|  |  |
| --- | --- |
| Regularization Assignment  Assignment 3 | Rebecca Leu  ALY 6015 – Intermediate Analytics |

Introduction

For this assignment I will be using the Heart dataset inside of the nvcreg package. The heart dataset is about the risk factors associated with heart disease. The dataset is a subset from the coronary risk factor study baseline survey from rural South Africa. The dataset has an X and a Y. The X is a matrix with 462 observations and 9 predictor variables. Each x variable is listed below.

* sbp: Systolic blood pressure
* tobacco: Cumulative tobacco consumption, in kg
* ldl: Low-density lipoprotein cholesterol
* adiposity: Adipose tissue concentration
* famhist: Family history of heart disease (1=Present, 0=Absent)
* typea: Score on test designed to measure type-A behavior
* obesity: Obesity
* alcohol: Current consumption of alcohol
* age: Age of subject

The Y is simple a 0 or 1 number for each of the participants with 0=No and 1=Yes to having heard disease.

Part 1– Loading the packages and attaching the heart dataset

> install.packages("ncvreg")

Error in install.packages : Updating loaded packages

> install.packages("bigmemory")

Error in install.packages : Updating loaded packages

> install.packages("biglasso")

Error in install.packages : Updating loaded packages

> install.packages("lars")

Error in install.packages : Updating loaded packages

> install.packages("glmnet")

Error in install.packages : Updating loaded packages

package ‘biglasso’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\rebec\AppData\Local\Temp\RtmpkZza6U\downloaded\_packages

> library(lars)

> library(glmnet)

> library(biglasso)

> library(bigmemory)

> library(ncvreg)

> data("Heart")

> attach(Heart)

> ?Heart

Part 2 – visual inspection of the data

To begin we plot each of the variables in X against the y. I created a loop that plots each x (x(i)) by the y to see how they correlate with each other. It is important to remember that the y dataset only has 0’s and 1’s because they represent yes or no to having heart disease. In these graphs the yes’s are up top and the no’s are on the bottom of each graph.

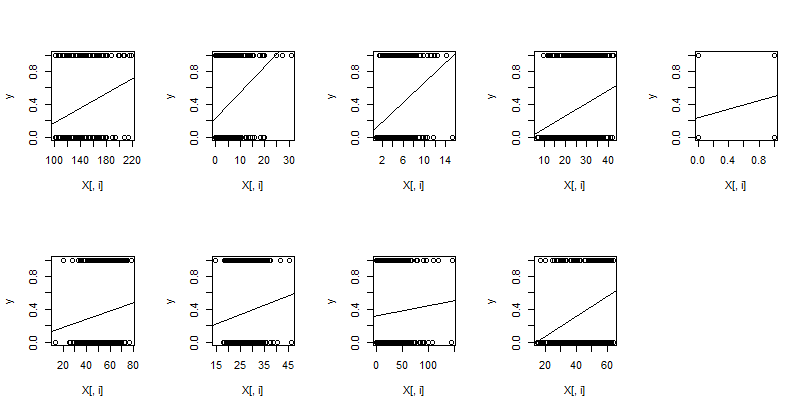
par(mfrow=c(2,5))

> for(i in 1:9){

+ plot(X[,i], y)

+ abline(lm(y~X[,i]))

+ }



The plots are in order of the x’s in the description. By the looks of it some of these factors are high predictors of heart disease. The first is systolic blood pressure. As the blood pressure rises so does the probability that the person has heart disease. This is usually one of the best predictors of heart disease. Additionally, the second graph is of cumulative tobacco consumption again has a positive linear relationship, maybe one of the most positive of them all. The third graph is for low density lipoprotein cholesterol, or high cholesterol, again a high correlation. All three of these are known to be large factors for heart disease. It appears that out of all the observations, current alcohol consumption seems to have one of the lowest positive relationships. Next we are going to see how our initial predictions based off of these graphs lines up with the lasso regularization we are going to use to find a model of perfect fit.

Part 3 – Multiple regression using all x as predictors of heart disease.

> model\_ols <- lm(y ~ X)

> summary(model\_ols)

Call:

lm(formula = y ~ X)

Residuals:

Min 1Q Median 3Q Max

-0.7464 -0.3299 -0.1045 0.3766 1.0348

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.5080306 0.2045921 -2.483 0.013385 \*

Xsbp 0.0013387 0.0010581 1.265 0.206441

Xtobacco 0.0165841 0.0048610 3.412 0.000704 \*\*\*

Xldl 0.0331791 0.0106747 3.108 0.002001 \*\*

Xadiposity 0.0023026 0.0047696 0.483 0.629499

Xfamhist 0.1734290 0.0412746 4.202 3.19e-05 \*\*\*

Xtypea 0.0060817 0.0020369 2.986 0.002983 \*\*

Xobesity -0.0111711 0.0070315 -1.589 0.112821

Xalcohol -0.0002364 0.0008284 -0.285 0.775502

Xage 0.0068440 0.0019868 3.445 0.000625 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4205 on 452 degrees of freedom

Multiple R-squared: 0.236, Adjusted R-squared: 0.2208

F-statistic: 15.51 on 9 and 452 DF, p-value: < 2.2e-16

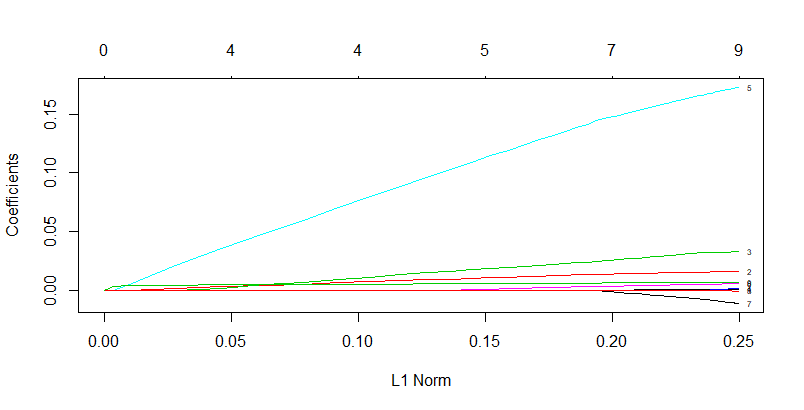
The table lists the most significant predictor values with \*\*\* or \*\*. From this we conclude tobacco use, family history and age are the biggest predictors of heart disease.

Part 4 – Plotting the path of each x against the y vector

> model\_lasso <- glmnet(X, y)

> par(mfrow=c(1,1))

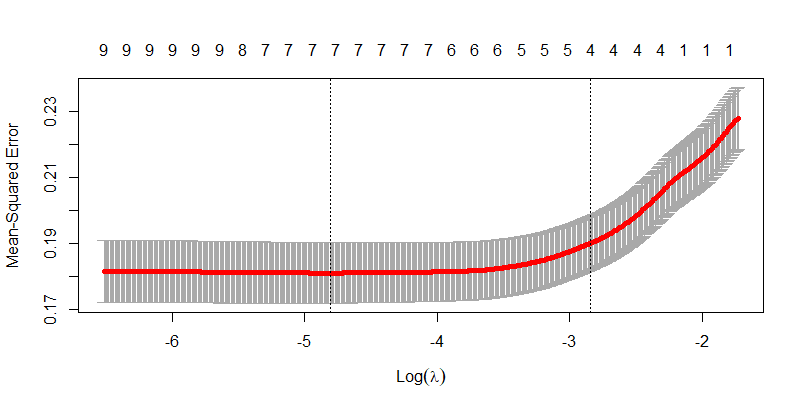
> plot(model\_lasso, xvar = "norm", label = TRUE)



This graph indicates at which stage each coefficient shrinks to zero.

Part 5 – cross validation curve

|  |
| --- |
| > cv\_fit <- cv.glmnet(X, y=y, alpha = 1, nlambda = 1000)  > plot(cv\_fit)  > cv\_fit$lambda.min  [1] 0.008161349 |
|  |
| |  | | --- | | > | |



This cv.glmnet function gives us the cross validation curve which is the red dotted line and the value of lambda that minimizes the mean cross validation error. The two vertical dotted lines are the two lambdas it gives us. One is the lambda.min that gives the *minimum* mean cross validated error. The other is lambda.1se which gives the most regularized model that is within *one standard error* to the minimum. It is with the lambda.min that we use to shrink the predictor values in the next section.

Part 6 – shrinking predictors to zero using lambda.min

> fit <- glmnet(X, y=y, alpha = 1, lambda=cv\_fit$lambda.min)

> fit$beta

9 x 1 sparse Matrix of class "dgCMatrix"

s0

sbp 0.000959776

tobacco 0.015351502

ldl 0.030345986

adiposity .

famhist 0.160383247

typea 0.004982246

obesity -0.005012872

alcohol .

age 0.007030239

Here we used the minimum value of lambda.min from part 5 to get the estimated beta matrix. Some of the coefficients have been shrunk to 0. The ones remaining are the important predictors. Here alcohol use and adipose tissue concentration have been zeroed out.

Part 7 – shrinking further using lamdbda.1se

|  |
| --- |
| > cv\_fit$lambda.1se  [1] 0.05815871  > fit <- glmnet(X, y=y, alpha = 1, lambda=cv\_fit$lambda.1se)  > fit$beta  9 x 1 sparse Matrix of class "dgCMatrix"  s0  sbp .  tobacco 0.008489500  ldl 0.013547774  adiposity .  famhist 0.088633271  typea .  obesity .  alcohol .  age 0.005656981 |
|  |
| |  | | --- | | > | |

For the final shrinking we used the higher value lambda.1se that is within one standard error of the minimum. The last coefficients that we are left with give us the best predictors for y. This time Obesity, the test score predictor and systolic blood pressure are removed from the predictors. This leaves us 4 predictors left in our model to predict heart disease. Age, tobacco use, cholesterol and finally family history is the best model to use moving forward to predict heart disease in other observations.