Logistic Regression to Predict Heart Disease

Olivia Zhang, Harshit Handa, Rory Davies

1 - Set-Up

1.1 Running Code

- Ensure that you have the necessary packages installed ('glmnet')
- Ensure seed is set to 1 ('set.seed(1)') in section 4.1
- Knit rmd file to HTML or pdf

1.2 Importing necessary packages

```
#install.packages(glmnet)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-1
```

1.3 Reading in Data

Run the below code to import the data from GitHub repository

```
#reading in data from github
urlfile<-'https://raw.githubusercontent.com/rory-davies/university-projects/main/framingham.csv'
heartDisease_NA <-read.csv(urlfile)
```

1.4 Cleaning Data

```
#Removing rows with NA values
colSums(is.na(heartDisease_NA))
```

```
##            male             age       education   currentSmoker      cigsPerDay
##               0               0             105               0              29
##          BPMeds prevalentStroke    prevalentHyp        diabetes         totChol
##              53               0               0               0              50
##           sysBP           diaBP             BMI       heartRate         glucose
##               0               0              19               1             388
##      TenYearCHD
##               0
```

```
heartDisease <- na.omit(heartDisease_NA)
```

2 Exploratory Data Analysis

2.1 Dataset Summary

```
summary(heartDisease_NA)
```

```
##       male              age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.006   Mean   :0.02962   Mean   :0.005896   Mean   :0.3106
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.0
##  Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5
##                    NA's   :50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.96   Mean   :0.1519
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
##  NA's   :19      NA's   :1        NA's   :388
```

2.2 Binary/Categorical Variable Counts

```
#sex of the participant
(gender <- table(heartDisease_NA$male))
```

```
##
##    0    1
## 2420 1820
```

```
#edcuation level
(ed <- table(heartDisease_NA$education))
```

```
##
##    1    2    3    4
```

```
## 1720 1253  689  473
```

```
#whether takes blood pressure medication
(bpMeds <- table(heartDisease_NA$BPMeds))
```

```
##
##    0    1
## 4063  124
```

```
#whether has ever had hypertension
(hyp <- table(heartDisease_NA$prevalentHyp))
```

```
##
##    0    1
## 2923 1317
```

```
#whether has ever had stroke
(strk <- table(heartDisease_NA$prevalentStroke))
```

```
##
##    0    1
## 4215   25
```

```
#whether has ever had diabetes
(diab <- table(heartDisease_NA$diabetes))
```

```
##
##    0    1
## 4131  109
```

```
#is a smoker
(smk<- table(heartDisease_NA$currentSmoker))
```

```
##
##    0    1
## 2145 2095
```

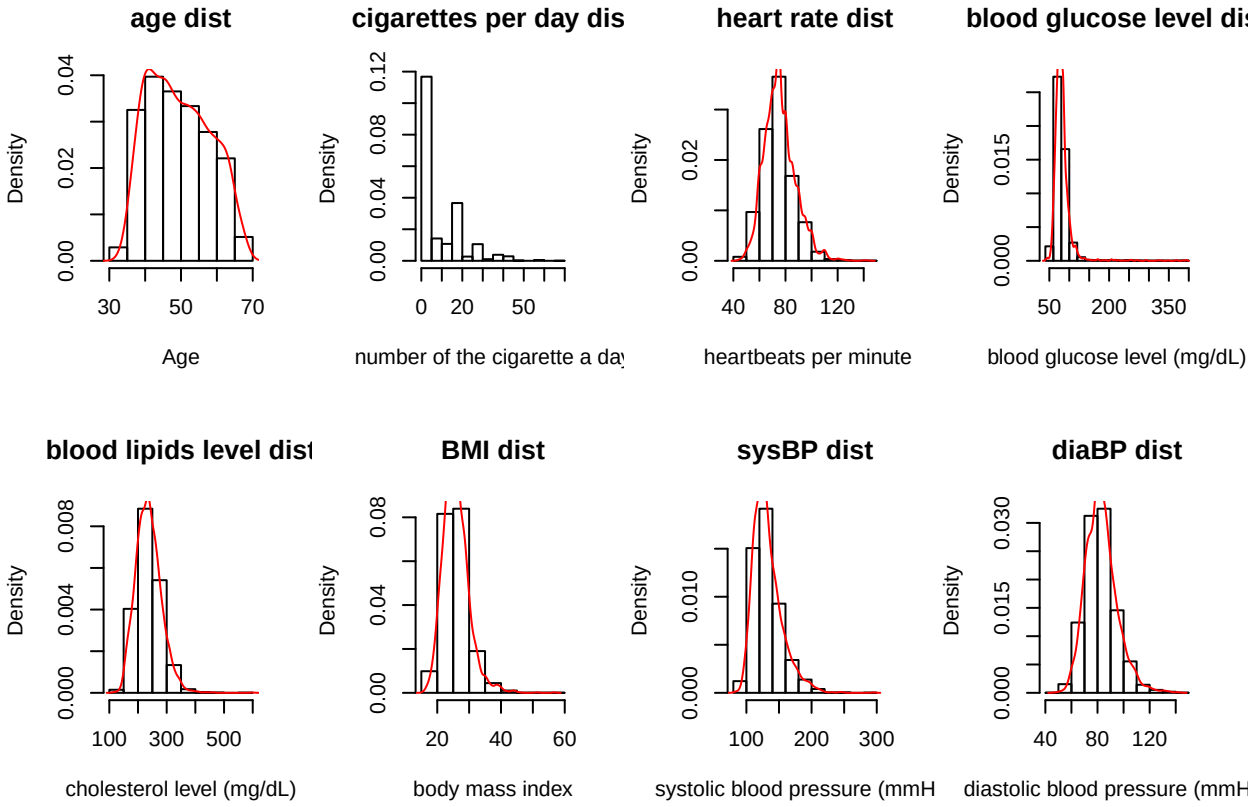
```
# 10 year coronary heart disease risk
(risk <- table(heartDisease_NA$TenYearCHD))
```

```
##
##    0    1
## 3596  644
```

2.3 Continuous Variable Density Plots

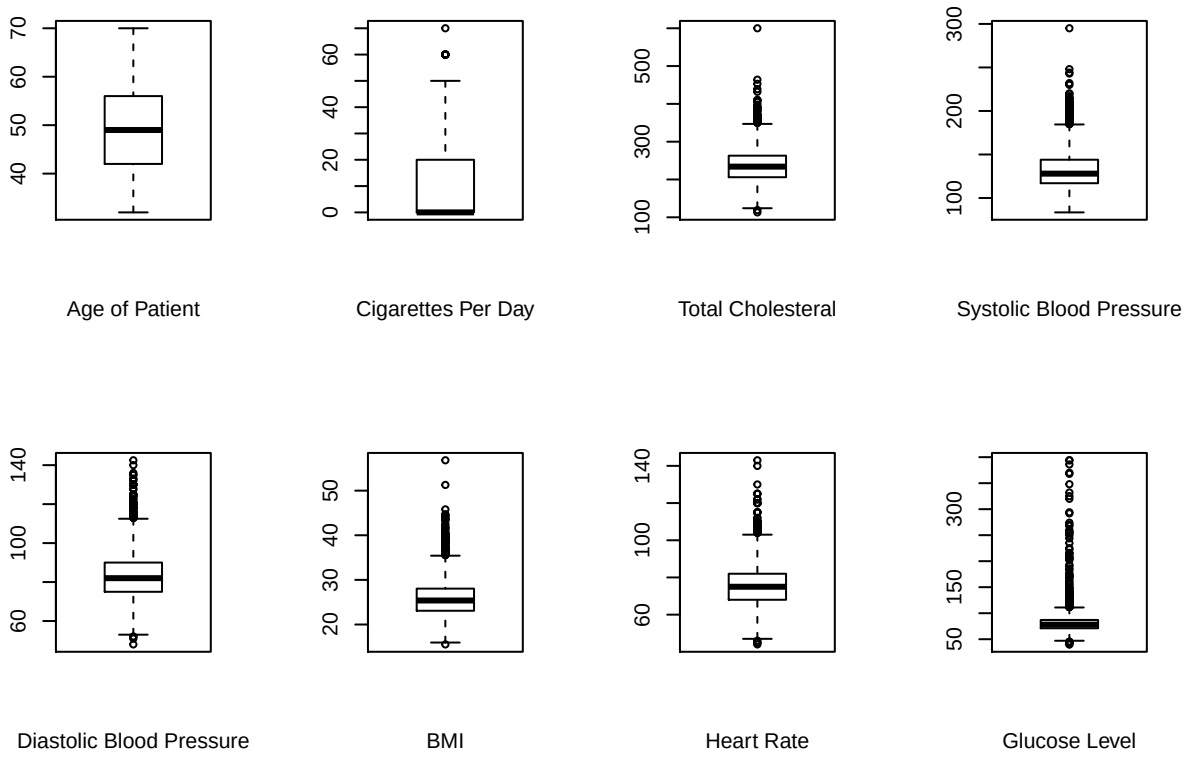
```
par(mfrow = c(2,4))
hist(heartDisease$age, main = "age dist", xlab = "Age", prob = TRUE)
lines(density(heartDisease$age), col = 2)
hist(heartDisease$cigsPerDay, main = "cigarettes per day dist", xlab = "number of the cigarette a day",
hist(heartDisease$heartRate, main = "heart rate dist", xlab = " heartbeats per minute", prob = TRUE)
HR.dist <- na.omit(heartDisease$heartRate)
lines(density(HR.dist), col = 2)
hist(heartDisease$glucose, main = "blood glucose level dist", xlab = "blood glucose level (mg/dL)", prob
glucose.dist <- na.omit(heartDisease$glucose)
lines(density(glucose.dist), col = 2)
hist(heartDisease$totChol, main = "blood lipids level dist", xlab = "cholesterol level (mg/dL)", prob =
```

```r
chol.dist <- na.omit(heartDisease$totChol)
lines(density(chol.dist), col = 2)
hist(heartDisease$BMI, main = "BMI dist", xlab = "body mass index", prob = TRUE)
BMI.dist <- na.omit(heartDisease$BMI)
lines(density(BMI.dist), col = 2)
hist(heartDisease$sysBP, main = "sysBP dist", xlab = "systolic blood pressure (mmHg)", prob = TRUE)
lines(density(heartDisease$sysBP), col = 2)
hist(heartDisease$diaBP, main = "diaBP dist", xlab = "diastolic blood pressure (mmHg)", prob = TRUE)
lines(density(heartDisease$diaBP), col = 2)
```



2.4 Continuous Variables Boxplots

```r
par(mfrow=c(2,4))
boxplot(heartDisease$age, xlab="Age of Patient")
boxplot(heartDisease$cigsPerDay, xlab="Cigarettes Per Day")
boxplot(heartDisease$totChol, xlab="Total Cholesteral")
boxplot(heartDisease$sysBP, xlab="Systolic Blood Pressure")
boxplot(heartDisease$diaBP, xlab="Diastolic Blood Pressure")
boxplot(heartDisease$BMI, xlab="BMI")
boxplot(heartDisease$heartRate, xlab="Heart Rate")
boxplot(heartDisease$glucose, xlab="Glucose Level")
```



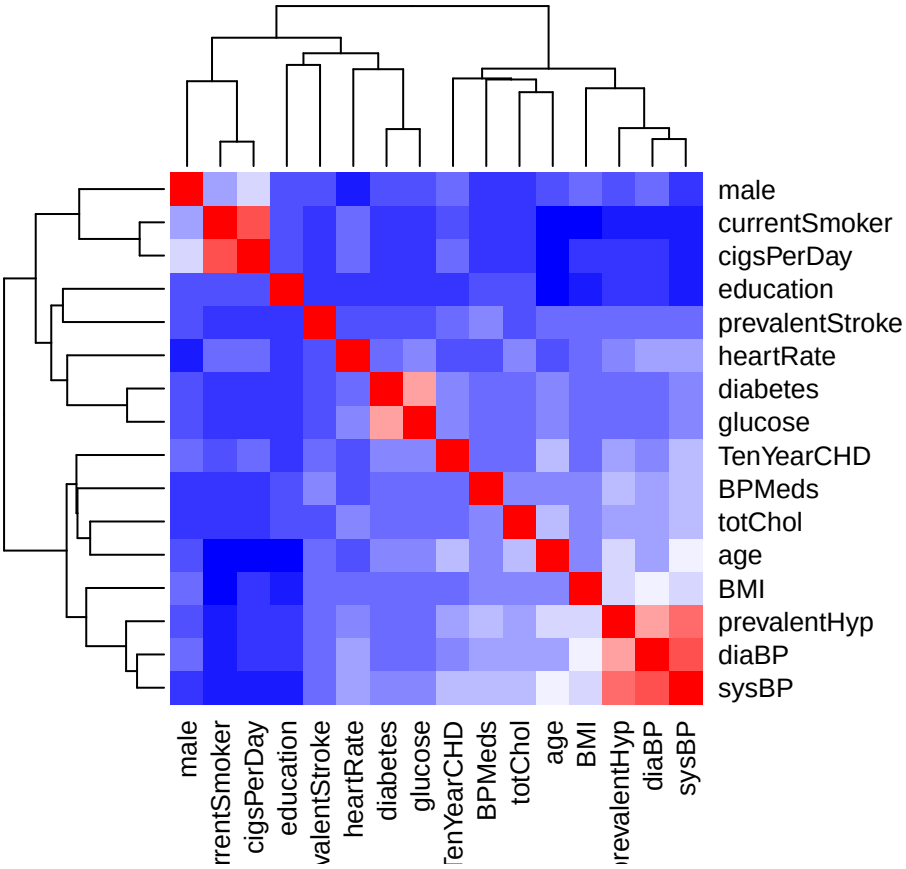
2.5 Correlation Heatmap

```
corr_matrix <- signif(cor(heartDisease),2)
corr_matrix
```

```
##                    male     age education currentSmoker cigsPerDay BPMeds
## male             1.0000 -0.0240     0.018         0.210      0.330 -0.052
## age             -0.0240  1.0000    -0.160        -0.210     -0.190  0.130
## education        0.0180 -0.1600     1.000         0.025      0.014 -0.014
## currentSmoker    0.2100 -0.2100     0.025         1.000      0.770 -0.052
## cigsPerDay       0.3300 -0.1900     0.014         0.770      1.000 -0.047
## BPMeds          -0.0520  0.1300    -0.014        -0.052     -0.047  1.000
## prevalentStroke -0.0023  0.0510    -0.030        -0.038     -0.036  0.110
## prevalentHyp     0.0014  0.3100    -0.079        -0.110     -0.070  0.260
## diabetes         0.0140  0.1100    -0.040        -0.042     -0.037  0.049
## totChol         -0.0700  0.2700    -0.014        -0.051     -0.030  0.094
## sysBP           -0.0450  0.3900    -0.120        -0.130     -0.095  0.270
## diaBP            0.0520  0.2100    -0.058        -0.120     -0.057  0.200
## BMI              0.0730  0.1400    -0.140        -0.160     -0.087  0.110
## heartRate       -0.1200 -0.0027    -0.064         0.051      0.064  0.013
## glucose          0.0029  0.1200    -0.032        -0.053     -0.054  0.054
## TenYearCHD       0.0920  0.2300    -0.063         0.019      0.052  0.089
##                 prevalentStroke prevalentHyp diabetes totChol  sysBP  diaBP
## male                    -0.0023       0.0014   0.0140  -0.070 -0.045  0.052
## age                      0.0510       0.3100   0.1100   0.270  0.390  0.210
## education               -0.0300      -0.0790  -0.0400  -0.014 -0.120 -0.058
## currentSmoker           -0.0380      -0.1100  -0.0420  -0.051 -0.130 -0.120
## cigsPerDay              -0.0360      -0.0700  -0.0370  -0.030 -0.095 -0.057
```

```
## BPMeds                   0.1100       0.2600   0.0490   0.094  0.270  0.200
## prevalentStroke          1.0000       0.0660   0.0096   0.013  0.061  0.056
## prevalentHyp             0.0660       1.0000   0.0810   0.170  0.700  0.620
## diabetes                 0.0096       0.0810   1.0000   0.048  0.100  0.051
## totChol                  0.0130       0.1700   0.0480   1.000  0.220  0.170
## sysBP                    0.0610       0.7000   0.1000   0.220  1.000  0.790
## diaBP                    0.0560       0.6200   0.0510   0.170  0.790  1.000
## BMI                      0.0360       0.3000   0.0890   0.120  0.330  0.390
## heartRate               -0.0170       0.1500   0.0610   0.093  0.180  0.180
## glucose                  0.0160       0.0870   0.6100   0.050  0.130  0.064
## TenYearCHD               0.0480       0.1800   0.0930   0.091  0.220  0.150
##                     BMI heartRate glucose TenYearCHD
## male              0.073   -0.1200  0.0029      0.092
## age               0.140   -0.0027  0.1200      0.230
## education        -0.140   -0.0640 -0.0320     -0.063
## currentSmoker    -0.160    0.0510 -0.0530      0.019
## cigsPerDay       -0.087    0.0640 -0.0540      0.052
## BPMeds            0.110    0.0130  0.0540      0.089
## prevalentStroke   0.036   -0.0170  0.0160      0.048
## prevalentHyp      0.300    0.1500  0.0870      0.180
## diabetes          0.089    0.0610  0.6100      0.093
## totChol           0.120    0.0930  0.0500      0.091
## sysBP             0.330    0.1800  0.1300      0.220
## diaBP             0.390    0.1800  0.0640      0.150
## BMI               1.000    0.0740  0.0840      0.082
## heartRate         0.074    1.0000  0.0970      0.021
## glucose           0.084    0.0970  1.0000      0.120
## TenYearCHD        0.082    0.0210  0.1200      1.000
```

```r
col <- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(corr_matrix, col=col, symm=TRUE)
```



3 Testing Significance of Predictors

3.1 Logistic Regression

```
mod2<- glm(TenYearCHD~male + age + cigsPerDay + prevalentStroke +  prevalentHyp + totChol + sysBP + glu
summary(mod2)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
##     prevalentHyp + totChol + sysBP + glucose, family = binomial(link = "logit"),
##     data = heartDisease)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0063  -0.5972  -0.4291  -0.2841   2.8634
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.745885   0.522456 -16.740  < 2e-16 ***
## male             0.553297   0.107018   5.170 2.34e-07 ***
## age              0.065411   0.006442  10.153  < 2e-16 ***
## cigsPerDay       0.019579   0.004181   4.683 2.82e-06 ***
## prevalentStroke  0.751698   0.483585   1.554   0.1201
```

```
## prevalentHyp     0.225762   0.135085   1.671   0.0947 .
## totChol          0.002257   0.001122   2.011   0.0443 *
## sysBP            0.014218   0.002857   4.976 6.50e-07 ***
## glucose          0.007317   0.001673   4.374 1.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3121.2  on 3657  degrees of freedom
## Residual deviance: 2757.5  on 3649  degrees of freedom
## AIC: 2775.5
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(mod2)
```

```
## [1] 2775.499
```

3.2 Confidence Intervals

```
confint(mod2)
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %       97.5 %
## (Intercept)     -9.782417e+00 -7.733620198
## male             3.439588e-01  0.763649345
## age              5.284871e-02  0.078112737
## cigsPerDay       1.135931e-02  0.027757026
## prevalentStroke -2.337237e-01  1.687457087
## prevalentHyp    -4.000911e-02  0.489727856
## totChol          4.562306e-05  0.004448414
## sysBP            8.627093e-03  0.019836944
## glucose          4.058943e-03  0.010641788
```

Model Selection

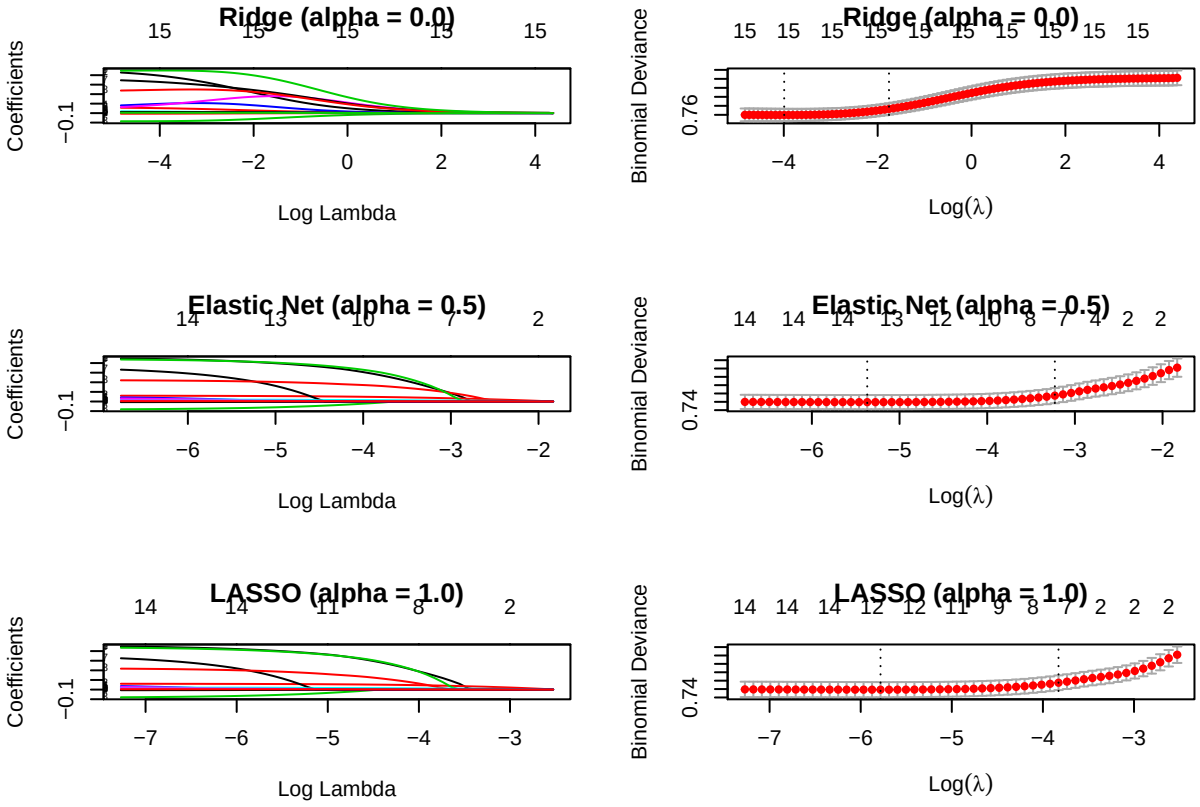
4.1 Elastic Net Model

```
#Setting seed
set.seed(1)

#Creating Prediction and Target Matrices
x <- as.matrix(heartDisease[ , -grep("TenYearCHD", colnames(heartDisease))])
y <- as.matrix(heartDisease[ , grep("TenYearCHD", colnames(heartDisease))])

#Splitting data into training (80%) and testing (20%) sets
train_rows <- sample(1:nrow(heartDisease), 0.8*nrow(heartDisease))

x.train <- x[train_rows, ]
x.test <- x[-train_rows, ]
```

```
y.train <- y[train_rows]
y.test <- y[-train_rows]

#Fitting regularization models for plots
fit.lasso <- glmnet(x.train, y.train, family="binomial", alpha=1)
fit.ridge <- glmnet(x.train, y.train, family="binomial", alpha=0)
fit.elnet <- glmnet(x.train, y.train, family="binomial", alpha=.5)

#10-fold cross validation on each alpha = 0.0, 0.1, 0.2,...,0.9, 1.0
for(i in 0:10){
  assign(paste("fit", i, sep=""), cv.glmnet(x.train, y.train,  type.measure="deviance", alpha=i/10, fami
}
```

4.2 Plotting The Effect of Lambda on Model Coefficients and Deviance

```
#Plotting solution paths:
par(mfrow=c(3,2))
plot(fit.ridge, xvar="lambda", label = TRUE, main="Ridge (alpha = 0.0)")
plot(fit0, main="Ridge (alpha = 0.0)")

plot(fit.elnet, xvar="lambda", label = TRUE, main="Elastic Net (alpha = 0.5)")
plot(fit5, main="Elastic Net (alpha = 0.5)")

plot(fit.lasso, xvar="lambda", label = TRUE, main="LASSO (alpha = 1.0)")
plot(fit10, main="LASSO (alpha = 1.0)")
```



4.3 Prediction & Model Comparison

```
#Predicting target variable from x.test data for each model
yhat0 <- predict(fit0, s='lambda.min', newx=x.test, type = 'response')
yhat1 <- predict(fit1, s='lambda.1se', newx=x.test, type = 'response')
yhat2 <- predict(fit2, s='lambda.1se', newx=x.test, type = 'response')
yhat3 <- predict(fit3, s='lambda.1se', newx=x.test, type = 'response')
yhat4 <- predict(fit4, s='lambda.1se', newx=x.test, type = 'response')
yhat5 <- predict(fit5, s='lambda.1se', newx=x.test, type = 'response')
yhat6 <- predict(fit6, s='lambda.1se', newx=x.test, type = 'response')
yhat7 <- predict(fit7, s='lambda.1se', newx=x.test, type = 'response')
yhat8 <- predict(fit8, s='lambda.1se', newx=x.test, type = 'response')
yhat9 <- predict(fit9, s='lambda.1se', newx=x.test, type = 'response')
yhat10 <- predict(fit10, s='lambda.1se', newx=x.test, type = 'response')


logloss0 <- -mean(y.test*log(yhat0) + (1-y.test)*log(1-yhat0))
logloss1 <- -mean(y.test*log(yhat1) + (1-y.test)*log(1-yhat1))
logloss2 <- -mean(y.test*log(yhat2) + (1-y.test)*log(1-yhat2))
logloss3 <- -mean(y.test*log(yhat3) + (1-y.test)*log(1-yhat3))
logloss4 <- -mean(y.test*log(yhat4) + (1-y.test)*log(1-yhat4))
logloss5 <- -mean(y.test*log(yhat5) + (1-y.test)*log(1-yhat5))
logloss6 <- -mean(y.test*log(yhat6) + (1-y.test)*log(1-yhat6))
logloss7 <- -mean(y.test*log(yhat7) + (1-y.test)*log(1-yhat7))
logloss8 <- -mean(y.test*log(yhat8) + (1-y.test)*log(1-yhat8))
logloss9 <- -mean(y.test*log(yhat9) + (1-y.test)*log(1-yhat9))
logloss10 <- -mean(y.test*log(yhat10) + (1-y.test)*log(1-yhat10))
logloss <- c(logloss0, logloss1, logloss2, logloss3, logloss4, logloss5, logloss6, logloss7, logloss8, l

# #Calculating classificaion rate for each model (probability cut off = 0.5)
class0 <- mean(ifelse(yhat0 > 0.5, 1, 0) == y.test)
class1 <- mean(ifelse(yhat1 > 0.5, 1, 0) == y.test)
class2 <- mean(ifelse(yhat2 > 0.5, 1, 0) == y.test)
class3 <- mean(ifelse(yhat3 > 0.5, 1, 0) == y.test)
class4 <- mean(ifelse(yhat4 > 0.5, 1, 0) == y.test)
class5 <- mean(ifelse(yhat5 > 0.5, 1, 0) == y.test)
class6 <- mean(ifelse(yhat6 > 0.5, 1, 0) == y.test)
class7 <- mean(ifelse(yhat7 > 0.5, 1, 0) == y.test)
class8 <- mean(ifelse(yhat8 > 0.5, 1, 0) == y.test)
class9 <- mean(ifelse(yhat9 > 0.5, 1, 0) == y.test)
class10 <- mean(ifelse(yhat10 > 0.5, 1, 0) == y.test)
class <- c(class0,class1,class2,class3,class4,class5,class6,class7,class8,class9,class10)

models <- c('alpha = 0.0','alpha = 0.1','alpha = 0.2','alpha = 0.3','alpha = 0.4','alpha = 0.5','alpha =

df <- data.frame("Elastic Net Model"=models, 'Log Loss' = logloss, 'Classification Rate'=class)
df
```

```
##    Elastic.Net.Model  Log.Loss Classification.Rate
## 1        alpha = 0.0 0.3916706           0.8456284
## 2        alpha = 0.1 0.4070810           0.8346995
## 3        alpha = 0.2 0.4070473           0.8346995
## 4        alpha = 0.3 0.4054806           0.8360656
## 5        alpha = 0.4 0.4033640           0.8360656
```

```
## 6        alpha = 0.5 0.4019896           0.8360656
## 7        alpha = 0.6 0.4058736           0.8360656
## 8        alpha = 0.7 0.4081113           0.8360656
## 9        alpha = 0.8 0.4075168           0.8360656
## 10       alpha = 0.9 0.4100745           0.8360656
## 11       alpha = 1.0 0.4012105           0.8360656
```

4.4 Analyzing Best Model (Ridge)

```
#Converting predicted probabilities into binary values (threshold = 0.5)
yhat0_b <- ifelse(yhat0 > 0.5, 1, 0)

#Viewing coefficients of predictors
coef(fit0, s='lambda.1se')
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)     -5.262600969
## male             0.190565353
## age              0.024996620
## education       -0.057095883
## currentSmoker    0.073772583
## cigsPerDay       0.006690691
## BPMeds           0.177906635
## prevalentStroke  0.228844827
## prevalentHyp     0.212000216
## diabetes         0.378231413
## totChol          0.001704907
## sysBP            0.006390526
## diaBP            0.005608720
## BMI              0.009908455
## heartRate       -0.002053719
## glucose          0.002845258
```

```
#Creating confusion matrix
table(yhat0_b, y.test)
```

```
##        y.test
## yhat0_b   0   1
##       0 610 112
##       1   1   9
```

```
(accuracy <- (610+9)/(610+112+1+9))
```

```
## [1] 0.8456284
```

```
(sensitivity <- 9/(9+112))
```

```
## [1] 0.07438017
```

```
(specificity <- 610/(610+1))
```

```
## [1] 0.9983633
```

```
(type1_error <- 1/(610+112+1+9))
```

```
## [1] 0.00136612
```

```r
(type2_error <- 112/(610+112+1+9))
```

```
## [1] 0.1530055
```