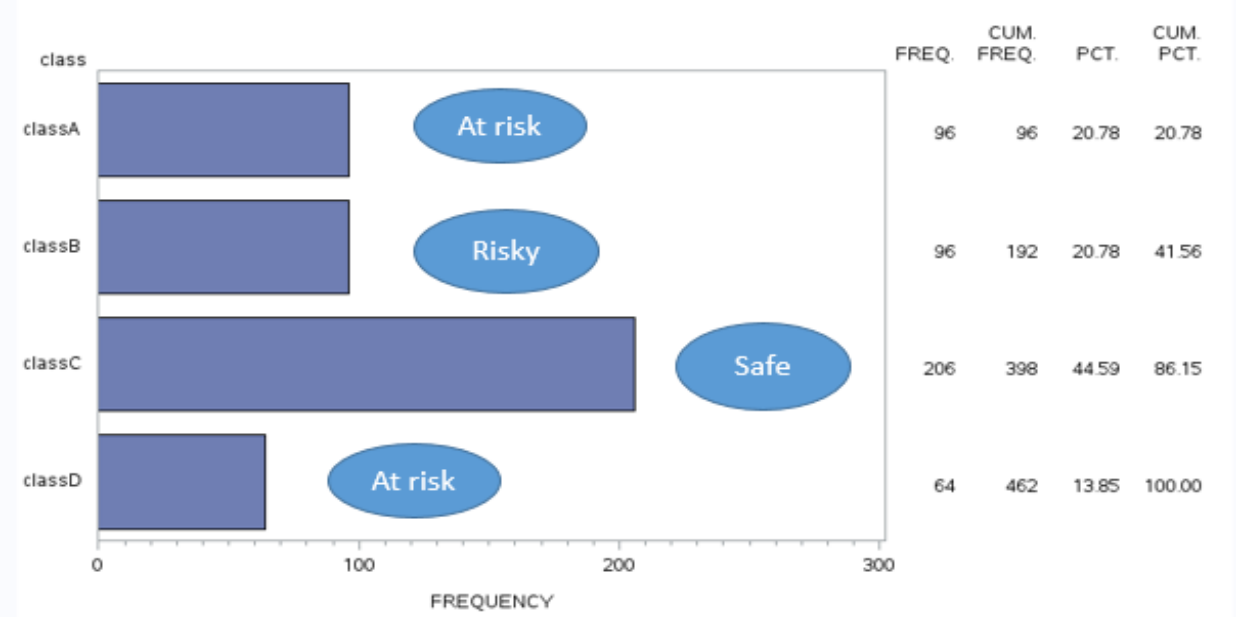
As part of my analysis I will try to answer the question, **what are the significant variables that influence the combination of patient and his family member getting the heart disease. If we can find answer to this question we can interpret the hereditary factor in heart disease.** So, for the required analysis we need to create new response variable based on the combination of **'chd'** and **'famhist'** as below.

Chd(0/1)	famhist(Present/Absent)	Class (New response)
0	Present	ClassA
1	Present	ClassB
0	Absent	ClassC
1	Absent	ClassD

Now we have the new response variable **'class'** with four levels, and we are dealing with multi-class classification problem. So, I used discriminant analysis model (LDA/QDA) for the analysis and to identify significant variables.

And if we can come up with a useful model than can explain and interpret several factors/predictors as mentioned above, we could apply this predictive model to future patients. We would be able to classify new patients in combination with their family history into several groups, which would be of help in treatment process.

Before moving to the discriminant analysis, we try to interpret the **'class'** variable for the analysis purpose. In **'class'** variable 'classC' is considered as "benign/safe", because neither the patient nor his family member has heart disease. 'classB' represents the patients who are "risky" because both patient and his family member has heart disease. And 'classA', 'classD' are "at risk" because either the patient or the family member has disease. Per this formulation of **'class'** variable our discriminant model should be able to classify different sections of patients correctly and QDA results will be discussed in the coming sections. The picture below shows the frequency of each class of patients with the total of 462 patients. We have majority of the patients (206 out of 462) are in 'classC' or "safe".



class	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
classA	96	96	20.78	20.78
classB	96	192	20.78	41.56
classC	206	398	44.59	86.15
classD	64	462	13.85	100.00

5) Discriminant Analysis and Results:

By using discriminant analysis, we are going to analyze how several predictor variables such as **'sbp', 'typea', 'tobacco', 'alcohol', 'ldl', 'adiposity', 'obesity', 'age'** will discriminate the response variable **'Class'**

STEP-1 (variable selection): In the analysis at first, we will perform stepwise variable selection and variables are selected based on the 0.05 significance level. Below is the final selection summary for the variables.

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	age		0.1607	29.23	<.0001	0.83929763	<.0001	0.05356746	<.0001
2	2	ldl		0.0431	6.86	0.0002	0.80311918	<.0001	0.06643638	<.0001
3	3	tobacco		0.0304	4.77	0.0028	0.77870328	<.0001	0.07563181	<.0001
4	4	typea		0.0216	3.35	0.0191	0.76189055	<.0001	0.08144358	<.0001

From the selection summary table, variables **'age', 'ldl', 'tobacco', 'typea'** have p-values less than the significance level (0.05), so we will retain these four significant variables in the final model.

STEP-2 (Homogeneity of Co-variance test): Now we will try to identify, either LDA or QDA will be ideal for the discriminant analysis and this can be done by using the homogeneity of variance test, below are results from the test.

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
130.913355	30	<.0001

Based on the results from Homogeneity within covariance matrices test, we have p-value (<0.0001) less than 0.05 significance level so we have evidence in support of alternate hypothesis (Ha: Homogeneity of covariance is not valid) over the null hypothesis (H0: Homogeneity of covariance is valid). Thus, we can say variance in the covariance matrices is not equal so we will opt for Quadratic discriminant analysis over Linear discriminant analysis.

STEP-3 (MANOVA test): As the next step, we can test whether the significant variables chosen in the previous step are good enough for discriminate analysis of the response variable**('Class').** MANOVA test can be used to check the possibility of discrimination using the significant variables and test results are as below.

The DISCRIM Procedure

Multivariate Statistics and F Approximations					
S=3 M=0 N=226.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.76189055	10.86	12	1204.1	<.0001
Pillai's Trace	0.24433075	10.13	12	1371	<.0001
Hotelling-Lawley Trace	0.30440681	11.52	12	791.95	<.0001
Roy's Greatest Root	0.27577493	31.51	4	457	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

The table above shows that, all the tests that are listed have p-values less than 0.05 significance level. Hence based on MANOVA results we have evidence in support of alternate hypothesis, which conveys the possibility of discrimination.

STEP-4 (Performing QDA): Till now, we performed several tests and obtained the set of significant variables. We continue, the Quadratic discriminant analysis with best set of predictors, proportional priors and on the whole dataset with 462 observations. QDA model performance will be evaluated based on cross-validation technique.

Below we have cross-validate summary for the QDA analysis and we also have contingency table/confusion matrix to evaluate model performance for each class in the response variable.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.HEART_DIS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into class					
From class	classA	classB	classC	classD	Total
classA	2 2.08	18 18.75	74 77.08	2 2.08	96 100.00
classB	3 3.13	41 42.71	45 46.88	7 7.29	96 100.00
classC	7 3.40	30 14.56	159 77.18	10 4.85	206 100.00
classD	2 3.13	22 34.38	33 51.56	7 10.94	64 100.00
Total	14 3.03	111 24.03	311 67.32	26 5.63	462 100.00
Priors	0.20779	0.20779	0.44589	0.13853	

Error Count Estimates for class					
	classA	classB	classC	classD	Total
Rate	0.9792	0.5729	0.2282	0.8906	0.5476
Priors	0.2078	0.2078	0.4459	0.1385	

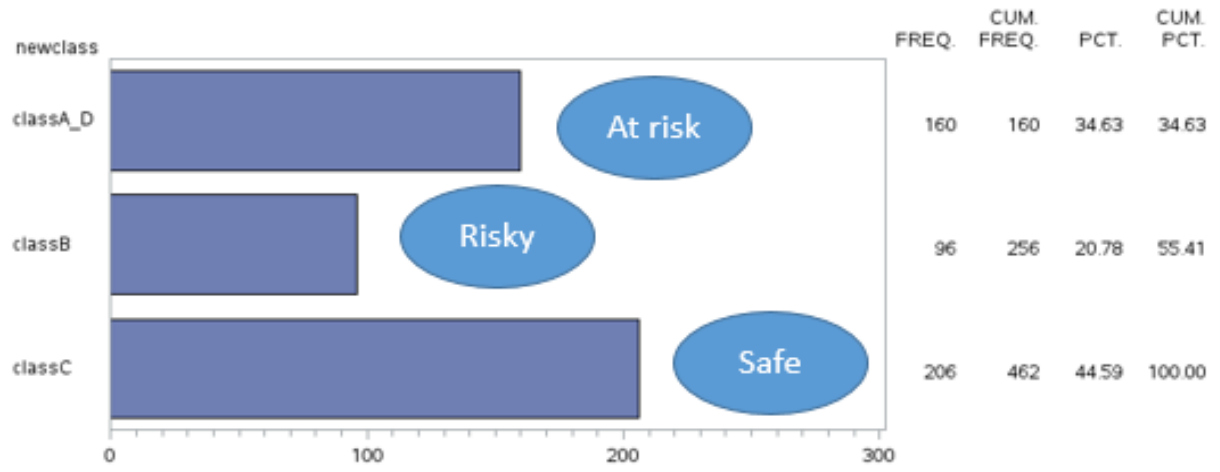
From the Error Count Estimates table, over-all error rate for the model is 54.76 percent. But for the 'classC' error rate is 22.82 percent which means the model is performing well at classifying the patients who are categorized as 'benign/safe'. For 'classB' error rate is 57.29 percent, the model is okay for this class and it is doing decent job in classifying patients those considered to be 'risky'. For 'classA' and 'classD' the model performance is bad with large error rates. Classes of patients who are 'At risk' needs attention and our model should predict these classes accurately.

Also, if we look at the classification summary table we can find that 'classA', 'classD' are often confused with 'classC' and it means to say that our QDA model is classifying the patients who are 'At risk' as 'safe'. So, the QDA model is being more optimistic and misclassified important 'At risk' patients with large error rates, and the analysis with four levels in 'class' variables is little useful hence we need to formulate the response variable again. Response variable **'class'** is converted into **'newclass'** variable by merging the 'classA', 'classD' where-as 'classB', 'classC' remains the same.

Merging different classes is subjective, and for the analysis I intended to do it is important to classify the patients who belongs to 'classA', 'classD' because these levels are regarded as

patients who are 'At risk'. And if we can correctly predict this class of patients using the model, we can make use of this model for predicting heart disease in future patients combined with their family history. As well as this model, would be of some help for the experts in medical domain to understand several factors that are relevant for making association between patient and family member, we can warn the patients who are 'At risk' so that some precautions or preventive measures can be taken to avoid heart disease.

Now onwards, **'newclass'** is the response variable and the discriminant analysis performed earlier will be repeated using the all continuous variables in the data. Below we have frequency plot for the response variable **("newclass")** and we can observe class A, D being merged. Still majority of observations are 'classC' or 'safe'.



<u>Repeating Discriminant analysis with 'newclass' as response:</u>

At first, we will do variable selection using **'newclass'** as response and **'sbp', 'typea', 'tobacco', 'alcohol', 'ldl', 'adiposity', 'obesity', 'age'** as predictor variables**.**

STEP-1 (variable selection): Like before Step-wise variable selection is done with 0.05 significance level, selection summary results are below.

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	age		0.1465	39.39	<.0001	0.85349495	<.0001	0.07325252	<.0001
2	2	ldl		0.0411	9.82	<.0001	0.81840706	<.0001	0.09189649	<.0001
3	3	typea		0.0177	4.12	0.0168	0.80390259	<.0001	0.09931652	<.0001

In this case variables **'age', 'ldl', 'typea'** have p-values (<0.001) less than 0.05 significance level, so these three variables will be used in the final model.

STEP-2 (Homogeneity of Co-variance test): Now we will test for the homogeneity of covariance matrices and based on the results we will decide, whether to use Quadratic discriminant analysis or Linear discriminant analysis. And from the test results below, we have p-value (<0.0001) less than 0.05 so we have evidence in support of alternate hypothesis. Based on alternate hypothesis we say that variance of the covariance matrices for the predictors are not equal, so we choose QDA for the analysis.

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
43.593065	12	<.0001

STEP-3 (MANOVA test): we use MANOVA test to check the possibility of discriminating the response variable **'newclass'** based on significant variables (**'age', 'ldl', 'typea')** selected from previous step, test results are shown below. And from the table we have p-values for all tests are less than 0.05 significance level, MANOVA results convey the possibility of discrimination.

The DISCRIM Procedure

Multivariate Statistics and F Approximations					
S=2 M=0 N=227.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.80390259	17.57	6	914	<.0001
Pillai's Trace	0.19863305	16.83	6	916	<.0001
Hotelling-Lawley Trace	0.24077765	18.32	6	607.56	<.0001
Roy's Greatest Root	0.22687502	34.64	3	458	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

STEP-4 (Performing QDA): By now, we have done stepwise variable selection for the discriminant analysis, tested homogeneity of variance and decided to use QDA. Also, MANOVA results suggest possibility of discrimination. So, we move forward and model Quadratic discriminant analysis with **'newclass'** as response and best set of predictors (**'age', 'ldl', 'typea')** from the stepwise selection, proportional priors. QDA results were validated based on cross-validation methodology and the results are below.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.HEART_DIS_3_CLS
Cross-validation Summary using Quadratic Discriminant Function

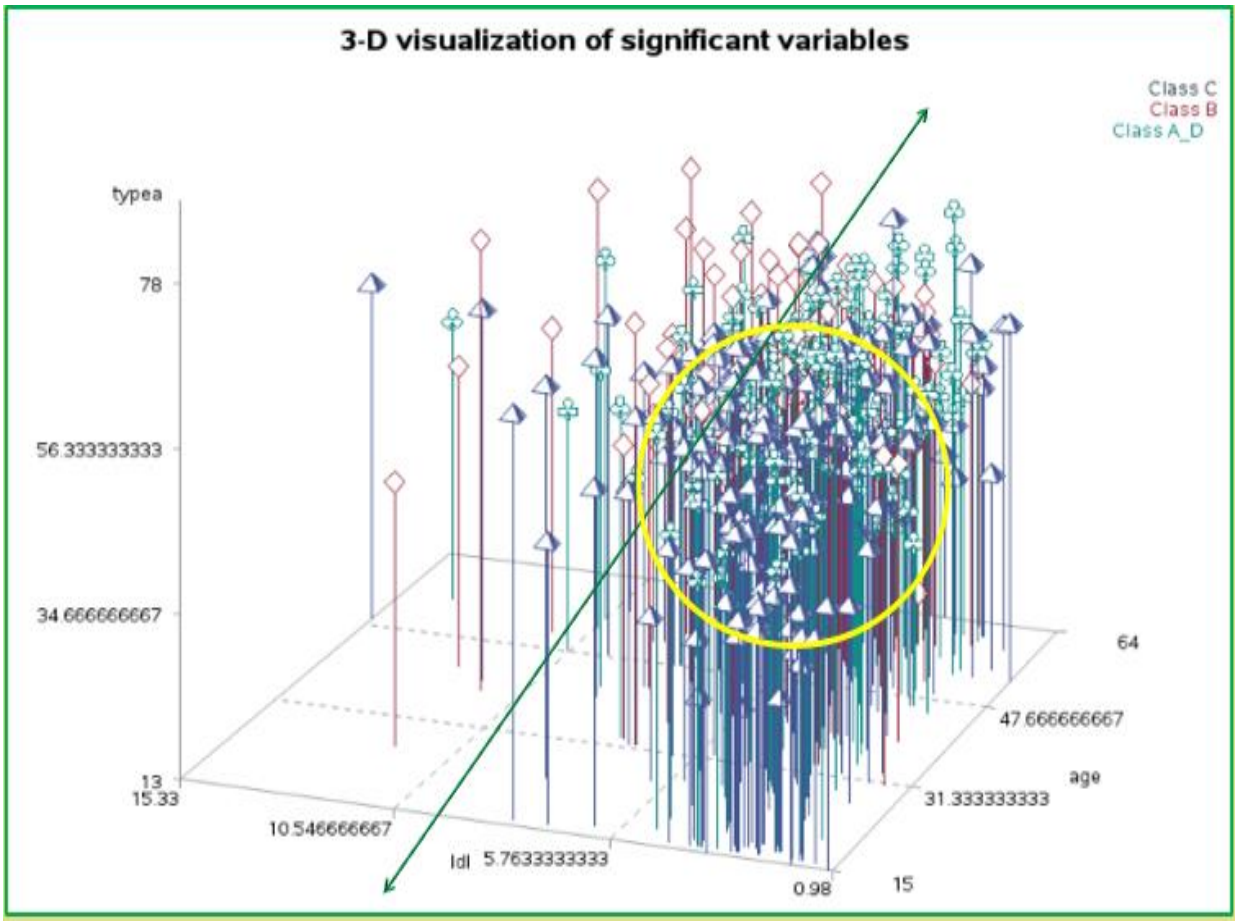
Number of Observations and Percent Classified into newclass				
From newclass	classA_D	classB	classC	Total
classA_D	71 44.38	20 12.50	69 43.13	160 100.00
classB	47 48.96	32 33.33	17 17.71	96 100.00
classC	57 27.67	26 12.62	123 59.71	206 100.00
Total	175 37.88	78 16.88	209 45.24	462 100.00
Priors	0.34632	0.20779	0.44589	

Error Count Estimates for newclass				
	classA_D	classB	classC	Total
Rate	0.5563	0.6667	0.4029	0.5108
Priors	0.3463	0.2078	0.4459	

From the cross- validation results above, the over-all error rate is 51.08 percent and the individual class error rates are 'classA_D' has 55.63 percent error rate, 'classB' has 66.67 percent error rate, 'classC' has 40.29 percent error rate. If we look at the 'Classification summary' table, 'classA_D' is often confused with 'classC' based on the number of observations that are misclassified. Even though the misclassification error rate for 'classA_D' is little over 55 percent, the QDA model is useful because with this model nearly half of the times we correctly classify the patients who are 'At risk'. So, merging the 'classA', 'classD' has produced meaningful results but still the model is misclassifying a lot between 'classA_D' and 'classC'. Hence, answering the combination of patient and his family member having the heart disease is a complex thing, we may need some more predictors related to the genetic information so that we can achieve better results and explain the hereditary factor in the heart disease.

6) Conclusions:

From the Quadratic discriminant analysis, based on the step wise selection we identified three variables are (**'age', 'ldl', 'typea')** as significant. Now we will try to infer how variables are discriminating three classes ('classA_D', 'classB', 'classC') in the response variable **'newclass'** and make some conclusions**.**



Above is the 3-D visualization for the three significant variables with all 462 patients/ observations are visualized in 3-D space. We can see that three variables **'age', 'ldl', 'typea'** represent three different axes. And we also have three different symbols, colors for the classes such that 'classC' is represented by 'pyramid' shape and 'blue' color, 'classB' is represented by 'diamond' shape and 'red' color, 'classA_D' is represented by 'leaf' shape and 'light green' color.

In the visualization, we have a 'dark green' line which can cut the 3-D plot along **'ldl'** axis. And towards the left of this line we have many observations with 'diamond' shape corresponding to 'classB'. From this we can derive

Conclusion 1: Patients with higher **'ldl'** (lipoprotein density) levels are the 'Risky' class (classB) of patients.

Considering the 'dark green' line again, towards the right side of this line we have majority of patients from 'classC' (pyramid shape) and 'classA_D' (leaf shape) with no exact separation between the two classes. So, we can infer that

Conclusion 2: Majority of Patients with lower **'ldl'** (lipoprotein density) levels are the 'At risk' or 'Safe' classes (classA_D, classC) of patients. But the other two axes **'age', 'typea'** could not separate the classA_D and classC, it can be inferred that these two classes are often confused.

Back to the visualization, we have a yellow circle and this can be treated as a projection of the plane representing the **'ldl'** and **'age'**. And this circle is aligned towards the lower right corner of the plane **'ldl'** and **'age'** and it means to say this circle represents the lower **'ldl'** levels and smaller **'age'.** We can observe that yellow circle has majority of the observations with 'pyramid' shape(classC). So, we can interpret

Conclusion 3: Patients with Smaller **'age'** and lower **'ldl'** levels are of the 'Safe' class even though they have varying **'typea'** scores.

7) Issues and remedies:

Major issue with the analysis is that combining **chd** (coronary heart disease) and **famhist** (family history) is a complex thing i.e.; hereditary factor in heart disease is difficult to infer. From the QDA results we observed that we are often confused between the classes 'Safe', 'At risk' and misclassifying them. So, one of the possible remedy to handle this situation would be considering/adding variables related to genetic information, which can associate the patient and his family. Also, we have smaller sample of patients for the analysis, if we could use much bigger sample for the analysis that may solve the current issue and there is also a possibility of making more robust conclusions about coronary heart disease.

8) Appendices:

This section includes additional details on the analysis such as details on QDA, SAS procedures.

Mathematical details on Discriminant Analysis:

Discriminant analysis is a classification algorithm in which, each of the class is separately modelled to a distribution. And posterior probabilities for the distribution are obtained from the prior based on the Bayes theorem. We assume the distribution to be normal so it has probability density function of multivariate Gaussian distribution with K classes and it is as follows

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma^{-1}(x-\boldsymbol{\mu}_k)\right] \qquad \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

And there are two variants of discriminant analysis LDA (linear discriminant), QDA (quadratic discriminant), the difference is LDA assume constant variance across classes where-as QDA

doesn't make this assumption. We can derive the decision boundary that separate each of the by using the log likelihood function of condition distribution $\mathcal{N}(\mu_k, \Sigma_k)$.

For QDA we don't assume constant Σ (co-variance) across class so the discriminant function becomes and it is called quadratic discriminant because of the quadratic term in the function.

$$\arg\max_k \; -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^{\mathsf{T}}\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)$$

$$x^{\mathsf{T}}\mathbf{W}_k x + \mathbf{w}_k^{\mathsf{T}} x + b_k$$

And the decision boundary between two classes k, l becomes

$$x^{\mathsf{T}}(\mathbf{W}_k - \mathbf{W}_l)x + (\mathbf{w}_k^{\mathsf{T}} - \mathbf{w}_l^{\mathsf{T}})^{\mathsf{T}}x + (b_k - b_l) = 0$$

Similarly, we can find the decision boundary between several classes in case multi-classification problem. And with these decision boundaries we can discriminate between several classes.

Details on SAS procedures:

SAS programming was used to perform the intended analysis, and several SAS procedures were used for different purposes as below:

PROC IMPORT: For importing the dataset into SAS.

PROC UNIVARAITE: For generating descriptive statistics for continuous variables.

PROC FREQ: For generating descriptive statistics for categorical variables.

DATA STEP: To generate several intermediate SAS datasets for the analysis.

PROC GCHART: For generating bar graphs of the response variable.

PROC STEPDISC: To perform step-wise variable selection.

PROC DISCRIM: To perform Quadratic Discriminant Analysis.

PROC G3D: For plotting 3D visualization of three significant variables in the analysis.

Additional details about variables: Below table has units of measurement for several variables in the dataset.

Variable	Age	Sbp	Tobacco	ldl	Alcohol	Type-a	Adiposity	Obesity
Units	Years	mmHg	kg	mmol/l	Uints/week	points	Percent of fat	Kg/m2

*** THE END ***