MM924 Project 2

Question 1

The data contains 452 observations with ten variables the dimension of the data is 452 by 10. The nature of the data is data frame with four discrete variable, six continuous variable and categorical variable. The four discrete variables are unique number of individuals ("ind"), systolic blood pressure ("sbp"), type-A behaviour pattern score ("typea") and age. The six continuous variables are the yearly tobacco use ("tobacco"), low density lipoprotein cholesterol ("ldl"), percentage of body fat ("adiposity"), body mass index ("obesity") and current alcohol consumption ("alcohol"). The only categorical variable is the diagnosed coronary heart disease ("CHD"), where 1 is given that the have been diagnosed and 0 is given to those who have not.
The structure of the data is coherent to the variables, such that we have four integer types, six numerical types and one factor.

The unique number for individuals does not hold much meaning since this data number is assigned to every individual.
Systolic blood pressure has a mean of 138.4 and a median of 134.0, with a minimum of 101.0 and a maximum of 218.0.
Tobacco has a mean of 3.6 and a median of 2.0, with the minimum of 0 and maximum of 31.2.
The low-density lipoprotein has a mean of 4.8 and a median of 4.4, with a minimum of 0.98 and a maximum of 15.3.
The percentage of body fat has a mean of 25.6 and a median of 26.2, with a minimum of 6.7 and a maximum of 42.5.
Type-A behaviour pattern score has a mean of 53.1 and a median of 53.0, with a minimum of 13.0 and a maximum of 78.0.
The body mass index has a mean of 26.1 and a median of 25.8, with a minimum of 14.7 and a maximum of 46.6.
The current alcohol consumption has a mean of 16.8 and a median of 6.995, with a minimum of 0.0 and a maximum of 147.2.
Age has a mean of 42.95 and a median of 45.0, with a minimum of 15.0 and a maximum of 64.0.
The number of individuals who have been diagnosed with coronary heart disease is 155 and the number of individuals who have not been diagnosed with coronary heart disease is 297.

We visualise the variables by observing their histograms which will lead us to any transformations that is needed. Figure 1 shows that both systolic blood pressure is slightly skewed. The percentage of body fat, body mass index and type-A behaviour pattern score seem to be normally distributed. The tobacco, the lipoprotein cholesterol and current alcohol consumption have a very skewed distribution meaning these must be transformed with a log transformation with a plus one for those who have zeros as an observation. The age has a uniform distribution which shows that the similar numbers of individuals are from different ages.

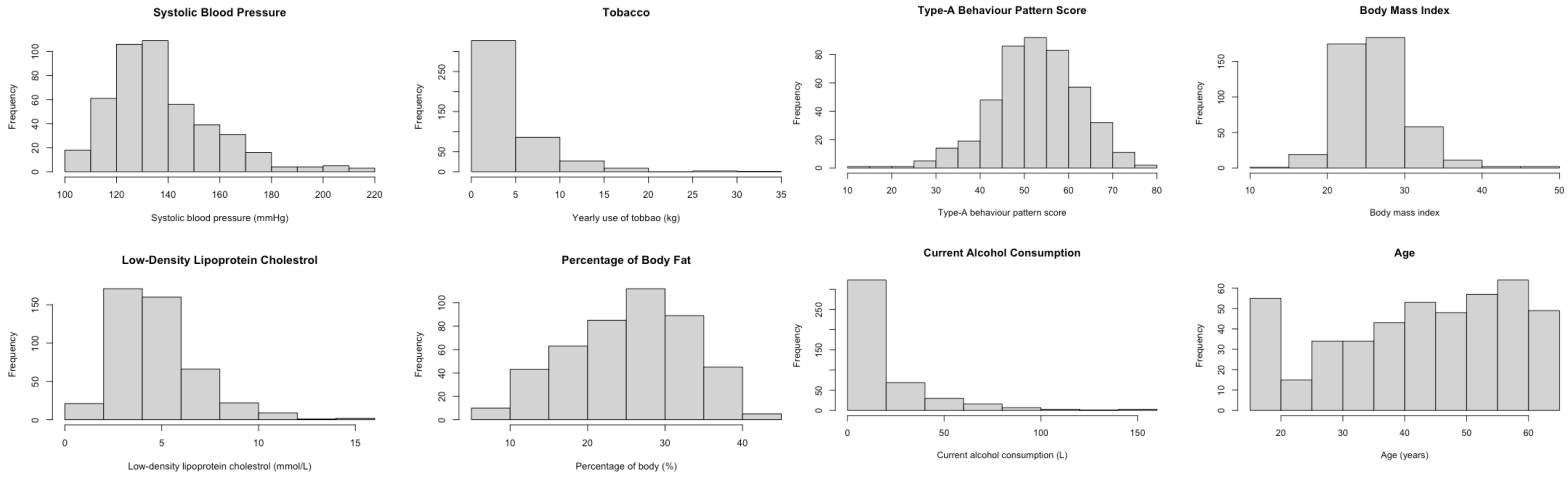

Figure 1: Eight histograms of systolic blood pressure, tobacco, low density lipoprotein cholesterol and percentage of body fat, type-A behaviour score, Body mass index, current alcohol consumption and age.

Figure 2 shows the new transformations of tobacco, lipoprotein cholesterol and alcohol consumption. The tobacco and alcohol consumption have many zeros meaning we plus one. Ignoring the large numbers of the ones, we see a more normal distribution for both variables. The lipoprotein cholesterol does not have zeros therefore we do not add one before applying the log transformation. The histogram produces a normal distribution as seen in Figure 2 graph 2. The final is the bar chart of the coronary heart disease which shows that there a more individuals who have not been diagnosed with CHD than there are.

Figure 3 shows the relationship with the eight variables with the coronary heart disease status of “0” or “1”. The group with CHD has a higher median of systolic blood pressure than without. The spread of box of group with CHD is larger than of the group without CHD meaning the variance is larger.
The group with CHD has a higher median of log tobacco plus one than the group without. The variance of the group without CHD is larger than that of group with CHD.

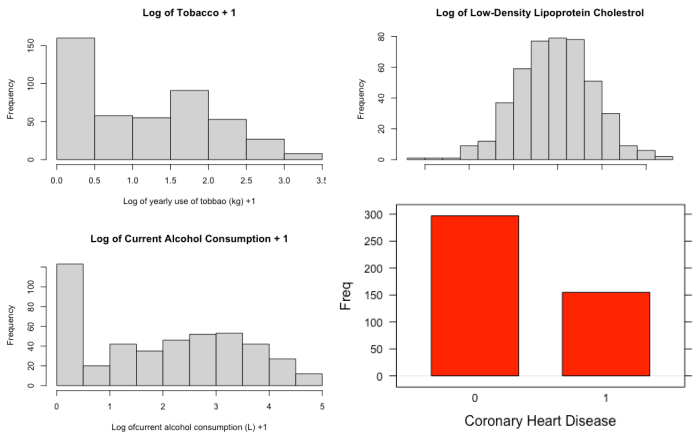

Figure 2: Three histogram of log of tobacco +1, log of low density of lipoprotein cholesterol, log of current alcohol consumption +1, and a bar chart of coronary heart disease.

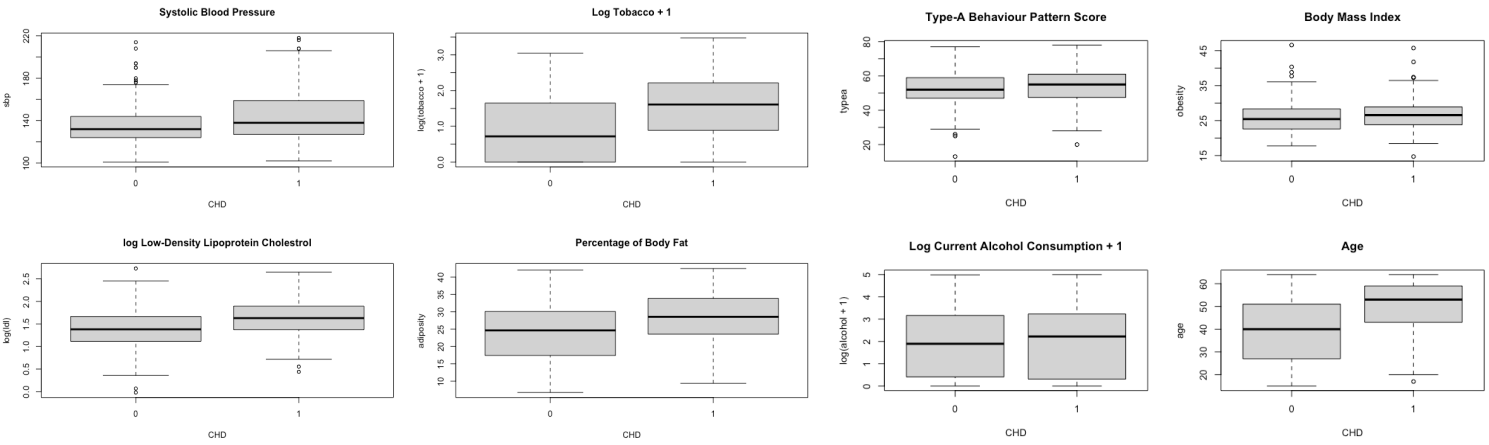

Figure 3: Eight boxplot of systolic blood pressure, tobacco, low density lipoprotein cholesterol and percentage of body fat, type-A behaviour score, body mass index, current alcohol consumption and age with the coronary heart disease.

The group with CHD has a higher median of log lipoprotein cholesterol than those without. The variance of individuals of CHD is the same as those individuals without CHD.
The group with CHD has a higher median of body fat than without CHD. The variance of individuals with CHD is smaller than those without.
The group with CHD and without CHD has similar median of type A pattern score and similar variance since the spreads looks the same.
The group with CHD and without has similar median of body mass and similar variance.
The group with CHD has a higher median of log alcohol plus one than without CHD. The spread of the box is very similar therefore the variances are as well.
The group with CHD has a higher median of age than without CHD. The spread of the group without CHD is much larger than that of the group without meaning their variance is larger.

From a simple look we see the potential risk factor that contribute to CHD are from systolic blood pressure, tobacco, lipoprotein, body fat and age. Since they had higher medians and differing sizes of variances.

Figure 4 shows the correlation between the eight variables. The largest positive correlation amongst them is between body fat and the BMI which has a correlation coefficient of 0.71. Type-A and log tobacco have no correlation since the correlation coefficient is -0.004. Many of the correlations are in between the range of 0.2 and 0.6.
The covariance shows that many differences in between two variables such as age and systolic blood pressure with a variance of 117.9. This is from the differences in measurements where age is measured in years and systolic blood pressure is measured in mmHg [1]. The variables that would be relevant for the fitted model would be all except log alcohol, log of tobacco and maybe the type A behaviour pattern score since the have low to no correlations with other variables. The covariances of these variables are also very large or very small.

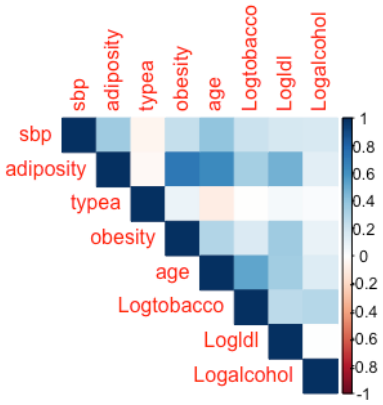

Figure 4: A correlation plot including log functions.

Question 2

The data is split into a training set and a testing set. Two thirds of the data will go to the training set and the other one third to the testing set. The training set has dimensions 294 by 9, and the testing set has dimensions 158 by 9.
The logistic regression uses the CHD status as a response variable therefore the full model is,

$$CHD = \beta_0 + \beta_1 * sbp + \beta_2 * adiposity + \beta_3 * typea + \beta_4 * obesity + \beta_5 * age + \beta_6 * Logtobacco + \beta_7 * Logldl + \beta_8 * Logalcohol + \varepsilon \quad (1)$$

A backwards selection process using a Chi-Squared Test is used, we start with our full model to a simple and more useful model.
The hypothesis test is the following:

$$H_0: \beta_i = 0$$
$$H_A: \beta_i \neq 0$$

With a significance level of 0.05 [2]. Meaning if the null hypothesis is not rejected then we remove the covariate with the least evidence against the null hypothesis [3]. In other words, the largest p-value.

First round of the backwards selection removes $Logalcohol$ since the p value is 0.9575160.

$$CHD = \beta_0 + \beta_1 * sbp + \beta_2 * adiposity + \beta_3 * typea + \beta_4 * obesity + \beta_5 * age + \beta_6 * Logtobacco + \beta_7 * Logldl + \varepsilon$$

Carrying on the selection process we remove $adiposity$ since the p-value is 0.9316559.

$$CHD = \beta_0 + \beta_1 * sbp + \beta_2 * typea + \beta_3 * obesity + \beta_4 * age + \beta_5 * Logtobacco + \beta_6 * Logldl + \varepsilon$$

Carrying on the selection process we remove $obesity$ since the p-value is 0.6964975.

$$CHD = \beta_0 + \beta_1 * sbp + \beta_2 * typea + \beta_3 * age + \beta_4 * Logtobacco + \beta_5 * Logldl + \varepsilon$$

Carrying on the selection process we remove sbp since the p-value is 0.5726638.

$$CHD = \beta_0 + \beta_1 * typea + \beta_2 * age + \beta_3 * Logtobacco + \beta_4 * Logldl + \varepsilon$$

Carrying on the selection process we remove $Logtobacco$ since the p-value is 0.1102771.

$$CHD = \beta_0 + \beta_1 * typea + \beta_2 * age + \beta_3 * Logldl + \varepsilon$$

This is the end of the selection process therefore this will be our final logistic regression for CHD status. Since no variable has a p-value greater than our significance level of 0.05 which rejects the null hypothesis in favour of the alternative, therefore we stop.

We assess the fit of the final model by comparing the deviance,
The hypothesis test is the following:

$$H_0: \ model\ 1\ fits\ the\ data$$
$$H_A: model\ 1\ is\ inadequate\ compared\ to\ model\ 2$$

With a significance level of 0.05 [4].
If the p-value is greater than 0.05 then model 1 is adequate.

The full model and the final model are compared using the hypothesis test above. The p-value is 0.6952 which is greater than the significance level therefore we don't reject the null hypothesis. This means that there is significant change compared to the full model and model 1 fits the data. Therefore, the variables removed were insignificant. The deviance of the final model is 311.19 while the full model has a deviance is 308.16 there is a deviance difference of 3.03 [5]. The Akaike's information criterion (AIC) of the final model is 319.19 and the final model is 326.16, this shows that AIC has been minimised [6].

The final model with the coefficients is,

$$CHD = -7.39580 + \ 0.04047typea + \ 0.05607age + \ 1.25030Logldl + \ \varepsilon$$

The coefficient of the intercept is the log odds when CHD equals to 0 [6]. The coefficients of the slope are change in the log odds with a unit of change in the variables [6]. Let's look at the odds and the confidence interval of each coefficient. Table 1 shows that the intercept is 0.0006138267 this is the odds ratio of not being diagnosed with CHD.
The odds ratio of type A behaviour pattern score being a potential risk factor of being diagnosed with CHD compared to not being diagnosed CHD is 1.041, The confidence interval is [1.011207, 1.072278771]. The confidence interval does not include 1. Therefore, there is a significant effect of type A behaviour pattern score on the odds of risk of being diagnosed of CHD [7]. There is a 4.1% increase of being diagnosed with CHD because of type A behaviour pattern score.
The odds ratio of age being a potential risk factor of being diagnosed with CHD compared to not being diagnosed CHD is 1.058. The confidence interval is [1.033241, 1.082676129]. The confidence interval does not include 1. Therefore, there is a significant effect of age on the odds of risk of being diagnosed of CHD [7]. There is a 5.8% increase of being diagnosed with CHD because of age.
The odds ratio of log of low-density lipoprotein cholesterol being a potential risk of CHD compared to not being a potential risk CHD is 3.491. The confidence interval is [1.714214, 7.111065574]. The confidence interval does not include 1 [7]. Therefore, there is a significant effect of log of low-density lipoprotein cholesterol on the odds of risk of being diagnosed of CHD. There is a 249% increase of increase of being diagnosed with CHD because of log of low-density lipoprotein cholesterol.
The odds of the log of low-density lipoprotein cholesterol is the largest risk factor. The probability of this event is about 0.778. For the age and type-A behaviour pattern score are 0.514 and 0.510 respectively.

Table 1: Odds and confidence interval of the coefficients

	Odds ratio	2.5%	97.5%
Intercept	0.0006138267	0.00005680572	0.006632839
Type-A	1.0412951628	1.011207	1.072278771
Age	1.0576699998	1.033241	1.082676129
logLDL	3.4914020429	1.714214	7.111065574

Table 2: Sensitivity, specificity, and classification rate on different classification techniques training sets

Classification techniques	Sensitivity	Specificity	Correct Classification Rate
Binary (cut-off = - 0.7)	69.1%	67.5%	68.0%
ROC (cut-off = -0.967)	85.1%	58.5%	66.7%
LDA	35.1%	88.5%	71.4%
QDA	41.1%	84.0%	70.4%
LDA CV	35.1%	88.0%	71.0%
QDA CV	41.4%	83.0%	69.0%
KNN K= 11	39.7%	82.0%	66.5%
KNN CV K = 11	43.2%	82.7%	69.7%

To find the optimal balance point for the sensitivity and specificity of the model of the fitted data we use different classification approaches such as binary classification, ROC, linear (LDA) and quadratic discriminant analysis (QDA), and the K-nearest neighbour classifications (KNN). We will also use cross validation (CV) of the LDA, QDA and KNN. The sensitivity is the proportion of cases that have been correctly identified [8]. The specificity is the proportion of non-cases that have been correctly identified [8].

The binary classification uses a linear predictor to find a cut-off point which is -0.7 [9], this is shown in Figure 5. The boxplot shows that the cut-off point is somewhere between -0.5 and -1. However, the density plot shows that there is an overlap between the two different statuses. This means that wherever we cut off there will be a large amount of the other status taking out as well [9]. The sensitivity, specificity and correct classification rate are shown in Table 2.

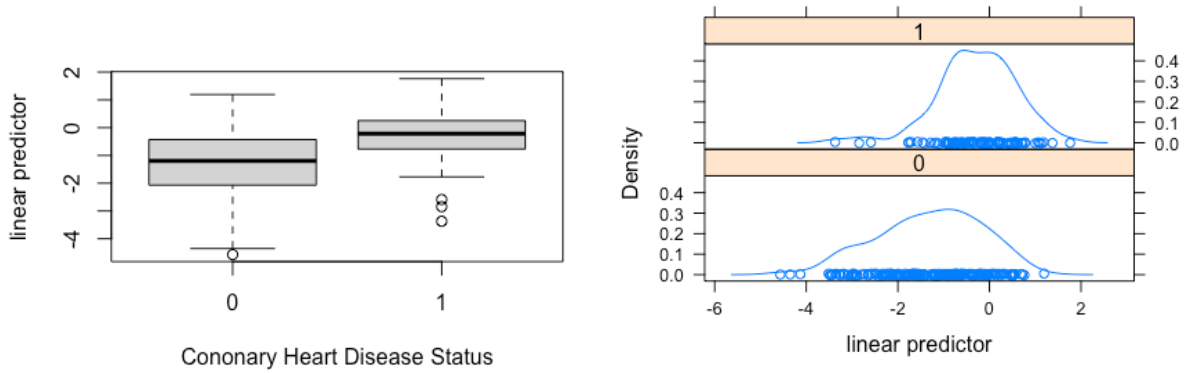

Figure 5: Boxplot and a density plot of the CHD status with a linear predictor

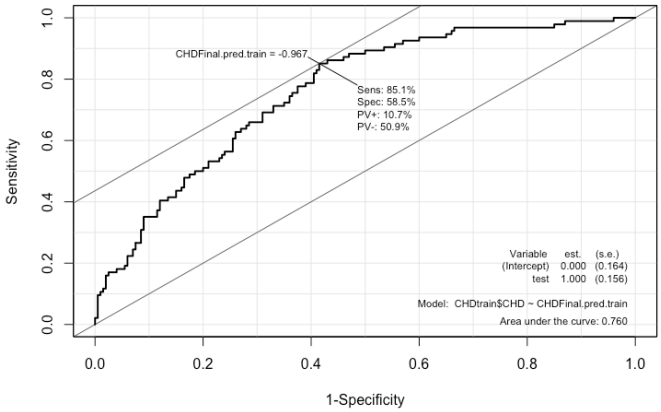

Figure 6: ROC plot for the training set.

Figure 6 shows the ROC plot where we find the best cut off point which is -0.967 for the training set this provides us with the sensitivity, and the specificity as recorded in Table 2 [8].

Table 2 also has other techniques such as LDA, LDA CV, QDA and QDA CV the sensitivity in these is very small and the specificity very large. They also have higher correct classification rate especially as shown by the LDA [10].
The KNN and the KNN cross validation with an appropriate k equalling 11 (an elbow point), also has a similar very small sensitivity and large specificity. However, a lower classification rate [11].
This shows that the best classification for the most optimal balance point for sensitivity and specificity of the model for the fitted data is the ROC.
The correct classification rate for the training set is 66.7% as shown in Table 2. The specificity is 58.5% and a sensitivity of 85.1% which is the largest amongst the ither approaches.
We also look at the testing set using the ROC, figure 7 shows that the cut off point for the optimal sensitivity and specificity is -0.809. The sensitivity is 75.4%, the specificity is 68.0% and the correct classification rate is 70.9%. This shows that the sensitivity is smaller by 9.7% and the specificity is larger by 9.5% compared to the training set. The correct classification rate increased by 4.2%

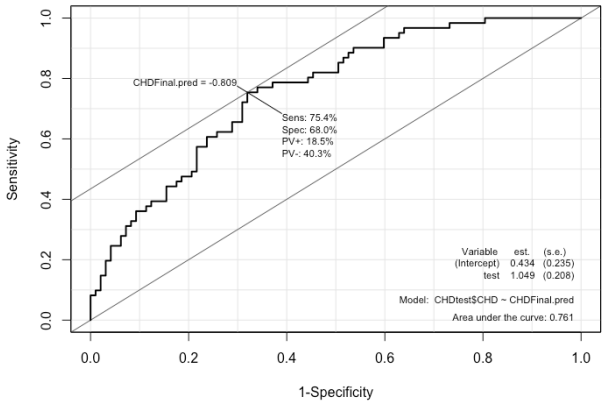

Figure 6: ROC plot for the testing set.

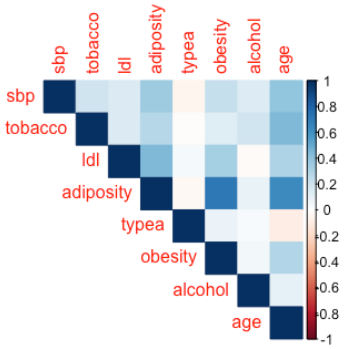

Figure 7: Correlation plot of the original data.

Question 3

We conduct a principal component analysis to grasp the variation in the original data without the CHD status.
We first check whether the principal component analysis will be using a correlation or covariance. The problem with the covariance is that there is varying scales and units used for example litres and kilograms [1]. Figure 7 shows the correlation of the data, most of the coefficients are within the range of 0.2 to 0.5. The data is mostly positively correlated. Therefore, a scaled correlation is used to carry out the principal component analysis instead of the covariance [1].

We also need to find out how many components provide a good representation of the data. This is done by looking at the cumulative proportion, PC1 has a cumulative proportion of 35.2%, PC2 has a cumulative proportion of 50.2%, PC3 has a cumulative proportion of 63.4%, PC4 has a cumulative proportion of 75.8%, PC5 has a cumulative proportion of 83.4%, PC6 has a cumulative proportion of 91.9%, PC7 has a cumulative proportion of 97.8% and PC8 has a cumulative proportion of 100%. This shows maybe 4 or 5 components would be suitable as it describes 75.8% to 83.4% of the variation of the data [12]. Kaiser criterion suggests that having 3 components since the variance is above 1 [12].
Looking at the Figure 8, we conclude that 4 components would be sufficient since the percentage of the explained variances does not change that much from 4 to 5 components (an elbow point) [12].

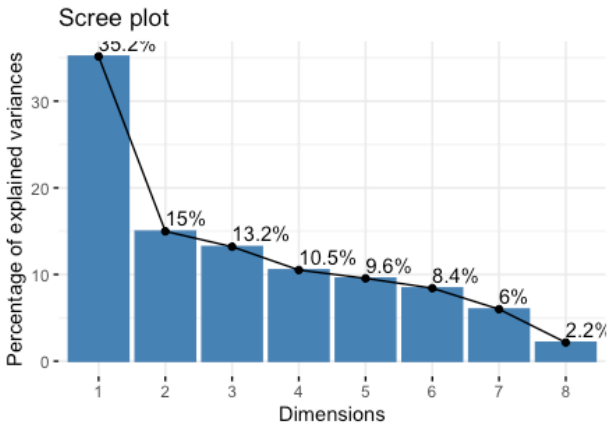

Figure 8: Scree plot of the eight principal components

Delving further into the loadings of the four principal components. We do not consider any loading less 0.2 positive or negative [12].
PC1 is an average of the variables since all the variables are negative, the percentage of body fat has the largest loading which means it contributes to PC1 the most.
PC2 is a contrast of the variables since there is mix of positive and negative loadings. The contrast is systolic blood pressure, tobacco, alcohol and age against lipoprotein cholesterol, type A and obesity. The alcohol contributes the most on the positive side and on the negative side obesity contributes the most to PC2.
PC3 is an average of two variables type A and alcohol with type A contributing the most to PC3.
PC4 is a contrast of variables. systolic blood pressure, obesity and alcohol against age, type A and lipoprotein cholesterol and tobacco. The tobacco contributes the most on the positive side and on the negative side it is alcohol that contributes the most to PC4 [12].

The scores of shows the position of the observation defined by the principal component. Figure 9 shows the six scores plot of the given principal components, the scores plot of PC1 vs PC2 shows that the scores are quite spread out [13]. This is seen by most the non-CHD (blue) patients have high value contribution with the PC1 and none with PC2, whereas the CHD patients have high value contributions with PC2 and a small value contribution to PC1, there are some with low value corresponding to the PC2 [13].
The other graph where there is spreading out is PC1 against PC3, this also consists of non-CHD patient high contributions with PC1 and none or low contributions with PC3. The CHD patients have a low contribution with PC3 and low contributions with PC1 (can be seen in the plot of PC2 vs PC3). The graphs of PC2 vs PC3, PC1 vs PC3, PC2 vs PC4 and PC3 vs PC4 shows that there is no or little to distinguish between them since the differing colours are mixed up [13].

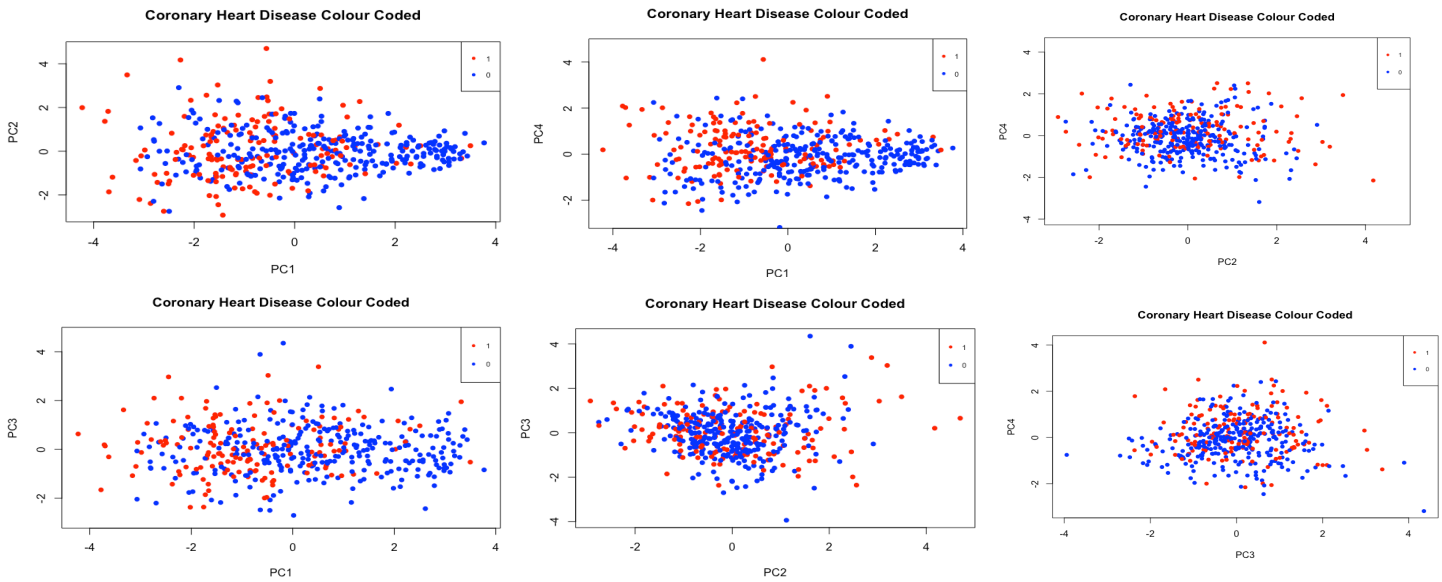

Figure 9: Six scores plot of the combinations of the four principal components

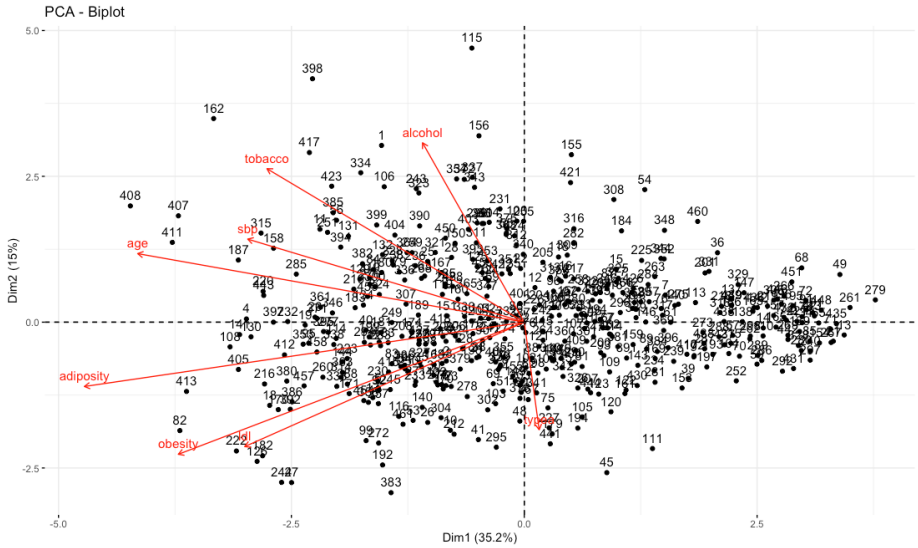

Figure 10: Biplot of PC1 and PC2

The biplot is used to show the relationships between the variables and the principal components. Figure 10 shows the biplots of the PC1 and PC2. All arrows are pointing to the left of PC1 which show that the high values of the variables correspond to low values of PC1. With the exemption of type-A behaviour pattern score, the arrow is on the right side which shows that that high values correspond to high values of PC1.
The four arrows point upwards have high values corresponding to high values PC2 and the other 4 arrows pointing downwards have high values corresponding to low values of PC2.

The angles between the variables and PC1 or PC2 shows high correlations.
For example, age has the smallest angle with PC1 which show that age has a large correlation with low values of PC1.
The same could be said with type-A behaviour pattern score where it has a high correlation with PC2. The angles between each variable also shows their correlation, body mass index and low-density lipoproteins cholesterol have a high positive correlation. Alcohol and low-density lipoproteins cholesterol have right angle which shows no correlation. The only obtuse angle is between alcohol and type-A behaviour pattern score which this shows a negative correlation.

The principal components show most of the variation is from PC1 and this corresponds to individuals who have not been diagnosed with CHD. The biplots show Type A has a high value with low values of PC2 and age has a high value corresponding to low values PC1. These to variables are also within the final model of the logistic regression. Therefore, the logistic regression and the principal components show that the type A behaviour pattern score, age and the log of the low-density lipoprotein cholesterol are all risk factors of coronary heart disease.

Abstract

Background: Coronary Heart disease (CHD) effects many males in Western Cape, South Africa. There are potential risk factors that could help identifies or predicts the diagnosis of CHD. We will use a logistic regression to fit a model with variables which are potential risk factors and a principal component analysis (PCA) to examine the variation in the data.

Method: The data contains 452 observations with ten variables the dimension of the data is 452 by 10. The data has eight different variable that could be potential risk factors. The data was summarised using different descriptive statistics and visualised using boxplots and histograms. An examination of covariance and a correlation of the variables was used. A logistic regression was carried out of the CHD status ("0" is not being diagnosed with CHD and "1" is diagnosed with CHD). Using a backwards selection for a suitable variable selection technique the appropriate model was fitted and assessed using deviance test and Akaike's information criterion (AIC) with analysis of the coefficients and confidence interval. The optimal sensitivity, specificity and correction rate was found by using different classification techniques such as a binary, ROC, linear, quadratic discriminant analysis, and k-nearest neighbour with cross validation.
The PCA is found by using observing the cumulative proportion with the help of the scree plot. This provides the best number of principal components need to carry on performing the PCA. The loadings analysed and the scores are plotted while the biplots are provided.

Result: We identified that there are up to three variables that are risk factors that could be used to predict CHD. The logistic regression provided a final model, $CHD = -7.39580 + 0.04047typea + 0.05607age + 1.25030Logldl + \varepsilon$, that had three different variables which were the age (OR=1.058, 95%CL= [1.033241, 1.082676129], P<0.00002) [7], type-A behaviour pattern score (OR=1.041, 95%CL= [1.011207, 1.072278771], P<0.0068) [7] and the log of the low-density lipoprotein cholesterol (OR=3.491, 95%CL=[1.714214, 7.111065574] , P<0.0006)[7]. We see that we four principal components were enough to describe 75.8% of the variation. This produced loadings that showed most the non-CHD have high value contribution with the PC1 whereas the CHD patients have high value contributions with PC2 and PC3. PC4 could not distinguished between them. Type-A behaviour pattern score had high contribution to low values of PC2 but had low contribution to high value PC1. Age has a high value corresponding to low values PC1.

Conclusion: The PCA showed the variations in the variables. These variables are also within the final model of the logistic regression. Therefore, the logistic regression and the principal components showed that the type-A behaviour pattern score, age and the log of the low-density lipoprotein cholesterol are all risk factors of coronary heart disease.

Key words: Coronary heart disease, Logistic regression, Principal components analysis, loadings, score, biplots.

Reference:

[1] Robertson, C.R, 2/6/2023, *10.8 Introducing the data-Correlation and covariance matrix*, Last Accessed 8/7/2023.
[2] Pyper, K.P, 07/03/2023, *8.7 Automatic variable Selection Backwards Selection*, Last Accessed 8/7/2023.
[3] Robertson, C.R, 2/6/2023, *7.17 Model Selection*, Last Accessed 8/7/2023.
[4] Robertson, C.R, 2/6/2023, *7.16 Deviance Test*, Last Accessed 8/7/2023.
[5] Robertson, C.R, 2/6/2023, *7.15 Comparing Model*s, Last Accessed 8/7/2023.
[6] Robertson, C.R, 2/6/2023, *7.11 Odd ratios, logistic regression and model fit*, Last Accessed 8/7/2023.
[7] Robertson, C.R, 2/6/2023, *7.12 Confidence intervals and linear regression*, Last Accessed 8/7/2023.
[8] Robertson, C.R, 2/6/2023, *7.27 Sensitivity, specificity, and ROC curves*, Last Accessed 8/7/2023.
[9] Robertson, C.R, 2/6/2023, *7.20 logistic regression for classification*, Last Accessed 8/7/2023.
[10] Robertson, C.R, 2/6/2023, *8.7 Model-based approach: LDA and QDA*, Last Accessed 8/7/2023.
[11] Robertson, C.R, 2/6/2023, *8.17 Applying knn-choice of k and cross validation*, Last Accessed 8/7/2023.
[12] Robertson, C.R, 2/6/2023, *10.10 Looking at numerical output-the PCs and loadings*, Last Accessed 8/7/2023.
[13] Robertson, C.R, 2/6/2023, *10.11 The scores and scores plot*, Last Accessed 8/7/2023.
[14] Robertson, C.R, 2/6/2023, *10.13 The biplots*, Last Accessed 8/7/2023.