A statistical Analysis for Genes and Environmental Factors Involved in Blood Pressure Variation

August 16, 2015

Table of Contents

List of Tables

List of Figures

Abstract

Purpose of this report is to identify true genetic markers along with clinical covariates as the best predictors of systolic blood pressure. Different regression models analyses were applied to this data. The goodness of fit for each model is also given to enable the comparison. We used multiple and logistic regression to model the relationship between predictors and systolic blood pressure. Forward, backward, stepwise and subset selection methods are used for variable selection. As we have large number of covariates in the data set, regularization methods: LASSO and Elastic net are applied in order to have more efficient results.

1. Introduction

Genes play key role in the development and growth of disease and they also affect how individuals react to medicines. A considerable part of recent medical research is dedicated to the detection of genetic markers that can be used to identify disease. These genetic markers can help diagnosis and risk assessment. In genetic and genome studies, usually hundreds of genetic markers, together with many clinical and environmental measurements, are collected. Statistical methods are helpful in predicting true genes. Predictive modeling is a statistical tool that builds a prediction function from the observed data. Regression is a commonly applied predictive modeling method that has been used to a wide range of application domains. In this report, we build several regression models of blood pressure using the data set of 500 subjects generated based on a complex genetic model developed at GSK. The data set for this report has been taken from the website of statistical society of Canada[1].

[1]http://www.ssc.ca/en/education/archived-case-studies/case-studies-for-the-2003-annual-meeting-blood-pressure#references

There are 500 predictors (483 genetic markers and 17 clinical covariates). The goal is to identify the best predictors among the 500 variables. For variable selection, Backward and Forward Stepwise variable selection methods, Cp and R-square adjusted subset selection and regularization methods are used. For analyses of data the statistical packages, R Version 3.2.1 (GLMNET, CAR and MASS packages) and SPSS Version 20 were used.

This report is organized as: In Section 2, a description of outcome variable used in the statistical analyses for this study is given. Descriptive analysis, including bar charts, pie charts and box plots, comparison of groups are discussed in Section 3. An outline of statistical methods used in the analysis is given in Section 4. Section 5 presents results of the statistical analyses. Statistical conclusions based on the results presented in Section 5 are given in Section 6. Finally, all relevant tables and figures generated from statistical analyses are provided in Appendix A and B, respectively. Relevant output for statistical analyses appears in Appendix C.

The data set contains 500 observations (subjects) and 501 variables. Of the 500 subjects, 250 had low blood pressure and 250 had high blood pressure (i.e. hypertension). The 501 variables consist of one response variable (systolic blood pressure) and 500 predictors (17 clinical covariates and 483 genetic markers). Among continuous covariates are age (in years), weight (in pounds), height (in inches) and BMI. Categorical variables are gender, marriage, smoking, stress, overweight, race, alcohol, treatment, exercise level, income, salt intake, child bearing and education level. For all the subjects, age range between 18 and 64. The study group consisted of 264 females and 236 males. One third of the subjects are younger than 50 years of age and only 7 percent are older than 60 years of age.

2. Outcome Variables

When heart beats, it contracts and pushes blood through the arteries to the rest of the body. This force creates pressure on the arteries. This is called systolic blood pressure. A normal systolic blood pressure is below 120. A systolic blood pressure of 120 to 140 indicates pre-hypertension, or borderline high blood pressure. Hypertension is arbitrarily defined as a systolic blood pressure greater

than 140, or commonly known as high blood pressure. Even people with pre-hypertension are at a higher risk of developing heart disease. High blood pressure (hypertension) can quietly damage a body for years before symptoms develop. Possible health consequences that can happen over time when high blood pressure is left untreated includes: Damage to the heart and coronary arteries, including heart attack, heart disease, congestive heart failure, aortic dissection and atherosclerosis (fatty build-ups in the arteries that cause them to harden), stroke, kidney damage, vision loss, memory loss, fluid in the lungs and angina. There are several risk factors that potentially contribute to the high blood pressure, such as age, race, weight or exercise level etc. Risk increases even more if one has high blood pressure along with other risk factors: age, heredity (including race), gender, overweight or obesity, smoking, high cholesterol, diabetes and physical inactivity. Hypertension is a classic example of a complex genetic attribute.

It is believed that there are several genes, which contribute to the variation in blood pressure. These genes interact with environmental and clinical factors such as, salt intake, stress, inactivity, excess alcohol consumption and body weight to reach the final disease. In this data set systolic blood pressure varies between 67 and 224 and average systolic blood pressure is 144.95. Detailed statistics for the blood pressure is given in Appendix A Table 1 and bar chart for blood pressure is shown as Figure 1, Appendix B. Of the 500 subjects, 11 are with hypotension (SBP less than 90), 72 are with normal blood pressure (SBP less than 120) 167 are with pre-hypertension (SBP between 120 and 140). In overall population 50 percent have high blood pressure (SBP greater than 140) out of which 13% can develop life threatening complications (as their SPB is higher than 180). In section 4, we define new categorical response variable y as 1 if systolic blood pressure is greater than 140 and 0 otherwise.

Main purpose is to check whether the incidence of hypertension (i.e high blood pressure > 140) can be predicted based on clinical covariates and genetic markers. A binomial logistic regression is run to determine whether the presence of heart disease could be predicted from the covariates.

3. Descriptive Analysis

In this section, a description of the data is presented through basic summary statistics and graphics. All the available data are used to compute these descriptive statistics. The exploratory analysis including the mean, median, mode,

standard deviation, minimum and maximum of the data is carried out and is presented in Table 1 to Table 3 Appendix A. To visualize the data, bar charts, box plots and pie charts are constructed. Box plots are a good way to identify outliers, so box plots are drawn for each categorical variable to detect unusual observations. Scatter plots are also constructed to show the rough relationship between each continuous factor and the systolic blood pressure. All these graphs are presented in Figures 1-5, Appendix B. Average age of subjects in this data set is 40 years and range between 18 to 64 years. One third of the subjects are younger than 50 years of age and only 7 percent are older than 60 years of age. In overall population 50 percent of the people are with high blood pressure out which out which 13% can develop life threatening complications (as their SBP is higher than 180). Hypertension group (systolic blood pressure >140) consists of 110 men with mean systolic blood pressure 167.74 and 140 women with mean 164.81. In this group 228 patients are not taking any treatment and their mean blood pressure is 167.64 and only 22 patients are taking treatment and have mean blood pressure 150.05.

Number of smokers in hypertension group are 142 with average blood pressure 169.83 and non-smokers 108 with mean blood pressure 161.19. In the category with systolic blood pressure ≤140, there are 126 males and 124 females, with average systolic blood pressure 125.19 and 122.40 respectively. Systolic blood pressure for smokers in this category is 127.35 that is higher than the non-smokers within the same group. Subjects that are not taking treatment in this group are 171 with average blood pressure 118.82 and those who taking some treatment are 79 with average blood pressure 134.59. Some preliminary analysis is given below to see which groups differ in systolic blood pressure level.

More females have participated in this study then males with average blood pressure 144.89 and145.02 respectively. To see whether male and female significantly differ on their systolic blood pressure level we use independent t-test assuming equal variances and find p-value 0.958 (Table 4) suggesting no significance difference of SBP among males and female. Average blood pressure for smokers is 150.03 and for non-smokers is 139.18. The two groups are significantly different as can be seen by the p-value near to zero in Table 4. The systolic blood pressure level does not vary among married and unmarried people as can be seen from Table 4. Box plot (Figure 3) for treatment and non-treatment groups depicts many outliers in the first group. There is big difference among the

number of subject receiving treatment and not receiving treatment. Only hundred 101 subjects are taking some treatment and this number gets even low for high blood pressure group as only 22 subjects with hypertension are going through some treatment. Table 2 shows that the mean SBP is higher for the subject who are not taking treatment as compare to those who are taking some treatment. Sufficiently small p-value in Table 4 suggests that SBP is significantly different among two groups. From the box plots (Figure 3) we can see that the group with high level exercise is different than the other two groups. For further verification we used one way ANOVA (Table 5) and conclude that the three groups are significantly different. For the overweight category people are divided into three categories: normal over weight and obese. Mean SBP is higher for obese group as compare to the other two groups. ANOVA (Table 5) supports the fact depicted by boxplot Figure 3 that there groups are different with p-value near to zero. Drinking too much alcohol can raise blood pressure levels. From the box plot (Figure 3) we can see that the group with high level of alcohol intake is different from the other two groups. ANOVA (Table 5) also verifies that the groups with different level of alcohol are significantly different in SBP.

4. Statistical Methodology

4.1. Using Clinical Covariates

Method1: Multiple Regression/Variable selection based on Forward, Backward and Stepwise Method

Method 2: Multiple Regression/Variable selection based on Subset Selection

Objective: To model the relationship between clinical and genetic markers predictors and systolic blood pressure and recognize best clinical predictors.

Multiple linear regression model describes a relationship between response variables in the case of more than one regressor variable. The term linear is used because the relationship is a linear function of the unknown parameters. To achieve the best model, first we suppose that all regressors included in the model are important and then we investigate about each variable [9].

In most studies, we expect only few variables are likely to be important. Finding an appropriate subset of regressors for the model is called the variable selection problem. Good variable selection methods are very important in the

presence of multicollinearity. The most common corrective technique for multicollinearity is variable selection. Variable selection does not guarantee elimination of multicollinearity [9].

In some applications based on prior information, we can select some variables which are more important than the others. For present study we do not have any extra information about the variables [9]. So we use another approach for variable selection which is not based on previous information of variables. These methods can be applied to select the best covariates depending on the statistical significances. We apply the classical approach to regression model selection.
The strategy to identify the best model is: First, fit the full model (the model with all of the regressors). Second, carry out a thorough analysis of this model, including a full residual analysis. Often, we should perform mentioned analysis to investigate possible collinearity. Third, determine if transformations of the response or of some of the regressors are necessary. Forth, using the t tests for the individual regressors to find the best the model. Fifth, perform a thorough analysis of the edited model, especially a residual analysis, to determine the model's adequacy [9].

The use of good model selection techniques help us to increase our confidence in the final model or models recommended. Although ideally fitting best model should be solved simultaneously, an iterative approach is often employed, in which (1) a particular variable selection strategy is employed and then (2) the resulting subset model is checked for correct functional specification, outliers, and influential observations. This may indicate that step 1 must be repeated. Several iterations may be required to produce an adequate model[9].
Various methods have been developed for selecting the number of subset regression. Three of the popular approaches are Stepwise, Backward and forward which works by adding or deleting regressors one at a time[9].
Forward selection begins with the assumption that there are no regressors in the model other than the intercept. Our aim is to find an optimal subset by inserting regressors into the model one at a time. Back ward selection begins with no regressors in the model and attempts to insert variables until a suitable model is obtained. Stepwise regression is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed.

Mallows's Cp Statistic Mallows has proposed a criterion that is related to the mean square error of a fitted value. When using the Cp criterion, it can be helpful to visualize the plot of Cp as a function of p for each regression equation.

We are looking for the number of covariates which is near to values of Cp (Cp = p)[9].

Coefficient of Multiple Determination is a measure of the adequacy of a regression model that has been widely used is the coefficient of multiple determination.Use *R*-squared values to find the point where adding more predictors is not worthwhile, because it yields a very small increase in *R*-squared. According to this criterion, the best regression model is the one with the largest adjusted *R*-squared[9].

We also have investigated on influenced data. Influence data, such a data value is not only remote in terms of the specific values for the regressors, but the observed response is not consistent with the values that would be predicted based on only the other points. We employed the DFBETAS for each model variable, DFFITS, covariance ratios, Cook's distances and the diagonal elements of the hat matrix. Cases which are influential with respect to any of these measures were marked[9].

4.2. Using Clinical Covariates and Genetic Markers

Method: Regularization methods / LASSO/Elastic Net.
Objective: model the relationship between predictors and systolic blood pressure and recognize best genetic marker predictors.

The linear regression model is proposed with $x_1, x_2, \ldots., x_p$ which are defined as p predictors. The response variables as y is predicted by:

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \cdots$$

Where $\{\hat{\beta}_p\}$ are coefficients which are obtained by the ordinary least squares (OLS). This method estimates parameters by calculating minimum of the residual sum of squares (RSS).

To evaluate the quality of a model, some criterion must be applied according to the circumstances. One of the important ones is accuracy of prediction on future data. However meeting this criteria is hard, but it is essential. Second, is that model should be Interpretable. Simpler model is preferred to understand more about relationship between response and covariates. Third, is making decision based on minimum mean square of the errors [2].

Estimating and choosing the parameter becomes an important issue when the number of predictors is large. As it is known that in such a case, OLS is achieved often poor estimate in both prediction and interpretation. There are still room for improvement, penalization techniques have been proposed to improve OLS. The technique called the lasso is a penalized least squares method. This method forces a L1 penalty on the regression coefficients. We should take into the account that the nature of the L1 penalty presents this opportunity that the lasso does both continuous shrinkage and automatic variable selection simultaneously. [3]

Although the lasso has shown success in many situations, it contains some limitations. We consider the two important cases: first one is when the p>n, the lasso selects at most n variables, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. In some situations, we have a group of variables among which the pair wise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected. [3]

Another regularization technique is known as elastic net. This method is similar to the lasso. The elastic net performs simultaneously automatic variable selection and continuous shrinkage, and is able to select groups of correlated variables.

GLMNET is a package in R program which is proposed for computing the entire regularization methods with the computational effort of a single OLS fit. This package does generalize linear model via penalized maximum likelihood. In this package based on our selection, regularization is employed by the lasso or elastic net penalty at a grid of values for the regularization parameter lambda. [4]

In the other word, GLMNET package solve the equation:

$$\min \frac{1}{N}\sum_{i=1}^{N} w_i I\left(y_i, \beta_0 + \beta^T x_i\right) + \lambda\left[\frac{(1-\alpha)\|\beta\|_2^2}{2} + \alpha\|\beta\|_1\right]$$

To find the minimum value λ, the elastic net penalty is controlled by lambda. The parameter λ manages the overall strength of the penalty. [4]

For the second parameter α in 1 elastic net method, mixes two passible choices of ridge penalty and lasso penalty. Based on correlation on predictors and trial and error, α is selected. It is called elastic net mixing parameter. The range of this parameter is between $\alpha \in (0,1)$. In this study alpha is assumed 0.6.

4.3. Regularization Methods with respect to Binomial Distribution of New Response Variable:

Method1: Regularization methods / LASSO/Elastic Net
Objective: To model the logistic relationship between predictors and indicator of blood pressure groups and recognize the best clinical and genetic marker predictors.

Logistic regression is applied to model categorical outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables. Dependent variable in logistic regression has two possible discrete outcomes. This model is used to predict the probabilities of the different possible outcomes of a categorically distributed response variable, given a set of independent variables. The independent variables could be continuous or categorical. [1]

Although systolic blood pressure is continuous. Assigning two categories to patient with high and low blood pressure instead of numeric systolic blood pressure, makes the outcome variable discrete. This assumption helps us to define a better model as well as a better prediction. Moreover, regularization and variable selection via the LASSO and the elastic net approach can be used in this situation. It is worthwhile to work with categorical value because the errors will be decreased in this situation. [1]

In the logit model, the log-odds of each response can be written as a linear model. Furthermore it is common that one of the response level, particularly, corresponded to zero, is fixed as the reference level. [2]

$$n_{ij} = log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha + X_i \beta_j$$

Where α_i is a constant and β_j is a vector of regression coefficients, for j = 1, 2, . . . J-1. The J − 1 multinomial logit equations contrast each of categories 1, 2, . . ., J − 1 with category J. [5]

For the analysis, the first category of outcome variable which correspond to "zero" is picked as a baseline and calculate the odds ratio and other estimated parameters based on this level.

The multinomial logit model may also be written in terms of the original probabilities π_{ij} rather than the log-odds. It can be written:

$$\pi_{ij} = \frac{\exp\{n_{ij}\}}{\sum_{k=1}^{J} \exp\{n_{ik}\}}$$

For j = 1, . . ., J. Based on these definitions, we will define and examine the model for the best fit and find the best predictors in section 5.2.3.

5. RESULTS

5.1. Variable Selection with respect to Clinical variables

5.1.1. Full Model

We have fitted the model based on all clinical variables which called full model. Table 7 shows that some of coefficient in the model are significant. It means that those coefficient are more predictive than the others for systolic blood pressure. The coefficients related to covariates smoke, exercise, alcohol, treatment are significant (the p-values are less than 0.05). This means that there are a linear relationship between systolic blood pressure and doing exercises, drinking alcohols, being under treatment and use of smoke. As shown in Table 6, R^2 equal to 0.2 indicates that about 20% of the variation in systolic blood pressure can be only explained by the relationship to Clinical predictors. It is noticeable that adjusted $R^2$is less than R^2 as well. This result leads us to improve our model. This may occur because of many reason such as multiliniarity. Based correlation of regressors, we can conclude that there is a huge multiliniarity in this model. We should perform a thorough analysis to investigate possible co-linearity.

However, based on predictors definition in model, overweight covariates should be highly correlated with weight, similarly Body Mass Index should be highly correlated with weight and height, excluding those variables did not help

us a better model. We preferred to keep those variables into the model. Table 7 and 8 show models which mentioned variables were excluded.
We should always consider the validity of the model assumptions to be doubtful and conduct analyses to examine the adequacy of the model. We have presented several methods which are useful for diagnosing violations of the basic regression assumptions (as seen in Figure 6). It exhibits there is some deviations from the normality in tail which is caused by truncated data. In this data set we have some influence data which are indicated in this Figure as well. We have used several methods to stabilize the variation including the transformation techniques for response and predictors. As seen in Figure 7, there is no specific trend in their model which we can remove them easily. This figure have been indicated that maybe there is no linear relationship between systolic blood pressure and clinical covariates, and we have to use some complex regression model to find this relationship. We still wants to work on linear relation, there is still some chance for improving the model.

To choose the appropriate model, checking the influence data is necessary. Among all patients, we were trying to diagnosis of influential data and eliminate them. We should omit influence data (Table 10) and fit the model again. By deleting the influence data (Sample Number 8, 32, 231, 243, 339, 355, 356, 366, 375, 403, 374 and 485) a little improvement accrued in model fitting. Adjusted R^2 were increased by 5%.

5.1.2. Variable selection via Stepwise, Backward, Forward:

The analyses has a rather large pool of possible candidate regressors, of which only a few are likely to be important. Finding an appropriate subset of regressors for the model is often called the variable selection problem.
Tables 11 to 13 present the results of using the stepwise, backward and forward selection methods. We have specified the alpha level for either adding or removing a regressors as 0.05.

However, it has been noted that forward selection tends to agree with all possible regressions for small subset sizes but not for large ones, while backward elimination tends to agree with all possible regressions for large subset sizes but not for small ones. Both the stepwise and backward selection techniques suggested the same variables: married, smoke, age, exercise, height, alcohol, treatment, BMI and income.

Note that variable selection methods do not necessarily lead to the same choice of final model. Although in this study all methods except forward

selection terminates with the model including covariates: married, smoke, age, height, alcohol, treatment, BMI and income.

We added interactions between the covariates selected from above procedure. There is some interactions effect, which is negligible because of small effect on the Model. As shown in Table 14.

Based on the final model, Residual standard error is 23.96. Multiple R-squared is 0.1994 and adjusted R-squared equal to 0.1843. Compared to full model Residual standard error reduced from 25.32 to 23.96. (Results in Table 15)

On the whole, we should feel comfortable recommending this model based on selected covariates: married, smoke, age, exercise, height, alcohol, treatment, BMI, stress and income. We have done the goodness of fit for this model as well (Figure 9). It seems the model fit well although the R-square is small and the mean square error is 543.9331.

Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while keeping other predictors in the model constant. For instance: If age or height differed by one unit, and other covariates do not differ, systolic blood pressure will differ by o.2 or 0.5 units, on average, respectively. In Smoke group, we would expect that a smoker will have 10 units higher systolic blood pressure than a Non-smoker, on average, keeping all other covariates same.

From this model we can interpret that by receiving the treatment, there are 13 units increased in systolic blood pressure. Doing strenuous exercise has fallen 15 units in systolic blood pressure. In contrast, smoke and alcohol increased the chance of high systolic blood pressure. Results shown in Table 15.

5.2. Variable Selection with respect of All Covariates:

5.2.1. Regularization Method (LASSO and Elastic Net) with respect to All Covariates

This section aims to choose an appropriate model for the systolic blood pressure based on the clinical variables and genetic markers. To achieve this aim, we should, firstly, add all independent variables into the model and find the most effective ones. We used two different approaches which are LASSO and Elastic net, both with respect penalty Factor and without penalty factor. Those approaches led us to obtain best predictors based on LASSO methods with respect to minimum mean square error. By comparing predicted values and

observed data and calculating mean of square error, we can conclude that using penalty factor might not have helped us to get better predictions. However, even without using this method, we were able to provide accurate predictions.

LASSO and Elastic net with respect to alpha equal to 0.6 has been applied to analyze the data. Alpha equal to 0.6 is obtained by trial and error. To fit generalized linear model, the number of lambda is defined by default in the GLMNET package in the R software.

Figures 10 illustrates a plot which is estimated by using the cross-validation. The optimal value of λ, which gives the minimum mean square error, is extracted. Base on this result, lambdas are 1.34 and 1.36 in LASSO and Elastic Net methods, respectively. Although both methods selected the same clinical and genetic markers variables as best predictors, but by comparing the mean square error of these methods, we conclude that LASSO method works a little better than Elastic Net. (MSE LASSO=381.43 and MSE Elastic net=385.96)

There is a coefficient function extractor which works on a cross validation object and picks the coefficient vector that corresponds to the best model. We retain only a subset of variables, eliminating the rest from the model. Consequently, the best subset regression model has been found. This subset gives the smallest residual sum of square. As Table 16 shows, we can extract the coefficient with respect to the obtained values of lambda. Mean square errors related to these models confirm that the predicted and actual values are rather close with using Lasso regularization, although there is not much difference between MSE from LASSO model and Elastic Net method(MSE LASSO=381.43 and MSE Elastic net=385.96).

Based on Table 15, we can say that 8 out of 17 non-zero clinical variables including: married, smoke, exercise, weight, overweight, alcohol, treatment, BMI and Stress and 58 out of 483non-zero genetic markers including:g7, g9, g10, g46, g48, g50, g59, g63, g86, g92, g108, g120, g122, g135, g137, g150, g160, g168, g169, g175, g179, g182, g187, g191, g200, g204, g222, g231, g232, g271, g279, g288, g289, g292, g295, g298, g309, g330, g337, g348, g356, g362, g364, g366, g371, g377, g391, g411, g412, g422, g425, g438, g443, g447, g453, g465, g469, g480, are identified to achieve the best model with respect to systolic blood pressure.

On the basis of the results, the coefficient values of the genetic marker 50 and 200 among other genes have greatest effect on systolic blood pressure (related coefficients are 13 and 10 respectively). The genetic markers number 298 and 453 have coefficient value 2.5 and 2.27 respectively. The rest of coefficient values are very small (values are between zero and one). In this model three high correlated covariates such as weight, overweight, BMI are chosen in the model. We have extracted one of these variables and fit the model. The results have not changed too much. We decided to keep them in the model and let the methods choose the best ones.

From this model, if we consider that other covariates are constant, one can expect that: systolic blood pressure of smokers is 6 units higher than Non-smokers; systolic blood pressure of obese patients is 6 units higher than normal patients; the use of alcohol increases the patients' systolic blood pressure by 2 units; patients who do exercise regularly have 6 units lower systolic blood pressure than others; and, patients under treatment have 12 units lower systolic blood pressure than patients who do not receive any treatment.

We fitted the model based on selected genetic markers. It is noticeable that mean square error with respect to LASSO equal to 381.63 which is less than mean square error based on clinical covariates model (MSE=532). Adjusted R-square increased as 0.52 which is really good in comparison to the model based on clinical covariates.

We assume that grouping the data, into high and low blood pressure study, might have been helpful to find the better model for the two groups. In the next section, we have divided the dataset in two groups with respect to information provided on website.

5.2.2. Regularization Method (Elastic Net) with respect to Low Blood Pressure / High Blood Pressure Patient Groups

From the provided information on the website all patients are divided into two different groups such as low blood pressure and high blood pressure. We attempt to find the best model with respect to this division. Considering all clinical and genetic markers variable: effective covariates and genetic markers are: smoke, overweight, alcohol, g50, g200, g377, g432, g458, g459 for high blood pressure group. Similarly, age, weight, treatment, BMI, g5, g7, g9, g47, g95, g111, g112, g132, g169, g177, g191, g200, g205, g207, g214, g216, g242, g253, g285, g327,

g357, g359, g377, g385, g393, g427, g434, g464, g474, g478, g480 are diagnosed as the significant factors and genetic markers in low blood pressure group. These genetic markers are obtained from Elastic net regularization method. When number of parameter are more than the number of sample, Elastic net is recommended strongly. Effective genetic markers for both groups are different. Systolic blood pressure related to patients in low blood pressure group is not influenced by same genetic markers (Table 17 and 18).

It is clear that smoke, overweight and alcohol have an effect on systolic blood pressure in high blood pressure group. In contrast, age, weight, treatment and BMI have an impact on systolic blood pressure in low blood pressure group. The effective genes number "g200" appeared in both model.

5.2.3. Regularization Methods with respect to Binomial Distribution of new Response Variable:

The objective of this section is to collect the most effective independent variables on the systolic blood pressure based on using categorical variables instead of continuous variables. As previously stated, there is some extra information about each patients, such as high blood pressure and low blood pressure. Based on this information data can be classified into two study groups.

Based on the information provided, we have two group: patients with high blood pressure and patient with low blood pressure. We assign two categorical values into these groups. For selecting best predictor, Regularization methods have been carried out, assuming binomial distribution for response variable. With this assumption, to obtain best model we use logistic regression. First step is to add all covariates into the model. In this step, we employed regularization method such as Lasso and Elastic net.

To assess the impact of the dependent variables on the independent variable, logistic regression is applied. From these two approaches, the LASSO regularization method with penalty has been resulted a minimum mean square error compare to obtained MSE in Elastic Net method (mean square error is equal to 0.62). The result shows that stress, BMI, treatment, overweight, exercise, married and the genetic markers including g10, g36, g49, g50, g65, g75, g86, g98, g120, g122, g137, g150, g168, g187, g191, g200, g204, g231, g279, g298, g309,

g330, g385, g391, g412, g425, g447, g450, g453, g460, g469, g475 are most important clinical variables and genetic markers respectively. This model indicates that the model is more reasonable than previous ones (Table 19). Both genes 50 and 200 appeared in the model. The selected clinical covariates are different from previous ones. As a results of this model, we can calculate the probability of having High or Low blood pressure given selected covariates.

6. Conclusions and Recommendations

In this report, using the given clinical and genetic markers information of 500 patients, we aimed to investigate which variables effect systolic blood pressure. Three approaches were adopted to find the most efficient predictors on systolic blood pressure. We used multiple regression and variable selection methods; for instance forward, backward, and stepwise variable selection, Cp and adjusted R-square considering only clinical variables. Regularization method, such as LASSO and Elastic Net, were employed to find the best predictors based on clinical and genetic markers variables. We calculated a categorical variable based on two different groups: the high blood pressure and the low blood pressure. Considering logistic regression, regularization methods were applied to the new categorical response variable. Using regularization method, we were able to find the best covariates and models to fit the dataset.

From descriptive analysis, we can conclude that treatment is an important factor. In overall data, only 20% subjects are taking treatment and, unfortunately, only 9% are taking some treatment in the hypertension group. Average blood pressure for subjects not being treated is significantly higher as compare to the patients being treated.

From the analysis of clinical predictors, no firm conclusion can be drawn on the linear effect of some variables. This may be resulted from the non-linear relationship between clinical variables and systolic blood pressure or highly correlation between clinical variables. Thus, to achieve a better conclusion on the results, complicated method in regression analysis should be used or genetic markers should be added in the model.

To address the inefficiency of models with the all clinical variables, we have applied regularization via lasso and elastic net method in order to find the association between systolic blood pressure and clinical and genetic markers

variables. This led us to conclude that there is an association between some of genetic markers and clinical variables and systolic blood pressure. Thus, we can introduce a model with selected coefficients, achieved through regularisation, as the best model with all patients and in both groups: high blood pressure and low blood pressure as well.

As a result, the model for all patients, which have included some of variables, including married, smoke, exercise, weight, overweight, alcohol, treatment, BMI and stress and 58 out of 483 genetic markers including g7, g9, g10, g46, g48, g50, g59, g63, g86, g92, g108, g120, g122, g135, g137, g150, g160, g168, g169, g175, g179, g182, g187, g191, g200, g204, g222, g231, g232, g271, g279, g288, g289, g292, g295, g298, g309, g330, g337, g348, g356, g362, g364, g366, g371, g377, g391, g411, g412, g422, g425, g438, g443, g447, g453, g465, g469, g480, were identified to achieve the best model with respect to systolic blood pressure. Based on this model, if we consider that other covariates are constant, we can expect that: 1) smokers have 6 units higher systolic blood pressure than Non-smokers; 2) Obese patients have 6 units higher systolic blood pressure than normal patients; 3) the use of alcohol increases the patients' systolic blood pressure by 2 units; 4) patients who do exercise regularly have 6 units lower systolic blood pressure than others; 5) and, patients under treatment have 12 units lower systolic blood pressure than patients who do not receive any treatment.

We can also conclude that two genetic markers, "g50" and "g200", seem to be very important in predicting systolic blood pressure as they appear in most of the models. Different combinations of these genetic markers increase systolic blood pressure by 13 and 10 units, respectively.

We recommend a further study to investigate the association between systolic blood pressure and clinical and genetic markers variables based on the principal component analysis and clustering methods to bring out strong patterns in predictors.

References

[1] *David W. Hosmer, Jr., Stanley Lemeshow; Rodney X. Sturdivant, "Applied Logistic Regression, 3rd Edition" , John Wiley & Sons, New Jersey, 2013.*

[2] *Zou, Hui, Hastie, Trevor, "Regularization and Variable Selection via the Elastic Net". Journal of the Royal Statistical Society , Series B: 301–320, 2005.*

[3] *Friedman, Jerome; Trevor Hastie; Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent", Journal of Statistical Software: 1–22, 2010.*

[4] *http://cran.r-project.org/web/packages/glmnet/index.html*

[5] *http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html*

[6] *Rodríguez, G. Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/wws509/notes/, 2007.*

[7] *Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome , "The Elements of Statistical Learning, Data Mining, Inference, and Prediction", Second Edition , 2009.*

[8] *http://www.ssc.ca/en/education/archived-case-studies/case-studies-for-the-2003-annual-meeting-blood-pressure,*

[9] *Douglas C.* Montgomery*, Elizabeth A. Peck, G. Geoffrey Vining,* Introduction to Linear Regression Analysis*, 3rd Edition 2013.*

Appendix A: Tables

Table 1. Descriptive for Continuous variables

	sbp	age	weight	height	bmi
N	500	500	500	500	500
Mean	144.95	40.20	166.64	65.33	27.66
Median	140.50	40.00	168.00	65.00	27.00
Std. Deviation	27.995	13.299	40.903	6.191	8.559
Minimum	67	18	90	54	11
Maximum	224	64	249	77	53

Table 2. Summary of systolic blood pressure by gender married smoke and treatment

Gender			Married		Smoke		Treatment	
	Male	Female	Yes	No	Yes	No	Yes	No
]Count	236	264	239	261	266	234	101	399
Mean	145.02	144.89	146.72	143.33	150.03	139.18	146.72	137.96
S. D	27.664	28.340	28.719	27.269	27.490	227.497	30.842	7.942
S. E	1.801	1.744	1.858	1.688	1.686	1.798	1.544	0.759

Table 3. Summary of systolic blood pressure by exercise overweight alcohol race stress salt child bearing income education

8/	N	SBP(Mean)	S.D	Lower Bound	Upper Bound	Minimum	Maximum
Exercise							
1	195	150.12	27.589	146.22	154.01	95	224
2	136	142.87	26.138	138.44	147.30	73	216
3	169	140.67	29.102	136.25	145.09	67	215
Overweight							
1	187	136.32	27.269	132.38	140.25	67	207
2	109	144.37	25.079	139.61	149.13	100	213
3	204	153.18	27.814	149.34	157.02	101	224
Alcohol							
1	160	141.46	29.539	136.84	146.07	72	224
2	167	142.59	26.148	138.60	146.59	67	222
3	173	150.46	27.567	146.33	154.60	102	215
Race							
1	355	144.79	28.817	141.78	147.79	67	222
2	99	145.86	26.735	140.53	151.19	84	224
3	25	148.28	27.149	137.07	159.49	101	191
4	21	139.52	20.469	130.21	148.84	104	183
Stress							
1	151	142.35	28.287	137.80	146.90	77	224
2	175	145.14	27.675	141.01	149.27	67	212
3	174	147.02	28.040	142.82	151.21	72	222
Salt							
1	166	145.79	27.465	141.58	150.00	77	216
2	157	145.34	30.063	140.60	150.08	72	224
3	177	143.82	26.677	139.87	147.78	67	210
Child							
1	236	145.02	27.664	141.47	148.57	72	224
2	143	142.73	26.953	138.27	147.18	67	210
3	121	147.45	29.806	142.08	152.81	77	222
Income							
1	176	143.61	28.585	139.36	147.87	73	216
2	167	144.64	27.283	140.47	148.81	72	224
3	157	146.78	28.155	142.34	151.22	67	221
Education							
1	171	145.64	29.331	141.21	150.07	73	224
2	159	144.10	28.737	139.60	148.60	67	216
3	170	145.06	25.988	141.12	148.99	72	224

Table 4**.** Independent samples t-test

Covariate	t-value	P-value	Std error diff.
gender	-0.05	0.958	2.510
married	1.355	0.176	2.504
smokers	4.4	0.000	2.464
treatment	2.829	0.005	3.097

Table 5. ANOVA table

	Sum of squares	df	Mean Squares	F-value	P-value
Exercise					
Btw gps	8895.499	2	4447.750	5.784	0.003
Within gps	382179.349	497	768.973		
Over weight					
Btw gps	27800.853	2	13900.426	19.017	0.000
Within gps	363273.99	497	730.934		
Alcohol					
Btw gps	8137.837	2	4068.919	5.281	0.005
Within gps	3891074.848	497	770.497		

Table 6. Residual standard error, Multiple R-squared, Adjusted R-squared, F-statistic relavant to Full model

Residual standard error	25.32	482 degrees of freedom
Multiple R-squared	0.2099	
Adjusted R-squared	0.1821	
F-statistic	7.534	17 and 482 DF/ p-value: < 2.2e-16

Table 7. Estimated coefficients, Standard error of coefficient and P-values related to each model parameters and Residuals.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	36.47519	48.53904	0.751	0.4527	
smokeY	10.74273	2.34442	4.582	5.89E-06	***
as.factor(x$exercise)2	-11.47064	2.9038	-3.95	9.00E-05	***
as.factor(x$exercise)3	-10.72057	2.7085	-3.958	8.71E-05	***
as.factor(x$overwt)2	8.98298	4.336	2.072	0.0388	*
as.factor(x$overwt)3	11.92226	5.92162	2.013	0.0446	*
as.factor(x$alcohol)3	13.1264	2.89709	4.531	7.45E-06	***
as.factor(x$trt)1	-15.26782	2.92514	-5.22	2.69E-07	***
bmi	1.39409	0.83965	1.66	0.0975	.
as.factor(x$stress)3	5.04992	2.86234	1.764	0.0783	.

Table 8. Examining the model after excluding overweight and BMI.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	132.07207	14.44624	9.142	< 2e-16	***
x$smokeY	10.84096	2.34461	4.624	4.86E-06	***
as.factor(x$exercise)2	-11.6367	2.91634	-3.99	7.64E-05	***
as.factor(x$exercise)3	-11.11275	2.715	-4.093	5.00E-05	***
x$weight	0.19238	0.02949	6.525	1.75E-10	***
x$height	-0.4511	0.19605	-2.301	0.0218	*
as.factor(x$alcohol)3	12.79744	2.8997	4.413	1.26E-05	***
as.factor(x$trt)1	-14.51539	2.91419	-4.981	8.87E-07	***
as.factor(x$stress)3	5.11625	2.86976	1.783	0.0753	.

a.Residual standard error: 25.36 on 476 degrees of freedom
b.Multiple R-squared: 0.2174
c.Adjusted R-squared: 0.1796

Table 9. Examining the model after excluding weight and height.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	115.73893	8.77319	13.192	< 2e-16	***
x$smokeY	10.70607	2.35215	4.552	6.77E-06	***
as.factor(x$exercise)2	-11.00422	2.8964	-3.799	0.000164	***
as.factor(x$exercise)3	-10.59772	2.70724	-3.915	0.000104	***
as.factor(x$overwt)2	7.82334	3.95312	1.979	0.04839	*
as.factor(x$overwt)3	10.81698	5.78371	1.87	0.062064	.
as.factor(x$alcohol)3	12.71883	2.89673	4.391	1.39E-05	***
as.factor(x$trt)1	-14.6408	2.92148	-5.011	7.64E-07	***
as.factor(x$stress)3	5.3133	2.86534	1.854	0.064311	.

a.Residual standard error: 25.32 on 475 degrees of freedom
b.Multiple R-squared: 0.2212
c.Adjusted R-squared: 0.1819
d.F-statistic: 5.622 on 24 and 475 DF
e. p-value: 5.377e-15

Table 10. Presenting the model after omitting the influence data

	Estimate	Std.Error	tvalue	Pr(>|t|)
(Intercept)	-2.86279	45.16763	-0.063	0.9495
x.ogender	2.66347	4.96471	0.536	0.5919
x.omarried	3.97342	2.1838	1.819	0.0695 .
x.osmoke	10.41848	2.19607	4.744	2.78e-06 ***
x.oexercise	-5.04562	1.28275	-3.933	9.64e-05 ***
x.oage	0.20797	0.08262	2.517	0.0122 *
x.oweight	-0.10789	0.12926	-0.835	0.4043
x.oheight	1.05696	0.67268	1.571	0.1168
x.ooverwt	2.56992	2.78197	0.924	0.3561
x.orace	-1.33113	1.42046	-0.937	0.3492
x.oalcohol	7.02912	1.37168	5.124	4.37e-07 ***
x.otrt	-13.43795	2.71304	-4.953	1.02e-06 ***
x.obmi	1.49231	0.75195	1.985	0.0478 *
x.ostress	2.19081	1.34812	1.625	0.1048
x.osalt	0.8659	1.31817	0.657	0.5116
x.ochldbear	1.97304	3.01025	0.655	0.5125
x.oincome	3.02	1.34375	2.247	0.0251 *
x.oeducatn	0.16287	1.32807	0.123	0.9024

a.Residual standard error: 23.68 on 469 degrees of freedom

b.Multiple R-squared: 0.2312
c.Adjusted R-squared: 0.2033
d-statistic: 8.295 on 17 and 469 DF,
e.p-value: < 2.2e-16

Table 11. Results of Step wise method

```
Step:  AIC=3103.86
y.o ~ married + smoke + age + height + alcohol + trt + bmi +
    stress + income
Df Sum of Sq    RSS    AIC
<none>                273931 3103.9
+ weight    1     745.4 273186 3104.5
- stress    1    1526.0 275457 3104.6
+ race      1     572.2 273359 3104.8
+ overwt    1     205.3 273726 3105.5
+ salt      1     187.1 273744 3105.5
+ chldbear  1      96.1 273835 3105.7
- income    1    2185.1 276116 3105.7
+ educatn   1      15.9 273915 3105.8
+ gender    1       0.3 273931 3105.9
- married   1    2488.1 276419 3106.3
- height    1    2753.6 276685 3106.7
- age       1    3600.8 277532 3108.2
- smoke     1   11287.9 285219 3121.5
- trt       1   12391.7 286323 3123.4
- alcohol   1   14493.3 288424 3127.0
- bmi       1   25952.6 299884 3145.9
Call:
lm(formula = y.o ~ married + smoke + age + height + alcohol +
trt + bmi + stress + income)
Coefficients:
(Intercept)     married       smoke         age      height     alcohol
   32.0811      4.5707      9.7840      0.2074      0.4808      6.8483
trtbmi      stress      income
 -12.6591      1.0859      2.2106      2.6245
```

Table 12. Results of backward selection

```
Step:  AIC=3103.86
y.o ~ married + smoke + age + height + alcohol + trt + bmi +
    stress + income
```

```
Df Sum of Sq    RSS    AIC
<none>              273931 3103.9
- stress   1    1526.0 275457 3104.6
- income   1    2185.1 276116 3105.7
- married  1    2488.1 276419 3106.3
- height   1    2753.6 276685 3106.7
- age      1    3600.8 277532 3108.2
- smoke    1   11287.9 285219 3121.5
- trt      1   12391.7 286323 3123.4
- alcohol  1   14493.3 288424 3127.0
- bmi      1   25952.6 299884 3145.9

Call:
lm(formula = y.o ~ married + smoke + age + height + alcohol +
trt + bmi + stress + income)

Coefficients:
(Intercept)     married       smoke         age      height     alcohol
   32.0811      4.5707      9.7840      0.2074      0.4808      6.8483
trtbmi      stress      income
  -12.6591      1.0859      2.2106      2.6245
```

Table 13. Results of Forward selection

```
Call:

lm(formula = y.o ~ gender + married + smoke + age + weight +
   height + overwt + race + alcohol + trt + bmi + stress + salt +
chldbear + income + educatn)

Coefficients:
(Intercept)      gender     married       smoke         age      weight
  -22.5406      3.2538      4.4428      9.7063      0.1976     -0.1538
    height      overwt        race     alcohol         trtbmi
    1.2238      2.5992     -1.3626      7.0589     -13.0594      1.7357
    stress        salt    chldbear      income     educatn
    2.3838      0.7250      2.0074      2.5743      0.3469
```

Table 14. Investigating interaction terms in final model

	Estimate	Std. Error t	value	Pr(>|t|)	
(Intercept)	-107.86265	92.98237	-1.16	0.24675	
x.o$smokeY	68.51315	39.78944	1.722	0.08589	.
as.factor(x.o$exercise)3	-78.96833	45.59482	-1.732	0.08408	.
x.o$age	3.37218	1.44631	2.332	0.02024	*
x.o$height	2.89273	1.21531	2.38	0.01779	*
x.o$bmi	3.82922	1.69866	2.254	0.02474	*
as.factor(x.o$stress)2	-130.09594	48.96873	-2.657	0.00822	**
x.o$smokeY:as.factor(x.o$trt)1	-13.00561	6.22561	-2.089	0.03736	*
as.factor(x.o$exercise)3:as.factor(x.o$trt)1	11.98766	6.94278	1.727	0.08503	.
as.factor(x.o$exercise)3:x.o$bmi	1.06422	0.4224	2.519	0.01216	*
as.factor(x.o$exercise)2:as.factor(x.o$income)3	14.96712	7.39871	2.023	0.04377	*
x.o$age:x.o$height	-0.04259	0.01795	-2.372	0.01816	*
x.o$age:x.o$bmi	-0.03028	0.01412	-2.144	0.03263	*
x.o$height:as.factor(x.o$stress)2	1.40316	0.58241	2.409	0.01645	*
as.factor(x.o$alcohol)3:as.factor(x.o$trt)1	-16.38835	7.13271	-2.298	0.02212	*
as.factor(x.o$trt)1:x.o$bmi	-1.0292	0.53977	-1.907	0.0573	.
x.o$bmi:as.factor(x.o$stress)2	0.97707	0.45079	2.167	0.03081	*

Table 15. Final Model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	58.87889	17.7727	3.313	0.000994	***
Smoke/Y	10.40449	2.18799	4.755	2.64E-06	***
Exercise/2	-9.52305	2.70537	-3.52	0.000473	***
Exercise/3	-10.12955	2.53552	-3.995	7.50E-05	***
age	0.21189	0.08147	2.601	0.009592	**
height	0.55435	0.21693	2.556	0.010916	*
Alcohol/3	13.9706	2.6886	5.196	3.03E-07	***
Treatment/1	-13.99667	2.70463	-5.175	3.38E-07	***

BMI	1.148	0.16038	7.158	3.15E-12	***
Income/3	5.77526	2.663	2.169	0.030604	*

Table 16. The parameter estimates, based on Lasso and elastic net regularization method

Covariate	LASSO	Elastic Net	Covariate	LASSO	Elastic Net
(Intercept)	98.62171275	97.96381	g200	-10.16908663	-9.82274
married	1.213024982	1.181459	g204	0.144891605	0.196882
smoke	6.241632804	6.070306	g222	0.173642996	0.17812
exercise	-2.561736925	-2.49205	g231	1.538892709	1.563042
weight	0.016918253	0.018782	g232	0.133575595	0.215477
overwt	6.101133049	5.330882	g271	0.072463224	0.100612
alcohol	2.582135105	2.491974	g279	-1.001582702	-1.02494
trt	-12.17618601	-11.5696	g288	0.939017266	0.836392
bmi	0.127373838	0.183008	g289	0.487137338	0.538304
stress	0.363644817	0.356072	g292	0.38024346	0.412306
g7	1.830074608	1.804598	g295	0.980644923	0.985043
g9	0.47023874	0.480819	g298	2.788953567	2.690553
g10	0.993889852	0.949791	g309	1.739311938	1.658337
g46	0.458541448	0.432034	g330	0.116400913	0.157259
g48	-0.618286444	-0.55245	g337	-0.970127524	-0.96891
g50	13.81948967	13.34604	g348	-0.554414413	-0.532
g59	0.720558504	0.689798	g356	0.69214981	0.674215
g63	-0.139378944	-0.16717	g362	0.055020437	0.05502
g86	-0.564274298	-0.52688	g364	-0.606134278	-0.60296
g92	-0.093392118	-0.10547	g366	0.093138254	0.07798
g108	0.196704248	0.165531	g371	0.201716641	0.219844
g120	-0.772939444	-0.76165	g377	-1.262438881	-1.28139
g122	-0.394313304	-0.39507	g391	-1.792120379	-1.76657
g135	0.627334714	0.588014	g411	-1.094084759	-1.02033
g137	1.016918872	1.022803	g412	-0.145624365	-0.12107
g150	0.833063633	0.788	g422	1.195815859	1.149338
g160	0.367542042	0.358586	g425	1.742065317	1.704052
g168	0.558274417	0.623828	g438	-0.272354116	-0.2832
g169	0.84382564	0.756787	g443	-1.92798866	-1.83685
g175	0.01857471	0.038428	g447	-1.107143019	-1.09474
g179	-0.929086786	-0.91618	g453	-2.274494065	-2.1281
g182	0.241317519	0.265798	g465	-1.826344393	-1.73153
g187	0.890428795	0.942401	g469	0.000250151	0.07466
g191	0.076566627	0.143505	g480	0.161150871	0.221406

Table 17. Estimated Coefficients for group high blood pressure

(Intercept)	112.3523324	g214	-0.87086178
age	0.04931164	g216	-1.87885995
weight	0.03484354	g242	0.80648087
trt	9.10661401	g216	-0.1210693
bmi	0.02627079	g242	-0.07134717
g5	0.43670059	g253	-0.84012946
g7	1.38800582	g285	0.54412912
g9	0.24026096	g327	0.63872101
g47	0.25377508	g357	-0.41726317
g95	1.24884173	g359	-0.07159938
g111	-0.12810864	g377	-0.68956028
g112	0.69135076	g385	0.08027192
g132	-0.60349761	g393	-1.32600769
g169	0.48755412	g427	0.02368833
g177	0.05735365	g434	-0.68330925
g191	0.34273117	g464	1.14815752
g200	-0.81803034	g474	0.089646
g205	-0.12139622	g478	-1.87885995
g207	-0.44789587	g480	0.80648087

Table 18. Estimated Coefficients for group low blood pressure

(Intercept)	140.1897608	g200	-4.2757269
smoke	2.66003	g377	-0.6687182
overwt	2.9656668	g432	0.1640836
alcohol	0.2337104	g458	0.215961
g50	13.9084737	g459	-0.1779234

Table 19. Relevant coefficient based on binary response variable

(Intercept)	-0.590942497	g120	-0.053674063
married	0.255988487	g122	-0.035817758
exercise	-0.080107103	g137	0.063438963
overwt	0.03619205	g150	0.016843234
trt	-1.418685495	g168	0.015018091
bmi	0.028479371	g187	0.092408609
stress	0.080702519	g191	0.009219666
g10	0.066552532	g200	-0.575575943
g36	-0.014449676	g204	0.022184393
g49	0.003576402	g231	0.056584005
g50	0.195049129	g279	-0.018736388
g65	-0.093599266	g298	0.074179733
g75	-0.000880806	g309	0.055582681
g86	-0.111798266	g330	0.013998067
g98	-0.059927625	g385	0.199842492
g391	-0.079077044	g453	-0.12178576
g412	-0.000508177	g460	-0.030333724
g425	0.048351833	g469	0.347407238
g447	-0.017509594	g475	0.018117494
g450	-0.01223533		

Appendix B: Figures

Figure 1. Bar chart of blood pressure

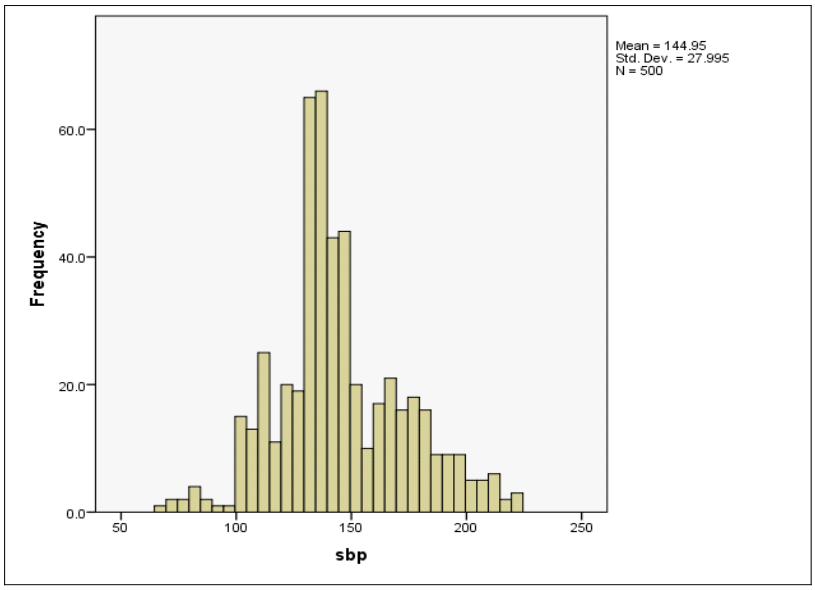


Figure 2. Correlation matrix plot for the continuous variables

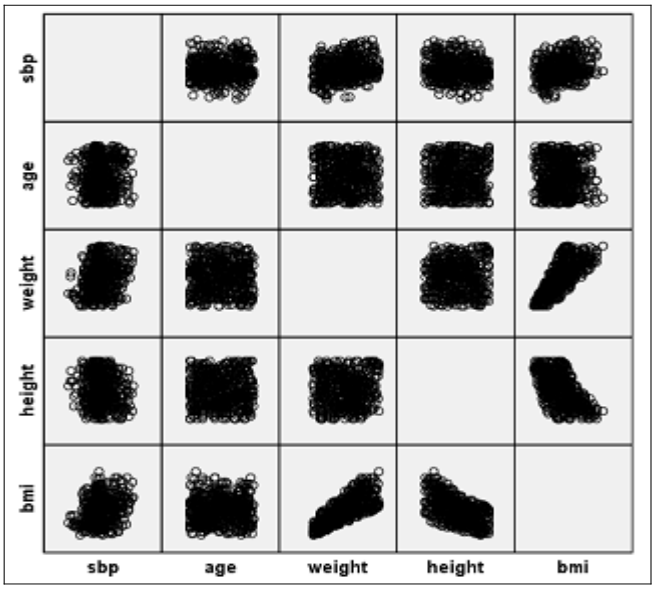


Figure 3. Box plots for systolic blood pressure by gender, married smoke treatment exercise and over weight

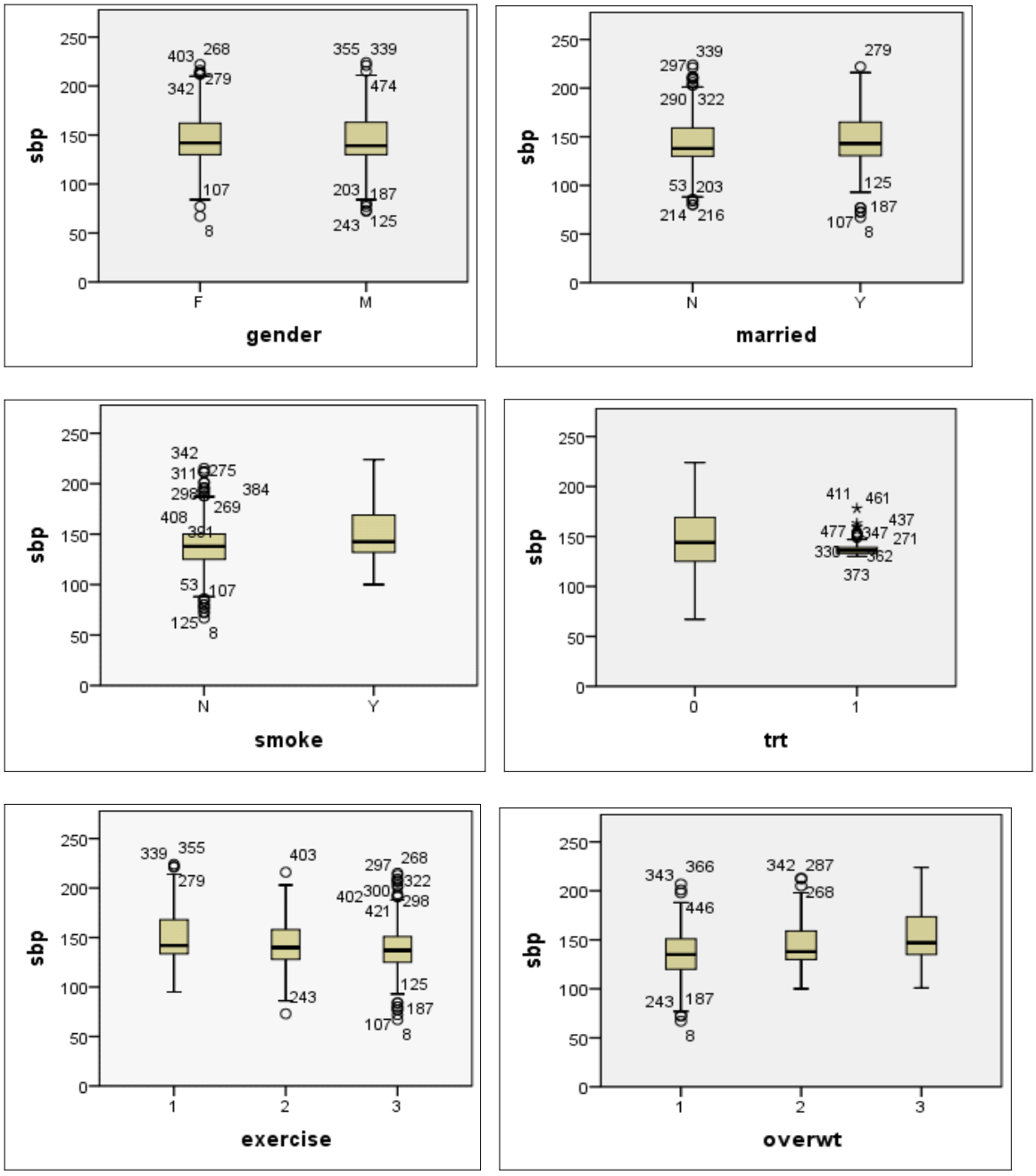


Figure 4. Box plots for systolic blood pressure by, income alcohol education stress salt child bearing and race

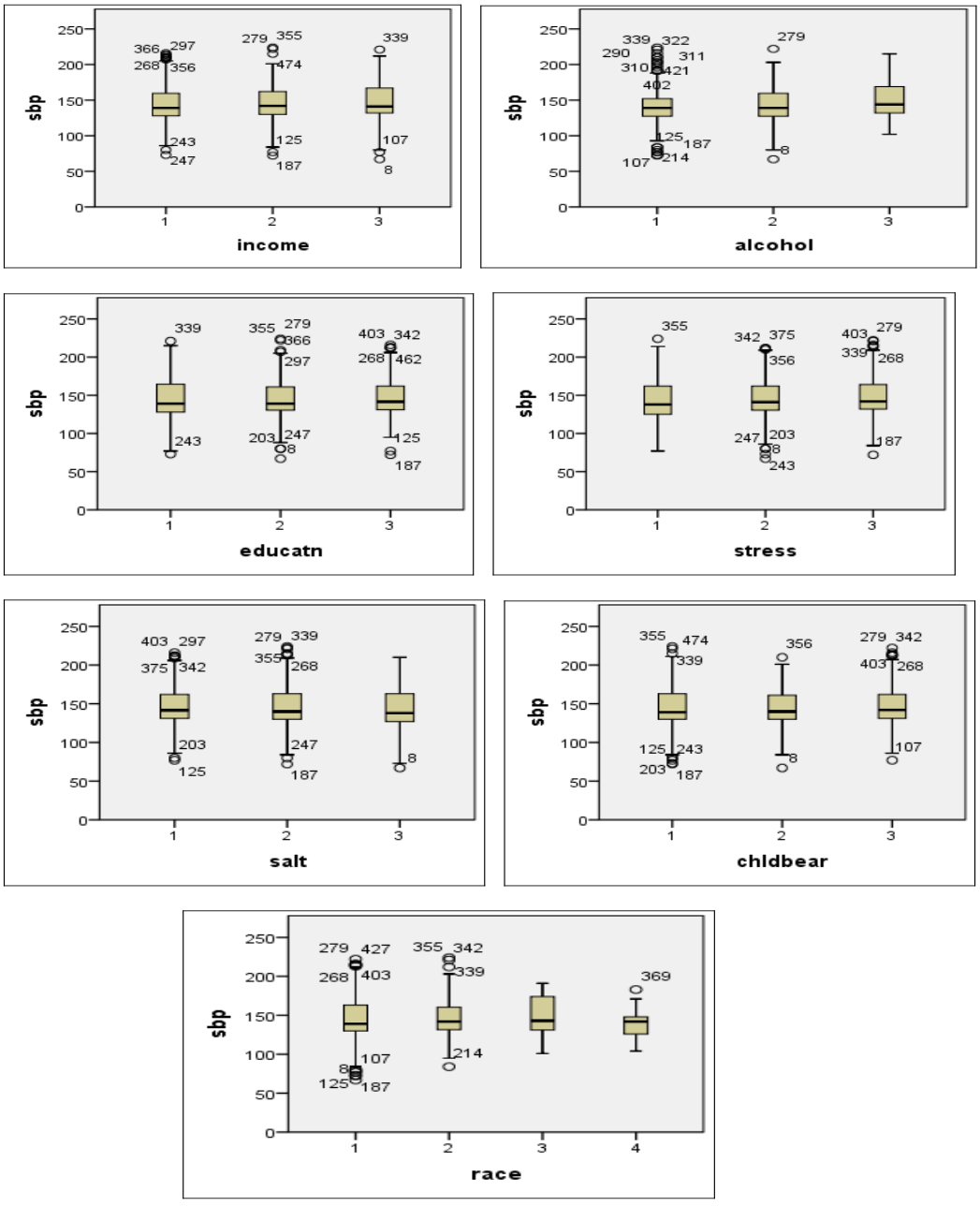
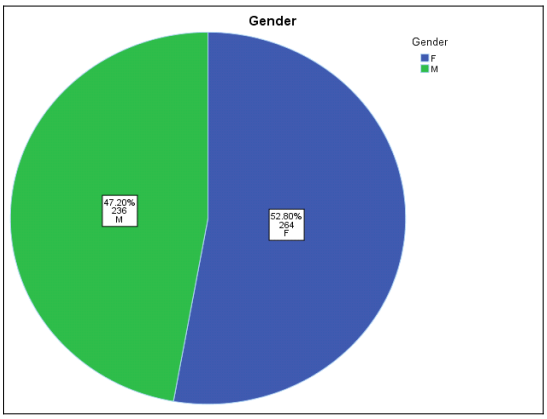
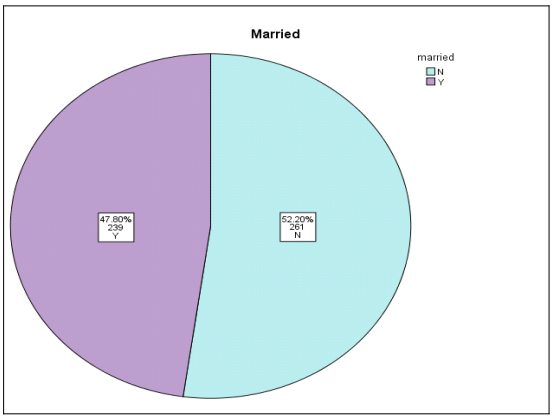
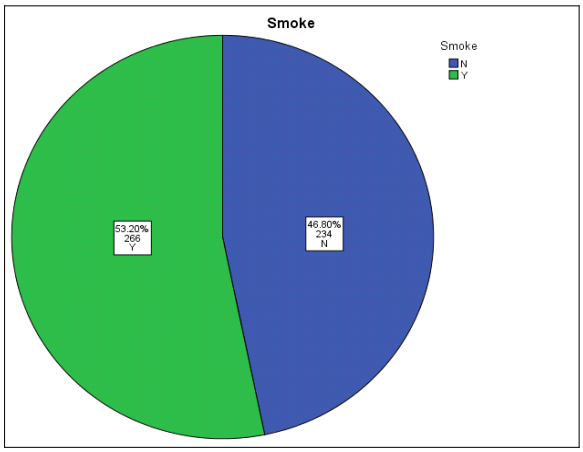
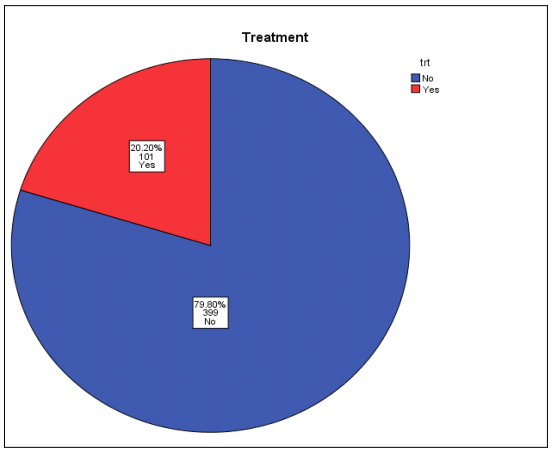


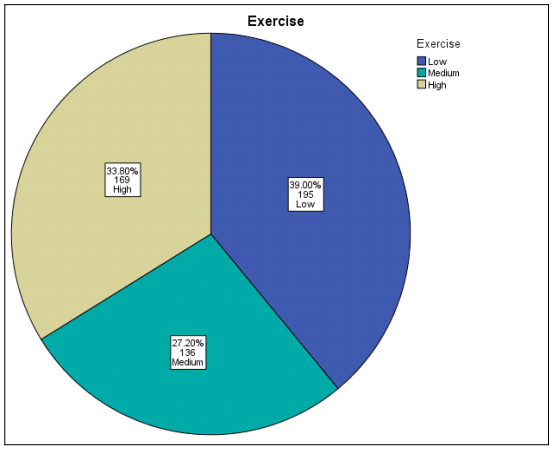
Figure 5. Pie charts for categorical variables

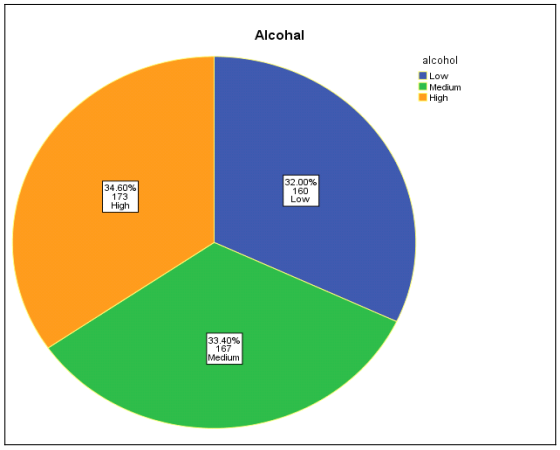


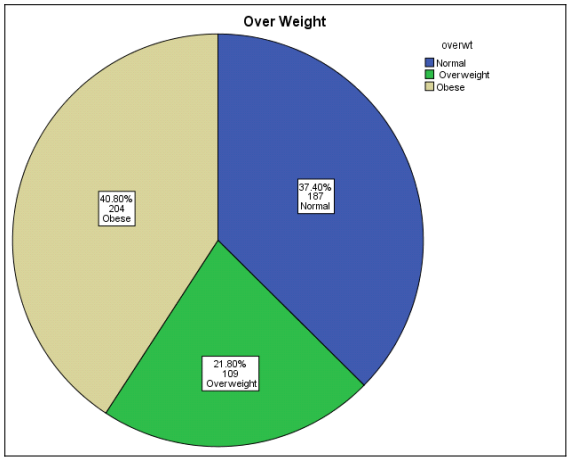


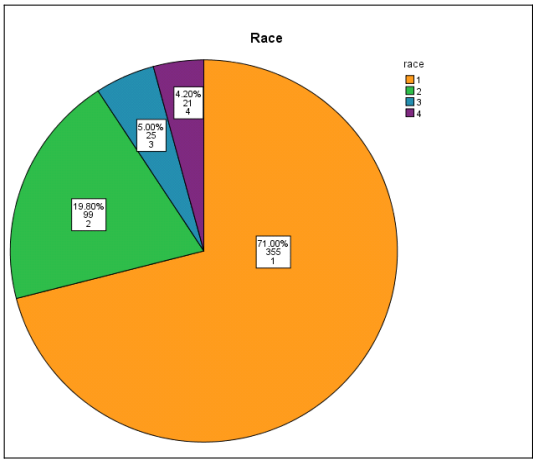


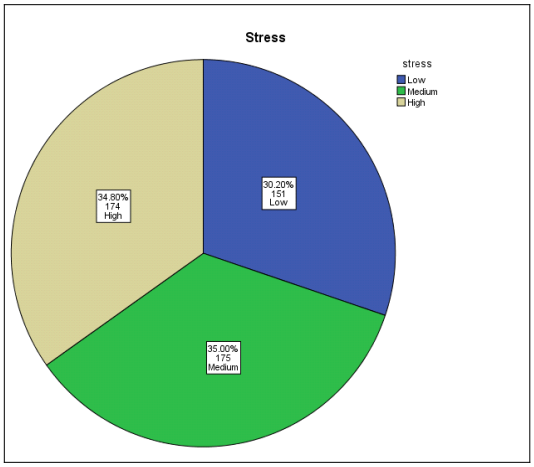







Over Weight
overwt
Normal
Overweight
Obese
37.40%
187
Normal
21.80%
109
Overweight
40.80%
204
Obese


Race
race
1
2
3
4
4.20%
21
4
5.00%
25
3
19.80%
99
2
71.00%
355
1


Stress
stress
Low
Medium
High
30.20%
151
Low
35.00%
175
Medium
34.80%
174
High

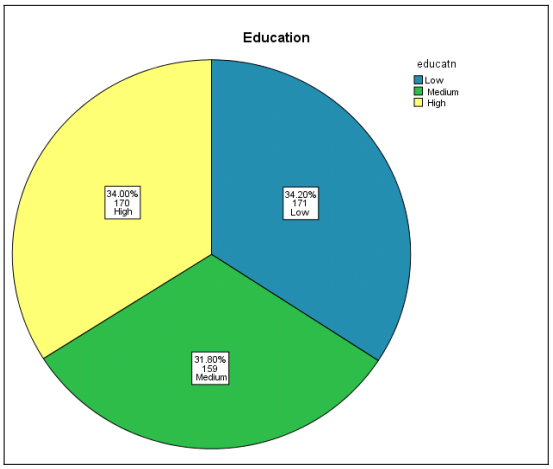

Education
educatn
Low
Medium
High
34.20%
171
Low
31.80%
159
Medium
34.00%
170
High

Figure 6. Checking the goodness of fit for the full model

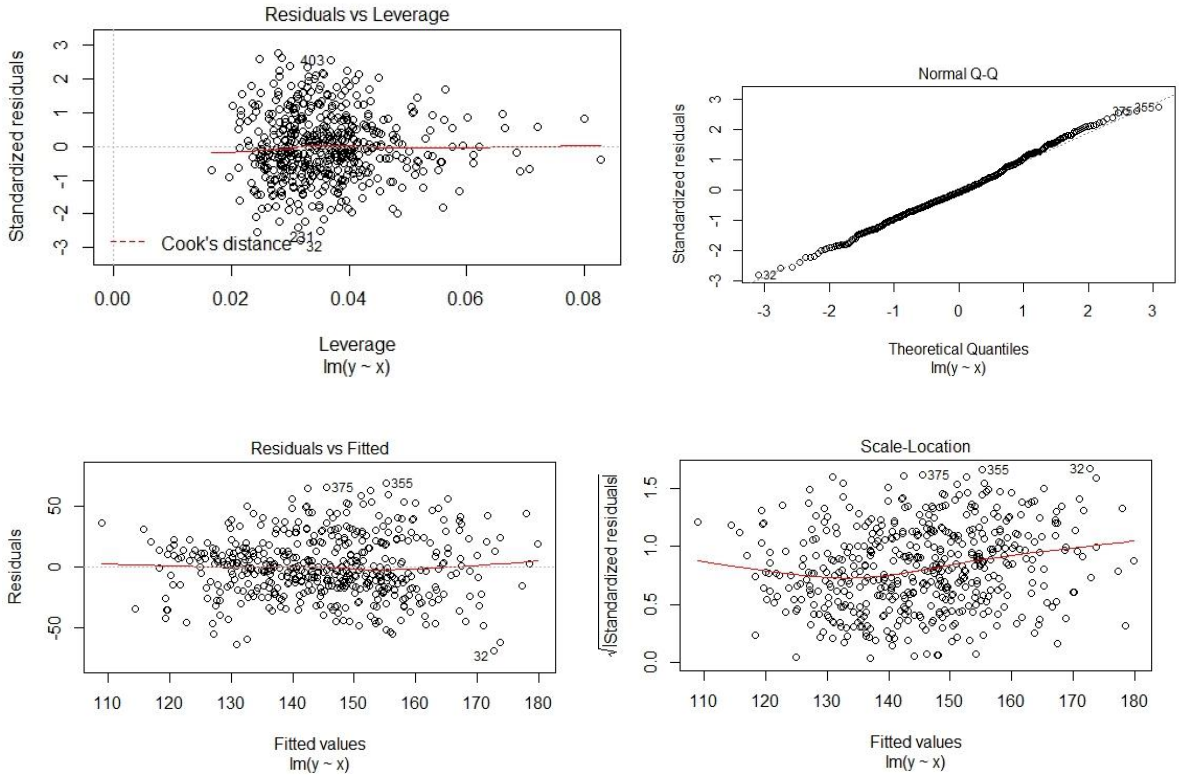
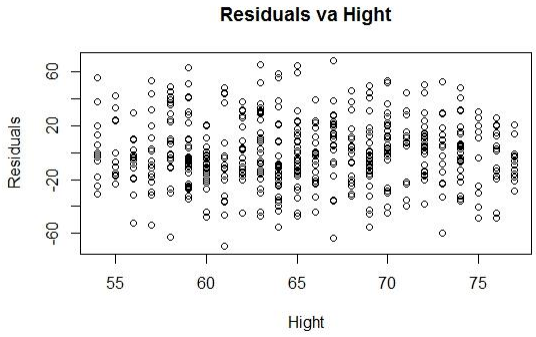
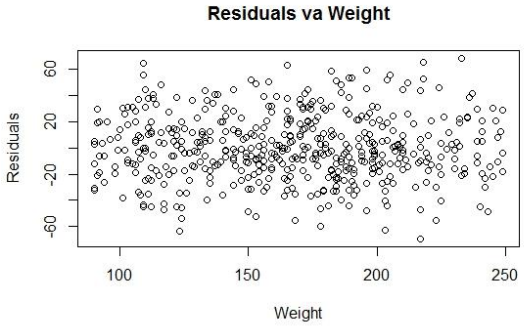


Figure 7. Investigating the plots of variables vs response for finding specific trend





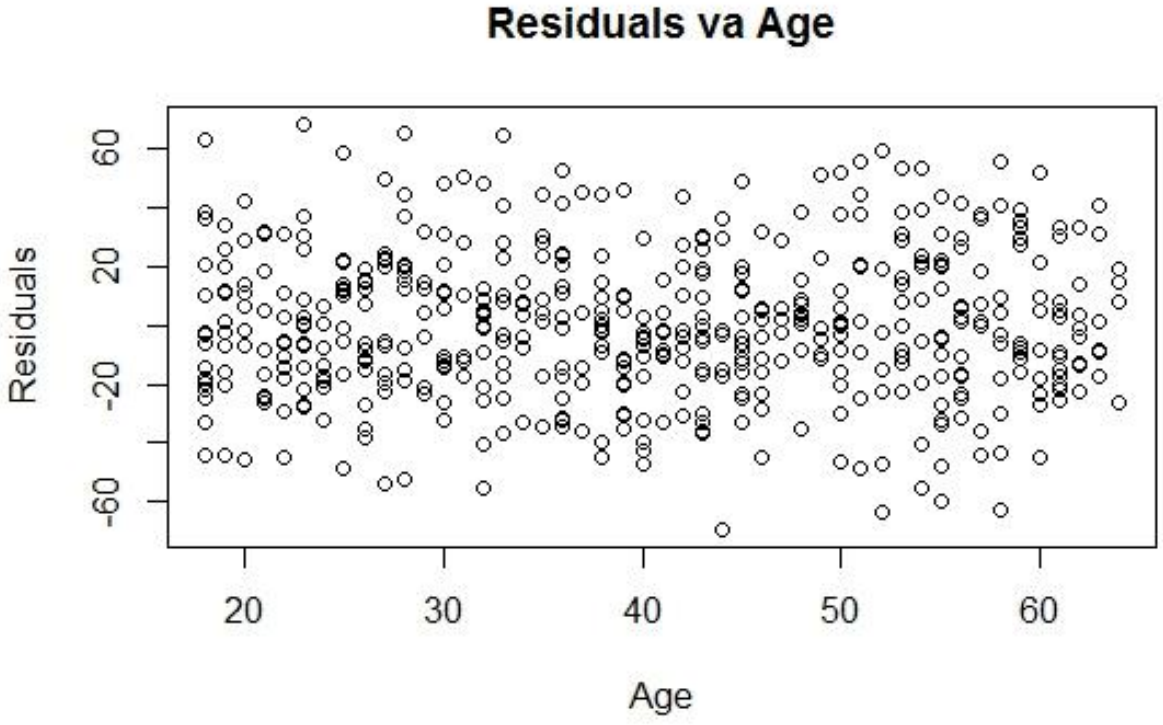


Figure 8. Plot relevant to Cp method

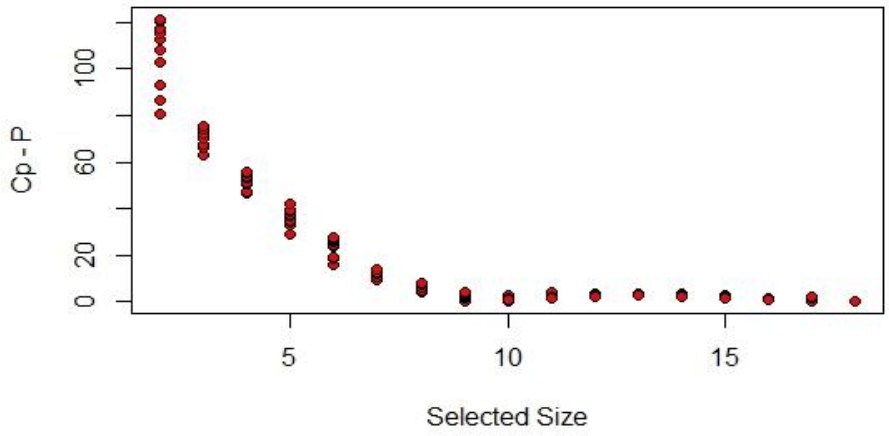


Figure 9. Goodness of fit for final model

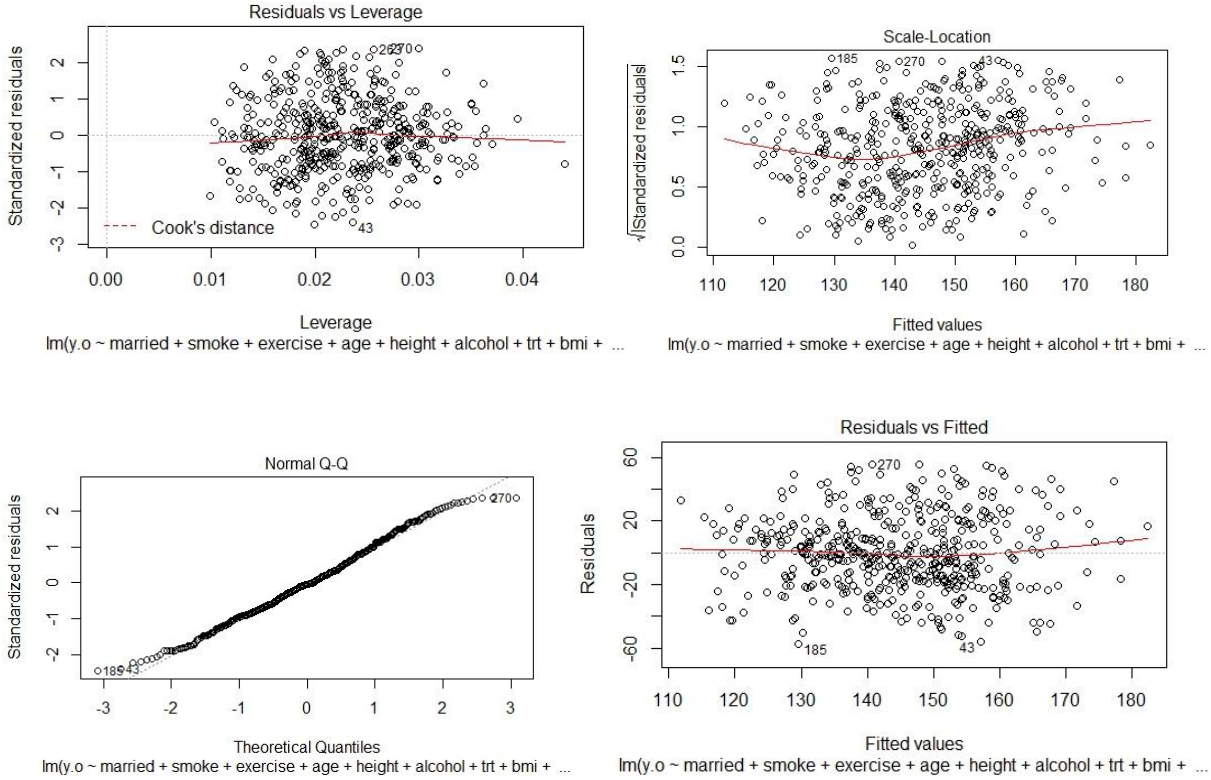
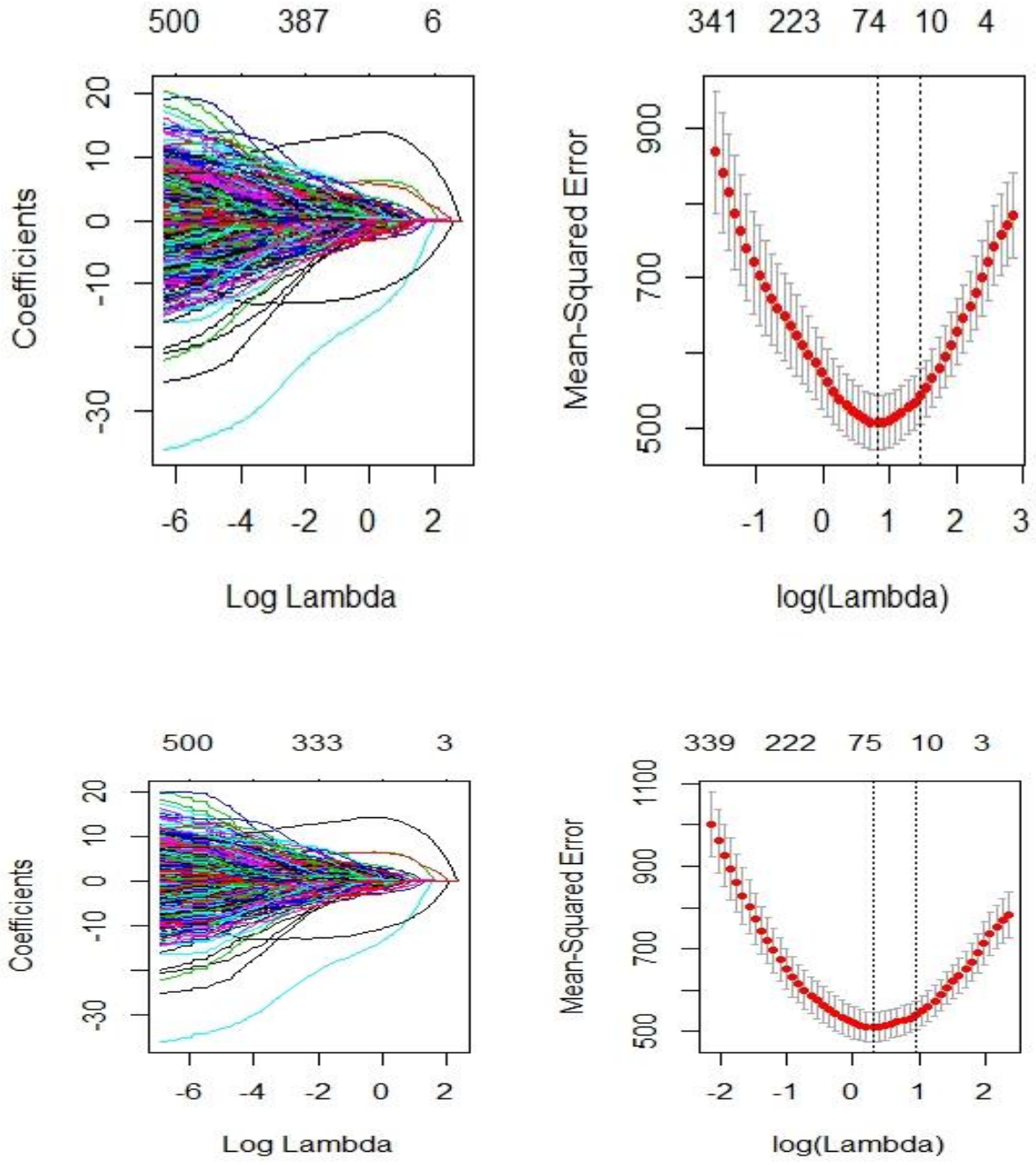


Figure 10. Coefficients values against the log-lambda value /Cross validation carve across lambda (First LASSO, Second Elastic Net)



Appendix C: R codes

```
#############data set#########
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="	")
blood=data.frame(blood)
library(glmnet)
attach(blood)
pairs(blood)
blood[1,]

###############descriptive analysis##############
par(mfrow=c(2,2))
boxplot(sbp~gender,xlab="Gender",ylab="Systolic Blood Pressure")
boxplot(sbp~married,xlab="Married",ylab="Systolic Blood Pressure")
boxplot(sbp~smoke,xlab="smoke",ylab="Systolic Blood Pressure")
boxplot(sbp~exercise,xlab="Exercise",ylab="Systolic Blood Pressure")
boxplot(sbp~overwt,xlab="overwt",ylab="Systolic Blood Pressure")
boxplot(sbp~race,xlab="race",ylab="Systolic Blood Pressure")
boxplot(sbp~alcohol,xlab="alcohol",ylab="Systolic Blood Pressure")
boxplot(sbp~trt,xlab="treatment",ylab="Systolic Blood Pressure")
boxplot(sbp~salt,xlab="Salt (NaCl) Intake Level",ylab="Systolic Blood Pressure")
boxplot(sbp~chldbear,xlab="Childbearing Potential",ylab="Systolic Blood Pressure")
boxplot(sbp~income,xlab="Income Level",ylab="Systolic Blood Pressure")
boxplot(sbp~educatn,xlab="Education Level",ylab="Systolic Blood Pressure")
plot(bmi,sbp)
plot(age,sbp)
plot(weight,sbp)
plot(height,sbp)
t.test(sbp~smoke)
summary(aov(sbp~exercise))
summary(aov(sbp~overwt))
summary(aov(sbp~race))
summary(aov(sbp~alcohol))
t.test(sbp~trt)


###############for only clinical data#########
library(MASS)
x=blood[,c(2:18)]

y=blood[,1]
y=as.numeric(y)
fit.lin=lm(y~x)
fit.lin=lm(y~x$gender+x$married+x$smoke+as.factor(x$exercise)+x$age+x$weight+x$height+
          as.factor(x$overwt)+as.factor(x$race)+as.factor(x$alcohol)+as.factor(x$trt)
         +x$bmi+as.factor(x$stress)+as.factor(x$salt)+
          as.factor(x$chldbear)+as.factor(x$income)+as.factor(x$educatn))
fit.lin=lm(y~x$gender+x$married+x$smoke+x$exercise+log(x$age)+log(x$weight)
        +x$height+
log(x$overwt)+x$race+x$alcohol+x$trt+x$bmi+x$bmi+x$stress+x$salt+
x$chldbear+x$income+x$educatn)
summary(fit.lin)
########exclud#########
xe=x[,-c(12,8)]
fit.lin=lm(y~xe)
summary(fit.lin)
library(xtable)
a=xtable(inf)
```

```
print(a,type= "html")
#fit.lin <- xtable(fit.lin)
#print(fit.lin,print.results = FALSE)
plot(y)
hist(y, xlab = "Systolic blood presure")
summary(y)
a=cor(x)
kappa(fit.lin)
########Chech the model######
library(car)
library(perturb)
plot(fit.lin)
inf=summary(influence.measures(fit.lin))
crPlots(fit.lin,terms = ~.)
studres(fit.lin)
fitted=fit.lin$fitted.values
plot(fitted,studres(fit.lin),xlab="Fitted values",ylab="Studentized Residual")
plot(x[,5],y)
plot(x[,6],y)
plot(x[,7],y)
str(fit.lin)
plot(x[,5],fit.lin$ residuals, main = "Residuals va Age", xlab = "Age",ylab = "Residuals")
plot(x[,6],fit.lin$ residuals,main = "Residuals va Weight", xlab = "Weight",ylab = "Residuals")
plot(x[,7],fit.lin$ residuals,main = "Residuals va Hight", xlab = "Hight",ylab = "Residuals")
###########omite the influence#######
x.o=blood[-c(8,32,204,231,243,339,355,356,366,375,403,474,485),c(2:18)]
y.o=blood[-c(8,32,204,231,243,339,355,356,366,375,403,474,485),1]
x.o=data.matrix(x.o)
y.o=as.numeric(y.o)
fit.lin.o=lm(y.o~x.o)
summary(fit.lin.o)
omit=xtable(fit.lin.o)
print(omit,type= "html")
plot(fit.lin.o)
summary(influence.measures(fit.lin.o))
crPlots(fit.lin.o,terms = ~.)
plot(x.o[,5],y.o)
plot(x.o[,6],y.o)
plot(x.o[,7],y.o)
str(fit.lin.o)
plot(x.o[,5],fit.lin.o$ residuals, main = "Residuals va Age", xlab = "Age",ylab = "Residuals")
plot(x.o[,6],fit.lin.o$ residuals,main = "Residuals va Weight", xlab = "Weight",ylab = "Residuals")
plot(x.o[,7],fit.lin.o$ residuals,main = "Residuals va Hight", xlab = "Hight",ylab = "Residuals")
##############variable selection###
data.v=data.frame(y.o,x.o)
x.o[1,]
attach(data.v)

a=step(lm(y.o~gender+married+smoke+exercise+age+weight+height+overwt+race+alcohol+
     trt+bmi+stress+salt+chldbear+income+educatn),direction = "both",data=data.v)
library(xtable)
a=xtable(a)

print(a,type= "html")
step(lm(y.o~gender+married+smoke+age+weight+height+overwt+race+alcohol+
       trt+bmi+stress+salt+chldbear+income+educatn),direction = "backward" ,data=data.v)

step(lm(y.o~gender+married+smoke+age+weight+height+overwt+race+alcohol+
       trt+bmi+stress+salt+chldbear+income+educatn),direction = "forward" ,data=data.v)
#######model based on stepwise##########
fit.step=lm(y.o~(x.o$married+x.o$smoke+as.factor(x.o$exercise)+x.o$age+x.o$height
```

```
      +as.factor(x.o$alcohol)+as.factor(x.o$trt)
        +x.o$bmi+as.factor(x.o$stress)+
as.factor(x.o$income))^2)
fit.step=lm(y.o~married+smoke+exercise+age+height+alcohol+
trt+bmi+stress+income)
extractAIC(fit.step)
newy=predict(fit.step,data.v$x.o,data=data.v)
mse = sum((newy - y.o) ^ 2) / length(y.o)
mse
summary(fit.step)
23*23

summary(fit.step)
plot(fit.step)
summary(influence.measures(fit.step))
crPlots(fit.lin.o,terms = ~.)
detach(data.v)
############subset selection#########
library(leaps)

ModelSel=leaps(x.o,y.o,method='Cp')
plot(ModelSel$size,abs(ModelSel$Cp-ModelSel$size),pch = 21, bg='red', xlab = "Selected Size", ylab="Cp - P")
aa = min(abs(ModelSel$Cp-ModelSel$size)[ModelSel$size==3])
bb = which(ModelSel$Cp == aa+3)
bb = which(ModelSel$Cp == 3-aa)
ModelSel$which[bb,]

ModelSel=leaps(x.o,y.o,method='adjr2')
aa = max(ModelSel$adjr2)
bb = which(ModelSel$adjr2 == aa)
ModelSel$which[bb,]

########################### LASSO#################
library(glmnet)

fit.lasoo = glmnet(x.o,y.o)
par(mfrow=c(1,2))
plot(fit.lasoo,label=TRUE)
plot(fit.lasoo,xvar="lambda",label=TRUE)

## slection of \lambda
fit.cv = cv.glmnet(x.o,y.o,lambda=fit.lasoo$lambda.min)
plot(fit.cv)
predict(fit.cv,newx=x[1:3,])
fit.cv$lambda.min
coef(fit.cv,s="lambda.min")
predict(fit.cv,newx=x[1:3,],s=c(0.001,0.002))
plot.cv.glmnet(fit.cv)
############################################################
##############################All covariates#########################
#############data set#########
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="           ")
blood=data.frame(blood)
library(glmnet)
x=blood[,c(2:501)]
y=blood[,1]
x=data.matrix(x)
y=as.numeric(y)
attach(blood)
fit.res.lasoo=lm(y~gender+married+smoke+exercise+age+weight+height+overwt+race+alcohol+
```

```
trt+bmi+stress+salt+chldbear+income+educatn+g7+g50+g137+g169+g179+g200+g298+g364
    +g391+g438+g447+g453+g465+g466)
summary(fit.res.lasoo)
###############all into regularization############
fit.gl=glmnet(x,y)
summary(fit.gl)
str(fit.gl)
plot(fit.gl, xvar = "lambda", label = TRUE)
cvfit=cv.glmnet(x, y)
plot(cvfit)
cvfit$lambda.min
newy=predict(cvfit, x, s = "lambda.min")
plot(newy)
points(y,col="red",pch=8)
coef(cvfit,s="lambda.min")
mse = sum((newy - y) ^ 2) / length(y)
mse
###############only genes into regularization############
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x,y, penalty.factor = p.fac)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x, y,penalty.factor = p.fac)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x, s = "lambda.min",penalty.factor = p.fac)
plot(newy)
points(y,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y) ^ 2) / length(y)
mse
#######################Elastic net alpha=0.6#################################
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x,y, penalty.factor = p.fac,alpha = 0.6)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x, y,penalty.factor = p.fac,alpha = 0.6)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x, s = "lambda.min",penalty.factor = p.fac)
plot(newy)
points(y,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y) ^ 2) / length(y)
mse

#########################Elastic without penalty#############################################
par(mfrow=c(1,2))

fit.gl.pen=glmnet(x,y,alpha = 0.6)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x, y,alpha = 0.6)
plot(cvfit.pen)
```

```
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x, s = "lambda.min")
plot(newy)
points(y,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
##############################################
##################logistic####################
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="	")
blood=data.frame(blood)
library(glmnet)
x=blood[,c(2:501)]
y=c(rep(0,250),rep(1,250))
x=data.matrix(x)
attach(blood)

###############all into regularization############min mse#########
par(mfrow=c(1,2))
fit.gl=glmnet(x,y, family = "binomial")
summary(fit.gl)
str(fit.gl)
plot(fit.gl, xvar = "lambda", label = TRUE)
cvfit=cv.glmnet(x, y,family = "binomial")
plot(cvfit)
cvfit$lambda.min
newy=predict(cvfit, x, s = "lambda.min",family = "binomial")
plot(newy)
points(y,col="red",pch=8)
coef(cvfit,s="lambda.min")
mse = sum((newy - y) ^ 2) / length(y)
mse
exp(0.66)
###############only genes into regularization############
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x,y, penalty.factor = p.fac,family = "binomial")
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x, y,penalty.factor = p.fac,family = "binomial")
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x, s = "lambda.min",penalty.factor = p.fac,family = "binomial")
plot(newy)
points(y,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y) ^ 2) / length(y)
mse
#######################Elastic net alpha=0.6###################################
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x,y, penalty.factor = p.fac,alpha = 0.6,family = "binomial")
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x, y,penalty.factor = p.fac,alpha = 0.6,family = "binomial")
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x, s = "lambda.min",penalty.factor = p.fac,family = "binomial")
```

```
plot(newy)
points(y,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y) ^ 2) / length(y)
mse

##########################Elastic without penalty#############################################
par(mfrow=c(1,2))

fit.gl.pen=glmnet(x,y,alpha = 0.6,family = "binomial")
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x, y,alpha = 0.6,family = "binomial")
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x, s = "lambda.min",family = "binomial")
plot(newy)
points(y,col="red",pch=8)
coef(cvfit.pen,s="lambda.min",family = "binomial")
mse = sum((newy - y) ^ 2) / length(y)
mse

###############################################
fit.res.lasoo=lm(y~gender+married+smoke+exercise+age+weight+height+overwt+race+alcohol+
trt+bmi+stress+salt+chldbear+income+educatn+g7+g50+g137+g169+g179+g200+g298+g364
            +g391+g438+g447+g453+g465+g466)
summary(fit.res.lasoo)
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="           ")
blood=data.frame(blood)
library(glmnet)
library(sm)

blood.low=blood[1:250,]
x1=blood.low[,c(2:501)]
y1=blood.low[,1]
x1=data.matrix(x1)
y1=as.numeric(y1)
attach(blood.low)
###############################################
par(mfrow=c(2,2))
boxplot(sbp~gender,xlab="Gender",ylab="Systolic Blood Pressure")
boxplot(sbp~married,xlab="Married",ylab="Systolic Blood Pressure")
boxplot(sbp~smoke,xlab="smoke",ylab="Systolic Blood Pressure")
boxplot(sbp~exercise,xlab="Exercise",ylab="Systolic Blood Pressure")
boxplot(sbp~overwt,xlab="overwt",ylab="Systolic Blood Pressure")
boxplot(sbp~race,xlab="race",ylab="Systolic Blood Pressure")
boxplot(sbp~alcohol,xlab="alcohol",ylab="Systolic Blood Pressure")
boxplot(sbp~trt,xlab="treatment",ylab="Systolic Blood Pressure")
boxplot(sbp~salt,xlab="Salt (NaCl) Intake Level",ylab="Systolic Blood Pressure")
boxplot(sbp~chldbear,xlab="Childbearing Potential",ylab="Systolic Blood Pressure")
boxplot(sbp~income,xlab="Income Level",ylab="Systolic Blood Pressure")
boxplot(sbp~educatn,xlab="Education Level",ylab="Systolic Blood Pressure")
plot(bmi,sbp)
plot(age,sbp)
plot(weight,sbp)
plot(height,sbp)
t.test(sbp~smoke)
summary(aov(sbp~exercise))
summary(aov(sbp~overwt))
```

```
summary(aov(sbp~race))
summary(aov(sbp~alcohol))
t.test(sbp~trt)

##################################################
library(MASS)
x1=blood.low[,c(2:18)]
y1=blood.low[,1]
x1=data.matrix(x1)
y1=as.numeric(y1)
fit.lin=lm(y1~x1)
summary(fit.lin)
#fit.lin <- xtable(fit.lin)
#print(fit.lin,print.results = FALSE)
plot(y1)
hist(y1)
summary(y1)



########Chech the model######
library(car)
plot(fit.lin)
summary(influence.measures(fit.lin))
crPlots(fit.lin,terms = ~.)
studres(fit.lin)
fitted=fit.lin$fitted.values
plot(fitted,studres(fit.lin),xlab="Fitted values",ylab="Studentized Residual")
str(fit.lin)
plot(x1[,5],fit.lin$ residuals, main = "Residuals va Age", xlab = "Age",ylab = "Residuals")
plot(x1[,6],fit.lin$ residuals,main = "Residuals va Weight", xlab = "Weight",ylab = "Residuals")
plot(x1[,7],fit.lin$ residuals,main = "Residuals va Hight", xlab = "Hight",ylab = "Residuals")
###########omite the influence#######
x.o=blood.low[-c(4,8,53,107,125,165,187,203,241,243,247),c(2:18)]
y.o=blood.low[-c(4,8,53,107,125,165,187,203,241,243,247),1]
x.o=data.matrix(x.o)
y.o=as.numeric(y.o)
fit.lin.o=lm(y.o~x.o)
summary(fit.lin.o)
plot(fit.lin.o)
crPlots(fit.lin.o,terms = ~.)
studres(fit.lin.o)
fitted=fit.lin.o$fitted.values
plot(fitted,studres(fit.lin.o),xlab="Fitted values",ylab="Studentized Residual")


###################transformation####################
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="           ")
blood=data.frame(blood)
library(glmnet)
library(sm)
blood.low=blood[1:250,]
x1=blood.low[-c(4,8,53,107,125,165,187,203,241,243,247),c(2:18)]
y1=blood.low[-c(4,8,53,107,125,165,187,203,241,243,247),1]

x1=data.matrix(x1)
y1=as.numeric(y1)
summary(y1)

y1=141-y1
```

```
y1=log(y1)

fit.lin=lm(y1~x1)
summary(fit.lin)
##############Chech the model######
library(car)
plot(fit.lin)
crPlots(fit.lin,terms = ~.)
studres(fit.lin)
fitted=fit.lin$fitted.values
plot(fitted,studres(fit.lin),xlab="Fitted values",ylab="Studentized Residual")
plot(x1[,5],fit.lin$ residuals, main = "Residuals va Age", xlab = "Age",ylab = "Residuals")
plot(x1[,6],fit.lin$ residuals,main = "Residuals va Weight", xlab = "Weight",ylab = "Residuals")
plot(x1[,7],fit.lin$ residuals,main = "Residuals va Hight", xlab = "Hight",ylab = "Residuals")
##############################Model selection#############
data.v=data.frame(y1,x1)
attach(data.v)
step(lm(y1~gender+married+smoke+age+weight+height+overwt+race+alcohol+
        trt+bmi+stress+salt+chldbear+income+educatn),direction = "both",data=data.v)

#############################################################################
###############all into regularization############
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="           ")
blood=data.frame(blood)
library(glmnet)
#library(sm)
#library(hydroGOF)
#blood.low=blood[1:250,]
#x1=blood.low[-c(4,8,53,107,125,165,187,203,241,243,247),c(2:501)]
#y1=blood.low[-c(4,8,53,107,125,165,187,203,241,243,247),1]
#x1=data.matrix(x1)
#y1=as.numeric(y1)

blood.low=blood[1:250,]
x1=blood.low[,c(2:501)]
y1=blood.low[,1]
x1=data.matrix(x1)
y1=as.numeric(y1)
attach(blood.low)
#######################Elastic net alpha=0.6#################################
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x1,y1, penalty.factor = p.fac,alpha = 0.6)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x1, y1,penalty.factor = p.fac,alpha = 0.6)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x1, s = "lambda.min",penalty.factor = p.fac)
plot(newy)
points(y1,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y1) ^ 2) / length(y1)
mse

##########################Elastic without penalty#############################################
par(mfrow=c(1,2))
```

```
fit.gl.pen=glmnet(x1,y1,alpha = 0.6)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x1, y1,alpha = 0.6)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x1, s = "lambda.min")
plot(newy)
points(y1,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y1) ^ 2) / length(y1)
mse
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="	")
blood=data.frame(blood)
blood.high=blood[251:500,]
library(glmnet)
attach(blood.high)
#############data sicription##############
par(mfrow=c(2,2))
boxplot(sbp~gender,xlab="Gender",ylab="Systolic Blood Pressure")
boxplot(sbp~married,xlab="Married",ylab="Systolic Blood Pressure")
boxplot(sbp~smoke,xlab="smoke",ylab="Systolic Blood Pressure")
boxplot(sbp~exercise,xlab="Exercise",ylab="Systolic Blood Pressure")
boxplot(sbp~overwt,xlab="overwt",ylab="Systolic Blood Pressure")
boxplot(sbp~race,xlab="race",ylab="Systolic Blood Pressure")
boxplot(sbp~alcohol,xlab="alcohol",ylab="Systolic Blood Pressure")
boxplot(sbp~trt,xlab="treatment",ylab="Systolic Blood Pressure")
boxplot(sbp~salt,xlab="Salt (NaCl) Intake Level",ylab="Systolic Blood Pressure")
boxplot(sbp~chldbear,xlab="Childbearing Potential",ylab="Systolic Blood Pressure")
boxplot(sbp~income,xlab="Income Level",ylab="Systolic Blood Pressure")
boxplot(sbp~educatn,xlab="Education Level",ylab="Systolic Blood Pressure")
plot(bmi,sbp)
plot(age,sbp)
plot(weight,sbp)
plot(height,sbp)
t.test(sbp~smoke)
summary(aov(sbp~exercise))
summary(aov(sbp~overwt))
summary(aov(sbp~race))
summary(aov(sbp~alcohol))
t.test(sbp~trt)

#####################################################
library(MASS)
x1=blood.high[,c(2:18)]
y1=blood.high[,1]
x1=data.matrix(x1)
y1=as.numeric(y1)
fit.lin=lm(y1~x1)
summary(fit.lin)
#fit.lin <- xtable(fit.lin)
#print(fit.lin,print.results = FALSE)
plot(y1)
hist(y1)
summary(y1)



########Chech the model######
```

```
library(car)
plot(fit.lin)
summary(influence.measures(fit.lin))
crPlots(fit.lin,terms = ~.)
studres(fit.lin)
fitted=fit.lin$fitted.values
plot(fitted,studres(fit.lin),xlab="Fitted values",ylab="Studentized Residual")
str(fit.lin)
plot(x1[,5],fit.lin$ residuals, main = "Residuals va Age", xlab = "Age",ylab = "Residuals")
plot(x1[,6],fit.lin$ residuals,main = "Residuals va Weight", xlab = "Weight",ylab = "Residuals")
plot(x1[,7],fit.lin$ residuals,main = "Residuals va Hight", xlab = "Hight",ylab = "Residuals")
##########omite the influence#######
x.o=blood.high[-c(29,105,116,125,153,180,211),c(2:18)]
y.o=blood.high[-c(29,105,116,125,153,180,211),1]
x.o=data.matrix(x.o)
y.o=as.numeric(y.o)
fit.lin.o=lm(y.o~x.o)
summary(fit.lin.o)

crPlots(fit.lin.o,terms = ~.)
studres(fit.lin.o)
fitted=fit.lin.o$fitted.values
plot(fitted,studres(fit.lin.o),xlab="Fitted values",ylab="Studentized Residual")


###################transformation####################
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="           ")
blood=data.frame(blood)
library(glmnet)
library(sm)
blood.high=blood[251:500,]
x1=blood.high[-c(29,105,116,125,153,180,211),c(2:18)]
y1=blood.high[-c(29,105,116,125,153,180,211),1]

x1=data.matrix(x1)
y1=as.numeric(y1)
summary(y1)

y1=log(y1)

fit.lin=lm(y1~x1)
summary(fit.lin)
##############Chech the model######
library(car)
plot(fit.lin)
crPlots(fit.lin,terms = ~.)
studres(fit.lin)
fitted=fit.lin$fitted.values
plot(fitted,studres(fit.lin),xlab="Fitted values",ylab="Studentized Residual")
plot(x1[,5],fit.lin$ residuals, main = "Residuals va Age", xlab = "Age",ylab = "Residuals")
plot(x1[,6],fit.lin$ residuals,main = "Residuals va Weight", xlab = "Weight",ylab = "Residuals")
plot(x1[,7],fit.lin$ residuals,main = "Residuals va Hight", xlab = "Hight",ylab = "Residuals")
##############################Model selection#############
data.v=data.frame(y1,x1)
attach(data.v)
step(lm(y1~gender+married+smoke+age+weight+height+overwt+race+alcohol+
       trt+bmi+stress+salt+chldbear+income+educatn),direction = "both",data=data.v)

#############################################################################
###############all into regularization############
```

```
rm(list = ls())
blood=read.table("D:/U of A/Courses/590/Final-Report/data.txt",header=T,sep="	")
blood=data.frame(blood)
library(glmnet)
library(sm)
library(hydroGOF)
blood.high=blood[251:500,]
x1=blood.high[-c(29,105,116,125,153,180,211),c(2:501)]
y1=blood.high[-c(29,105,116,125,153,180,211),1]
x1=data.matrix(x1)
y1=as.numeric(y1)
y1=log(y1)
fit.gl=glmnet(x1,y1)
summary(fit.gl)
str(fit.gl)
plot(fit.gl, xvar = "lambda", label = TRUE)
cvfit=cv.glmnet(x1, y1)
plot(cvfit)
cvfit$lambda.min
newy=predict(cvfit, x1, s = "lambda.min")
plot(newy)
points(y1,col="red",pch=8)
coef(cvfit,s="lambda.min")
mse = sum((newy - y1) ^ 2) / length(y1)
mse
###############only genes into regularization############
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x1,y1, penalty.factor = p.fac)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x1, y1,penalty.factor = p.fac)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x1, s = "lambda.min",penalty.factor = p.fac)
plot(newy)
points(y1,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y1) ^ 2) / length(y1)
mse
######################Elastic net alpha=0.6###################################
par(mfrow=c(1,2))
p.fac = rep(1, 501)
p.fac[c(1:17)] = 0
fit.gl.pen=glmnet(x1,y1, penalty.factor = p.fac,alpha = 0.6)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x1, y1,penalty.factor = p.fac,alpha = 0.6)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x1, s = "lambda.min",penalty.factor = p.fac)
plot(newy)
points(y1,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y1) ^ 2) / length(y1)
mse

#########################Elastic without penalty#############################################
```

```
par(mfrow=c(1,2))

fit.gl.pen=glmnet(x1,y1,alpha = 0.6)
summary(fit.gl.pen)
#str(fit.gl.pen)
plot(fit.gl.pen, xvar = "lambda", label = TRUE)
cvfit.pen=cv.glmnet(x1, y1,alpha = 0.6)
plot(cvfit.pen)
cvfit.pen$lambda.min
newy=predict(cvfit.pen, x1, s = "lambda.min")
plot(newy)
points(y1,col="red",pch=8)
coef(cvfit.pen,s="lambda.min")
mse = sum((newy - y1) ^ 2) / length(y1)
mse
```

Appendix D: SPSS Syntax

```
DATASET ACTIVATE DataSet1.

FREQUENCIES VARIABLES=sbp age weight height bmi

  /FORMAT=NOTABLE

  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN

  /ORDER=ANALYSIS.

* Bar chart

/GRAPHDATASET NAME="graphdataset" VARIABLES=sbp MISSING=LISTWISE

REPORTMISSING=NO

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: sbp=col(source(s), name("sbp"))
  GUIDE: axis(dim(1), label("sbp"))
  GUIDE: axis(dim(2), label("Frequency"))
  ELEMENT: interval(position(summary.count(bin.rect(sbp))), shape.interior(shape.square))
END GPL.

  * Correlation matrix plot
  /SCATTERPLOT(MATRIX)=sbp age weight height bmi

/MISSING=LISTWISE.

* Box Plots.


/GRAPHDATASET NAME="graphdataset" VARIABLES=gender sbp MISSING=LISTWISE
```

```
REPORTMISSING=NO

/GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  DATA: sbp=col(source(s), name("sbp"))
  DATA: id=col(source(s), name("$CASENUM"), unit.category())
  GUIDE: axis(dim(1), label("gender"))
  GUIDE: axis(dim(2), label("sbp"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: schema(position(bin.quantile.letter(gender*sbp)), label(id))
END GPL.


T-TEST GROUPS=gender('M' 'F')

/MISSING=ANALYSIS

  /VARIABLES=sbp

  /CRITERIA=CI(.95).
```