The Final Project - IE 5561
South African Heart Disease

May 12, 2021

I acknowledge that all I did for this project is my work and understanding of the course material. The sample data used in this project is obtained from *https://web.stanford.edu/ hastie/ElemStatLearn/*, and I have read several articles from different resources, which can be found in the reference section of this project, on the Internet to get myself familiar with the data.

1 Introduction

One ethnic group of people in South Africa is white Afrikaans - a society of Africans locating at the Western Cape of South Africa and speaking the local language but are of European descent. In late 1970s, many people in this segment of people, especially the male ones, were reported to suffer from a coronary heart disease, and a result of this incident, Rossouw et al. (1983) researched about the lifestyles of people from three different communities of this society to check what factors are mostly contributed to that disease.

The data on *https://web.stanford.edu/ hastie/ElemStatLearn/* is a small part of the whole data, which Rossouw et al. (1983) collected, with having some but not all of the factors as the variables and 462 observations from 3357. Since male morality rates in these three communities were about two and a half times that of the females, all observations are from males.

In the current data set, rows 1-261 have row numbers matching the source ”row.name”, thereafter the row number is one less than the source ”row.name”. It would appear that observation with

”row.name” 262 is absent from the source. Therefore, although the last number in the ”row.name” is 463, we are provided with only 462 observations.Furthermore, the 462 observations consisted of all 160 cases having had coronary heart disease as well as 302 controls samples from the remaining set of survey obsevations.

There are 10 variables (sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, age, chd; 9 features and 1 outcome) and as explained above 462 rows in the data set. The explanations of the variables are hereunder:

- ”sbp”: systolic blood pressure in milimeteres of mercury (mm Hg).
- ”tobacco”: cumulative tobacco use in kg. Appears to be lifetime cumulative.
- ”ldl”: low density lipoprotein cholesterol
- ”adiposity”: another measurement of obesity
- ”famhist”: a factor indicating whether there is a family history of heart disease (Present, Absent)
- ”typea” : type-A coronary prone personality. Possible total scores can range from 12 to 84 and those with score of 55 or more are classified as exhibiting type A behavior in Rossouw et al (1983).
- ”obesity”: a measure of obesity; body mass index (BMI). having BMI > 30 scored as obese in Rossouw et al (1983).
- ”alcohol”: current alcohol consumption, Unit of measurement is not mentioned anywhere.
- ”age”: age in years at time of study
- ”chd”: the response variable, representing coronary heart disease, a factor identifying whether the subject had been diagnosed as having coronary heart disease or not. It is 1 if that subject had had coronary heart disease, and 0 if that individual had not had that disease.

Considering this data set, let consider ”chd” as the response and the other nine variables as the predictors. Through the project, we explore how every single predictor is related to the response. We further check the extent to which the predictors can affect the response. In other words, we answer the question of whether we need all predictor variables to predict the response. Moreover, we examine our data to find any possible interactions between predictors or nonlinearity, which

provide us with a better and more accurate prediction of the response.

We start fitting linear models to our data using almost all available methods in classification, since our response is a qualitative variable. We investigate what methods in classification give us the best result, i.e., the lowest MSE test. In addition, these methods allow us to understand the influence of predictors on the response. For instance, if an individual is from a family with a history of this disease, we explain the likelihood of that person diagnosed with that disease. In order to obtain a more accurate linear model with the predictors to the data, we further use *the best subset selection method, forward and backward stepwise selection, PCA, PLS* along with the *validation set* and *cross-validation* approaches to have good collections of training and test data.

To complete our examination of the data, we finally apply *tree-based* methods to the data and study the structure of the tree, pruning.

Questions: Based on this data, does having a family history of coronary heart disease affect a patients chance of having coronary heart disease? Does this results for patients younger than 40 years old? What about for patients aged 40 years or older?

2 Analysis the Data

Before jumping into fitting any model, we first study the data in detail. Based on the summary of the data provided by **R**, we can see the possible range of all quantitative variables. For example, the range of variable ”typea” is between 13 and 78. Also, the correlations between variables are checked using cor() function on R. Below is the table contained all correlations:

```
##                   sbp     tobacco         ldl   adiposity       typea
## sbp        1.00000000  0.21224652  0.15829633  0.35650008 -0.05745431
## tobacco    0.21224652  1.00000000  0.15890546  0.28664037 -0.01460788
## ldl        0.15829633  0.15890546  1.00000000  0.44043175  0.04404758
## adiposity  0.35650008  0.28664037  0.44043175  1.00000000 -0.04314364
## typea     -0.05745431 -0.01460788  0.04404758 -0.04314364  1.00000000
## obesity    0.23806661  0.12452941  0.33050586  0.71655625  0.07400610
## alcohol    0.14009559  0.20081339 -0.03340340  0.10033013  0.03949794
## age        0.38877060  0.45033016  0.31179923  0.62595442 -0.10260632
## chd        0.19235411  0.29971754  0.26305268  0.25412139  0.10315583
##              obesity     alcohol        age        chd
## sbp       0.23806661  0.14009559  0.3887706 0.19235411
## tobacco   0.12452941  0.20081339  0.4503302 0.29971754
## ldl       0.33050586 -0.03340340  0.3117992 0.26305268
## adiposity 0.71655625  0.10033013  0.6259544 0.25412139
## typea     0.07400610  0.03949794 -0.1026063 0.10315583
## obesity   1.00000000  0.05161957  0.2917771 0.10009508
## alcohol   0.05161957  1.00000000  0.1011246 0.06253068
## age       0.29177713  0.10112465  1.0000000 0.37297334
## chd       0.10009508  0.06253068  0.3729733 1.00000000
```

Figure 1: Correlations

According to Figure 2, ”chd” has the highest (positive) correlation with ”age” and the lowest (positive) correlation with ”alcohol”, which shows the high impact of age (compared to other variables) on the likelihood of being diagnosed with that heart disease. Also, ”obesity” and ”typea” has the highest (positive) correlation between all variables.

More information about the data can be found on pages 1-3 of Project.pdf.

3 Linear Models

Let $X = (sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, age)^T$ Since the response is a qualitative variable, we first fit a logistic regression to the data using all variables as predictors as follows:

$$\log\left(\frac{P(chd=1|X)}{P(chd=0|X)}\right) = sbp + tobacco + ldl + adiposity + famhist + typea + obesity + alcohol + age$$

In Figure 2, the summary of the model is shown. Not all variables have statistically significant p-values. The variables with significant small p-values are ”tobacco”, ”ldl”, ”famhistPresent”, ”typea”, and ”age”, with ”famhistPresent” has the smallest p-values among all predictors.

```
##
## Call:
## glm(formula = chd ~ sbp + tobacco + ldl + adiposity + famhist +
##     typea + obesity + alcohol + age, family = binomial, data = SA.Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7781  -0.8213  -0.4387   0.8889   2.5435
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
## sbp             0.0065040  0.0057304   1.135 0.256374
## tobacco         0.0793764  0.0266028   2.984 0.002847 **
## ldl             0.1739239  0.0596617   2.915 0.003555 **
## adiposity       0.0185866  0.0292894   0.635 0.525700
## famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
## typea           0.0395950  0.0123202   3.214 0.001310 **
## obesity        -0.0629099  0.0442477  -1.422 0.155095
## alcohol         0.0001217  0.0044832   0.027 0.978350
## age             0.0452253  0.0121298   3.728 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

Figure 2: Summary of the Logistic Regression with All Variables

This means that these variables have the highest influences on the response, and we can remove all other variables such as ”sbp”, ”adiposity”, ”obesity”, and ”alcohol” from the model, and as a result, we have a new logistic regression with fewer variables involved in the model.

4 Reference

https://web.stanford.edu/ hastie/ElemStatLearn/

C. Wyndham (1982) ”Trends with time of cardiovascular mortality rates in the populations of the RSA for the period 1968-1977”, South African Medical Journal, 61, 987-993.

https://great-northern-diver.github.io/loon.data/reference/SAheart.html

http://www2.stat.duke.edu/ cr173/Sta102$_{S}p14/Project/heart.pdf$