**Individual Project report (Discriminant Analysis)**

***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 |

**Sections in the report:**

1. **Introduction.**
2. **Data Description.**
3. **Descriptive Statistics.**
4. **Question to Answer.**
5. **Discriminant Analysis and Results.**
6. **Conclusions.**
7. **Issues and remedies.**
8. **Introduction:**

For the final project, our group choose to work on dataset related to heart disease (health sciences). We know 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.

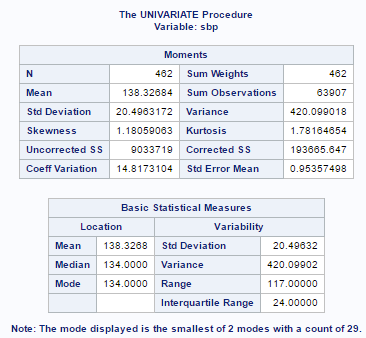
[URL: <http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data> ]

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 the high level these medical records have patient information related to Age, Drinking habits, Smoking habits, Eating habits, Family history and Personality traits etc.

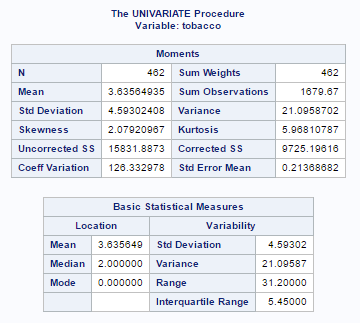
1. **Data description:**

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.

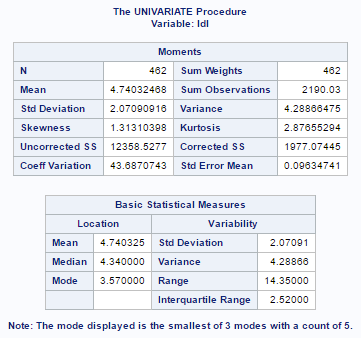
1. **Descriptive statistics for variables using PROC FREQ and UNIVARIATE:**



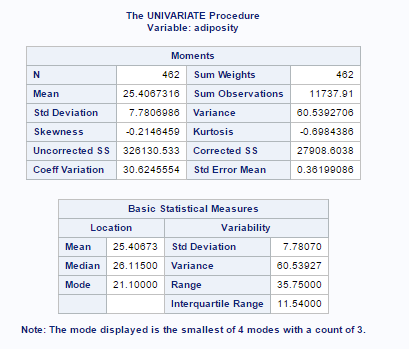
Some basic descriptive statistics for ‘**sbp**’(systolic blood pressure), Mean blood pressure of the patients is 138.32 and Median is 134. ‘**sbp**’ has large range i.e.; difference between minimum and maximum for ‘**sbp**’ is 117, also the standard deviation is 20.49 so ‘**sbp**’ varies a lot in the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



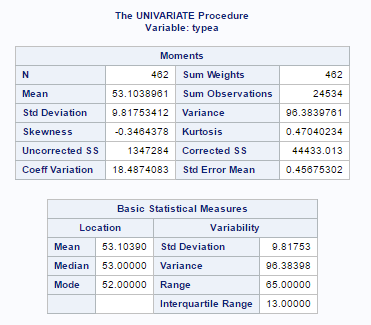
For ‘**tobacco**’ (cumulative tobacco in kg), Mean tobacco consumption of the patients is 3.635 kg and Median is 2 kg. ‘**tobacco**’ has large range 31.2 kg, also the standard deviation is 4.59 kg, so this variable varies a lot from the mean value in the sample of patients.



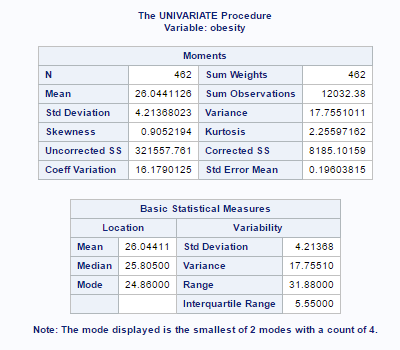
For ‘**ldl**’ (low density lipoprotein cholesterol), Mean for the sample patients is 4.74 and Median is 4.34. ‘**ldl**’ has range 14.35, also the standard deviation is 2.07.



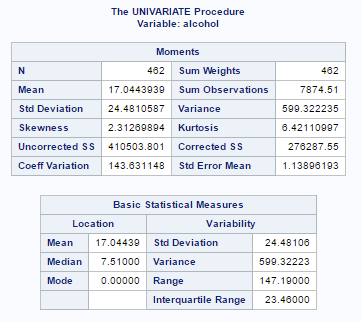
For ‘**adiposity**’ (fat levels), Mean for the sample patients is 25.40 and Median is 26.11. ‘**ldl**’ has range of 35.75, also the standard deviation is 7.78



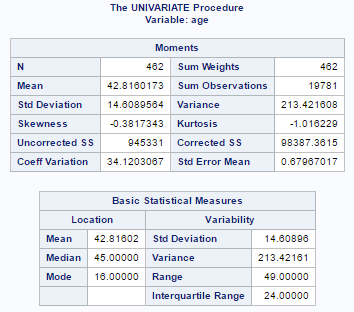
For ‘**typea**’(Personality trait that is characterized by excessive competitiveness and aggression) which is a score, Mean score for the sample patients is 53.10 and Median is 53. ‘**typea**’ has range of 65, also the standard deviation is 9.81



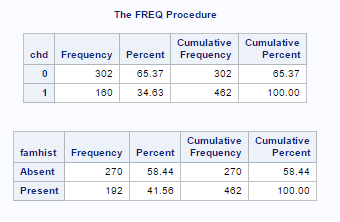
For ‘**obesity**’ variable, Mean value for the sample patients is 26.04 and Median is 25.80. ‘**obesity**’ has range of 31.88, also the standard deviation is 4.21



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.



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.



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).

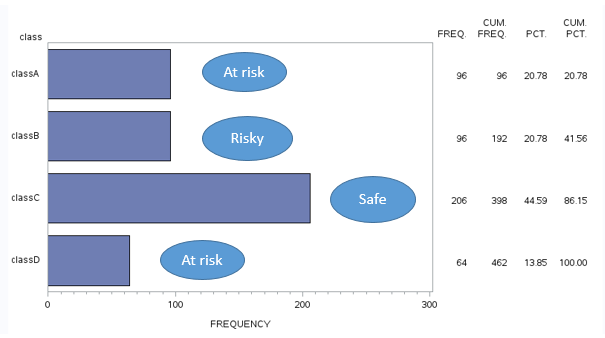
1. **Question to answer:**

Earlier we explored all variables in the dataset. We have two important categorical variables in the dataset **‘chd’** and **‘famhist’.** These respectively two variables have information whether the patient has heart disease or not, as well as any family member of the patient has heart disease or not.

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**. So, for the required analysis we need to create new response variable based on the combination of **‘chd’** and **‘famhist’** as below.

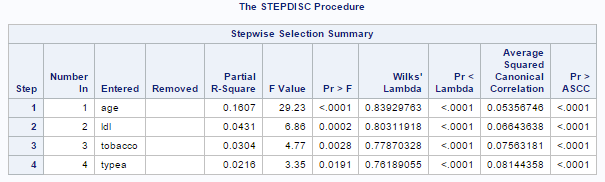
|  |  |  |
| --- | --- | --- |
| Chd(0/1) | famhist(Present/Absent) | Class (New response) |
| 0 | Present | ClassA |
| 1 | Present | ClassB |
| 0 | Absent | ClassC |
| 1 | Absent | ClassD |

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. 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. 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”.

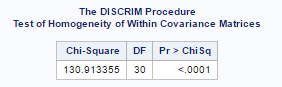


1. **Discriminant Analysis and Results:**

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’**. In the analysis at first, we will perform stepwise variable selection using the SAS procedure PROC STEPDISC and variables are selected based on the 0.05 significance level. Below is the final selection summary for the variables.

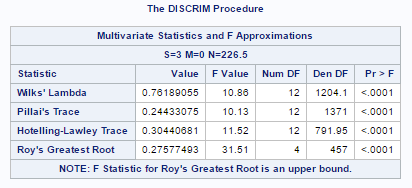


From the selection summary table, variables **‘age’, ‘ldl’, ‘tobacco’, ‘typea’** have p-values less than the significance level (0.05), so we will retain these four variables in the final model. Now we will try to identify, either LDA or QDA will be ideal for the discriminant analysis and this can be done by using the homogeneity of variance test, below are results from the test.



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.

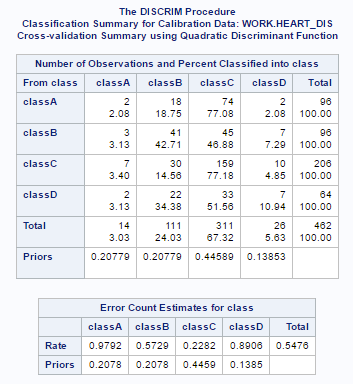
As the next step, we can test whether the significant variables chosen in the previous step are good enough for discriminate analysis of the response variable**(‘Class’).** MANOVA test can be used to check the possibility of discrimination using the significant variables and test results are as below.



The 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.

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.

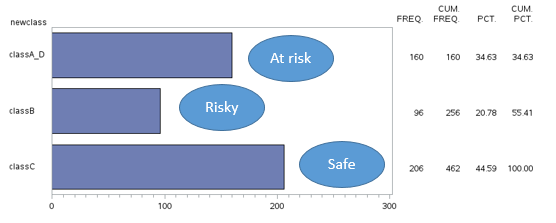


From the Error Count Estimates table, over-all error rate for the model is 54.76 percent. But for the ‘classC’ error rate is 22.82 percent which means the model is performing well at classifying the patients who are categorized as ‘benign/safe’. For ‘classB’ error rate is 57.29 percent, the model is 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. Actually , these 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.

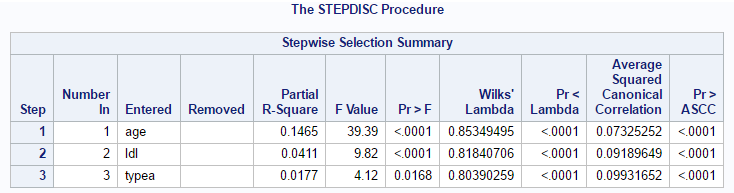
Merging different classes is subjective, and for the analysis I intended to do it is important to classify the patients who belongs to ‘classA’, ‘classD’ because these levels are regarded as patients who are ‘At risk’. And if we can correctly predict this class of patients using the model, we can make 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’.



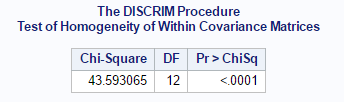
**Repeating Discriminant analysis with ‘newclass’ as response:**

At first, we will do variable selection using **‘newclass’** as response and **‘sbp’, ‘typea’, ‘tobacco’, ‘alcohol’, ‘ldl’, ‘adiposity’, ‘obesity’, ‘age’** as predictor variables**.** Like beforeStep-wise variable selection is done with 0.05 significance level, selection summary results are below.

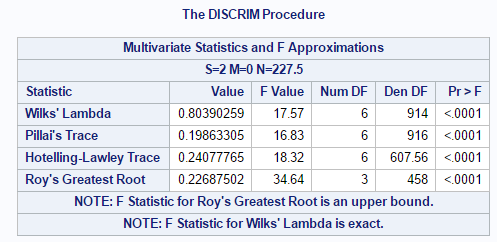


In this case variables **‘age’, ‘ldl’, ‘typea’** have p-values (<0.001) less than 0.05 significance level, so these three variables will be used in the final model.

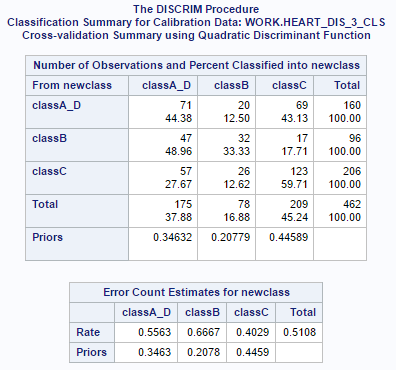
Now we will test for the homogeneity of covariance matrices and based on the results we will decide, whether to use Quadratic discriminant analysis or Linear discriminant analysis. And from the test results below, we have p-value (<0.0001) less than 0.05 so we have evidence in support of alternate hypothesis. 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.



And 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.



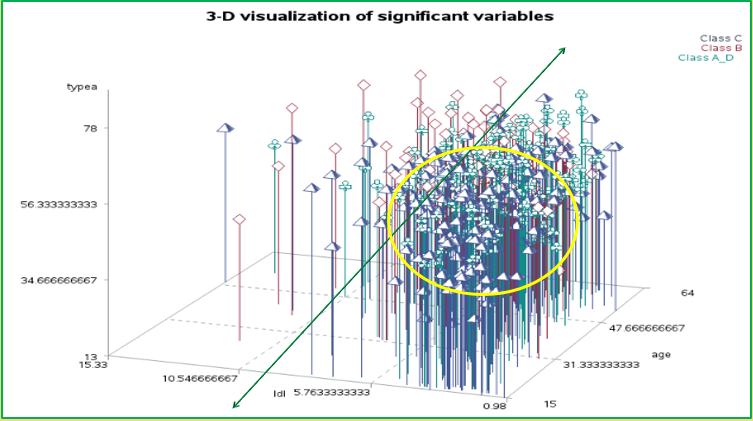
By now, we have done stepwise variable selection for the discriminant analysis, tested homogeneity of variance and decided to use QDA. Also, MANOVA results suggest possibility of discrimination. So, we move forward and model Quadratic discriminant analysis with **‘newclass’** as response and best set 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.



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 achieve better results.

1. **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 ‘darkgreen’ 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 make

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

Considering the ‘darkgreen’ 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 make

**Conclusion 2:** Majority ofPatients 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 infer that

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

1. **Issues and remedies:**

Major issue with the analysis is that combining **chd** (coronary heart disease) and **famhist** (family history) is a complex thing. 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 can use much bigger sample that may solve the current issue and there is also a possibility of making more robust conclusions about coronary heart disease.