STAT 489 HW3

Nicholas Thompson

9/27/2021

#Question 1  
set.seed(117649)  
heart.data=read.csv("heart.csv", header = TRUE)  
0.7\*918

## [1] 642.6

heart.ind=sample(1:918, 643)  
heart.train=heart.data[heart.ind,]  
heart.test=heart.data[-heart.ind,]  
str(heart.data)

## 'data.frame': 918 obs. of 12 variables:  
## $ Age : int 40 49 37 48 54 39 45 54 37 48 ...  
## $ Sex : chr "M" "F" "M" "F" ...  
## $ ChestPainType : chr "ATA" "NAP" "ATA" "ASY" ...  
## $ RestingBP : int 140 160 130 138 150 120 130 110 140 120 ...  
## $ Cholesterol : int 289 180 283 214 195 339 237 208 207 284 ...  
## $ FastingBS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ RestingECG : chr "Normal" "Normal" "ST" "Normal" ...  
## $ MaxHR : int 172 156 98 108 122 170 170 142 130 120 ...  
## $ ExerciseAngina: chr "N" "N" "N" "Y" ...  
## $ Oldpeak : num 0 1 0 1.5 0 0 0 0 1.5 0 ...  
## $ ST\_Slope : chr "Up" "Flat" "Up" "Flat" ...  
## $ HeartDisease : int 0 1 0 1 0 0 0 0 1 0 ...

#Numeric variables are Age, RestingBP, Cholesterol, MaxHR, Oldpeak  
  
  
cor.heart=heart.train[,c(1,4,5,8,10)]  
str(cor.heart)

## 'data.frame': 643 obs. of 5 variables:  
## $ Age : int 64 42 44 60 56 49 42 41 43 40 ...  
## $ RestingBP : int 140 120 108 132 170 120 134 110 120 130 ...  
## $ Cholesterol: int 313 198 141 218 388 297 240 250 175 275 ...  
## $ MaxHR : int 133 155 175 140 122 132 160 142 120 150 ...  
## $ Oldpeak : num 0.2 0 0.6 1.5 2 1 0 0 1 0 ...

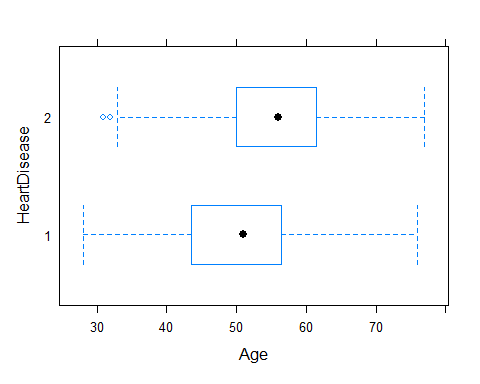
cor(cor.heart)

## Age RestingBP Cholesterol MaxHR Oldpeak  
## Age 1.0000000 0.30862506 -0.09485400 -0.3725609 0.25722143  
## RestingBP 0.3086251 1.00000000 0.07190617 -0.1204480 0.22334119  
## Cholesterol -0.0948540 0.07190617 1.00000000 0.2384678 0.04983467  
## MaxHR -0.3725609 -0.12044803 0.23846775 1.0000000 -0.15218520  
## Oldpeak 0.2572214 0.22334119 0.04983467 -0.1521852 1.00000000

#No signs of multi-collinearity, may need to check VIF  
vif.mod=lm(HeartDisease~Age+RestingBP+Cholesterol+MaxHR+Oldpeak,data = heart.train)  
vif(vif.mod,data=heart.train)

## Age RestingBP Cholesterol MaxHR Oldpeak   
## 1.305657 1.144057 1.080044 1.228859 1.108764

#No VIF above 5, no multi-collinearity  
  
bwplot(HeartDisease~Age, data = heart.train) #displays 5 number summary in graphical form



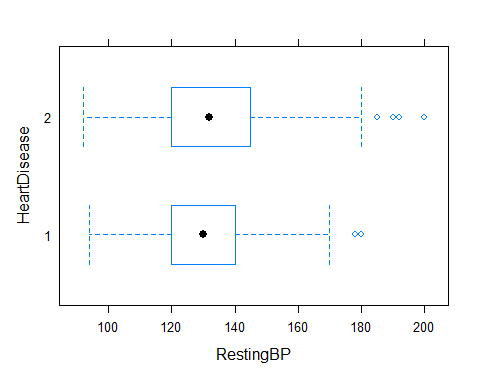
#for some reason, the levels of HeartDisease displayed are 1 and 2. This could just be a quirk to the mosaic boxplot function, but let's check by computing the 5 number summary numerically.  
  
numcomp1=filter(heart.train, HeartDisease == 0) #To give a 5 number summary for ages of the two response levels  
numcomp2=filter(heart.train, HeartDisease == 1)  
favstats(~Age, data = numcomp1) #displays 5 number summary in numerical form

## min Q1 median Q3 max mean sd n missing  
## 28 43.5 51 56.5 76 50.72509 9.248673 291 0

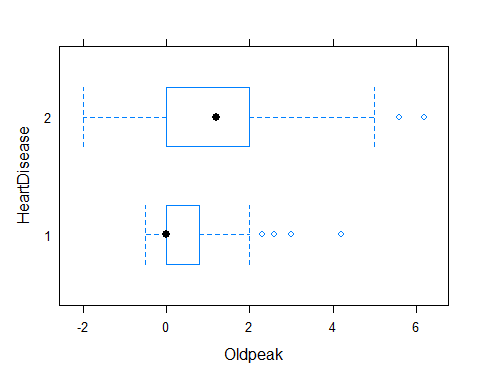
favstats(~Age, data = numcomp2)

## min Q1 median Q3 max mean sd n missing  
## 31 50 56 61.25 77 55.47443 8.846872 352 0

#Comparing the boxplots to the outputs from favstats, the lower boxplot on the display corresponds to HeartDisease = 0, and the upper one to HeartDisease = 1.   
#When plotting the response against age, we see that the medians for age in each state of the response lie close to the mean. Box for Age when HeartDisease=0 is fairly symmetrical with median close to mean. Median age (and also average age) is greater when the patient has heart disease. Based on the boxplots, Age may be a predictor to include in the model  
  
bwplot(HeartDisease~RestingBP, data=heart.train)



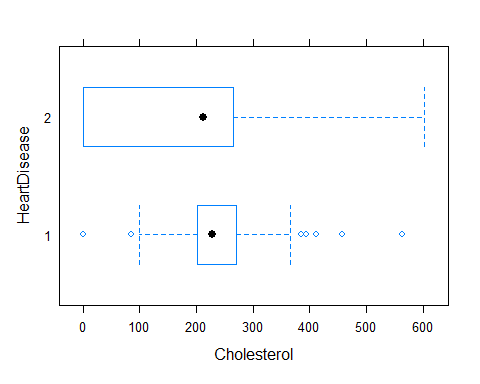
#A similar situation to Age when HeartDisease=0, however there are 2 outliers beyond the maximum. The median Resting BP for both states of the response look to be around the same value, but box for response=0 is shorter, suggesting that the data has a tighter distribution in this state. However, the median values being so close together suggest that RestingBP may not be needed to be in the model.   
  
bwplot(HeartDisease~Oldpeak, data=heart.train)



#By far the most interesting of the variables thus far, when HeartDisease=0, the median and the first quartile for old peak are equivalent (both are 0). When HeartDisease=1, however, the mean and median are 4 hundredths apart and the box has a slight skew towards the right with 2 outliers beyond the maximum. The difference in distributions suggest that this predictor should be included  
favstats(Oldpeak~HeartDisease, data=heart.train) #to see if median and mean are equal to each other when response equals 0

## HeartDisease min Q1 median Q3 max mean sd n missing  
## 1 0 -0.5 0 0.0 0.8 4.2 0.432646 0.6987177 291 0  
## 2 1 -2.0 0 1.2 2.0 6.2 1.241761 1.1360401 352 0

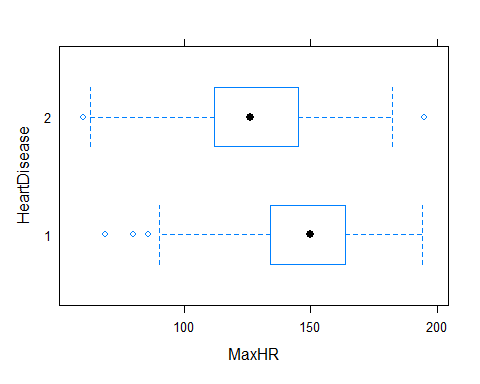
bwplot(HeartDisease~Cholesterol, data = heart.train)



#When HeartDisease=1, the minimum and 1st quartile are 0 and the data is quite varied and spread out, with the median being closer to the 3rd quartile and a lengthy range between Q3 and Q4. When HeartDisease=0, the data is more tightly distributed with several outliers (most beyond the maximum). Perhaps this suggests that Cholesterol levels are less varied when the patient does not have heart disease, which would make it a statistically significant predictor.  
favstats(Cholesterol~HeartDisease, data=heart.train)

## HeartDisease min Q1 median Q3 max mean sd n missing  
## 1 0 0 201 228 271.5 564 232.2509 70.75345 291 0  
## 2 1 0 0 213 265.0 603 172.2699 124.20062 352 0

bwplot(HeartDisease~MaxHR, data = heart.train)



#When HeartDisease=0, meadian max heart rate is 150, with more skew towards the minimum and 1st quartile. However, median max heart rate is lower when the patient has heart disease. Based on distribution of values, patients with heart disease achieve lower max heart rates than patients without heart disease. Inclusion as a predictor seems likely  
  
  
#

#Question 2  
log.reg1=glm(HeartDisease ~Age+ Sex+ChestPainType+RestingBP+Cholesterol+FastingBS+RestingECG+MaxHR+ExerciseAngina+Oldpeak+ST\_Slope, data=heart.train, family=binomial)  
summary(log.reg1)

##   
## Call:  
## glm(formula = HeartDisease ~ Age + Sex + ChestPainType + RestingBP +   
## Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +   
## Oldpeak + ST\_Slope, family = binomial, data = heart.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7050 -0.3894 0.1527 0.4677 2.5611   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.084164 1.683447 -0.644 0.519566   
## Age 0.013986 0.015894 0.880 0.378899   
## SexM 1.675329 0.334954 5.002 5.68e-07 \*\*\*  
## ChestPainTypeATA -2.004265 0.379733 -5.278 1.31e-07 \*\*\*  
## ChestPainTypeNAP -1.892084 0.327175 -5.783 7.33e-09 \*\*\*  
## ChestPainTypeTA -1.832757 0.480131 -3.817 0.000135 \*\*\*  
## RestingBP 0.002709 0.007729 0.351 0.725925   
## Cholesterol -0.005327 0.001328 -4.010 6.07e-05 \*\*\*  
## FastingBS 1.441661 0.330510 4.362 1.29e-05 \*\*\*  
## RestingECGNormal -0.336710 0.315876 -1.066 0.286443   
## RestingECGST -0.245898 0.409575 -0.600 0.548258   
## MaxHR -0.003008 0.005814 -0.517 0.604874   
## ExerciseAnginaY 0.847709 0.291727 2.906 0.003663 \*\*   
## Oldpeak 0.443574 0.146450 3.029 0.002455 \*\*   
## ST\_SlopeFlat 1.504780 0.524213 2.871 0.004098 \*\*   
## ST\_SlopeUp -0.617539 0.554260 -1.114 0.265207   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 885.59 on 642 degrees of freedom  
## Residual deviance: 423.93 on 627 degrees of freedom  
## AIC: 455.93  
##   
## Number of Fisher Scoring iterations: 6

#Looking at individual t-tests, we see that variables Age, RestingBP, RestingECG, and MaxHR are not statistically significant  
  
#HeartDisease = 1  
  
#Coefficient for cholesterol: -0.005327  
exp(-0.005327)

## [1] 0.9946872

#Interpretation for Cholesterol coefficient: An increase in 1 mm/dl of cholesterol multiplies the odds of developing heart disease by 0.9946872  
  
#Coefficient for Sex: 1.675329  
exp(1.675329)

## [1] 5.340552

#Interpretation for Sex coefficient (when sex is male): The male sex has 5.340552 times the odds of developing heart disease compared to the female sex  
  
  
logreg.probs=predict(log.reg1,heart.test)  
logreg.pred= rep(1, length(logreg.probs))  
logreg.pred[logreg.probs < 0.5] = 0  
table(logreg.pred, heart.test$HeartDisease)

##   
## logreg.pred 0 1  
## 0 108 23  
## 1 11 133

mean(logreg.pred != heart.test$HeartDisease)#test error rate

## [1] 0.1236364

mean(logreg.pred == heart.test$HeartDisease) #fraction of correct predictions

## [1] 0.8763636

#The amount of patients predicted to have heart disease when in fact they didn't was 11. This is juxtaposed with 23 predicted didn't have heart disease but actually did. In terms of mistakes, it gave more false negatives over false positives.

#Question 3  
#LDA  
lda.fit=lda(HeartDisease~., data = heart.train, subset = heart.ind)  
lda.fit

## Call:  
## lda(HeartDisease ~ ., data = heart.train, subset = heart.ind)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.4547511 0.5452489   
##   
## Group means:  
## Age SexM ChestPainTypeATA ChestPainTypeNAP ChestPainTypeTA  
## 0 50.23383 0.6268657 0.37313433 0.3283582 0.07960199  
## 1 55.68880 0.9087137 0.04564315 0.1369295 0.04564315  
## RestingBP Cholesterol FastingBS RestingECGNormal RestingECGST MaxHR  
## 0 129.9552 232.3930 0.1094527 0.6517413 0.1492537 147.6219  
## 1 134.4315 172.6515 0.3609959 0.5767635 0.2282158 127.0456  
## ExerciseAnginaY Oldpeak ST\_SlopeFlat ST\_SlopeUp  
## 0 0.1144279 0.3985075 0.2437811 0.7363184  
## 1 0.6058091 1.2307054 0.7219917 0.1867220  
##   
## Coefficients of linear discriminants:  
## LD1  
## Age 0.015838040  
## SexM 0.770180121  
## ChestPainTypeATA -1.234125665  
## ChestPainTypeNAP -1.139823249  
## ChestPainTypeTA -1.065937738  
## RestingBP -0.002726621  
## Cholesterol -0.001978274  
## FastingBS 0.628728807  
## RestingECGNormal -0.190670483  
## RestingECGST -0.087976960  
## MaxHR -0.003827708  
## ExerciseAnginaY 0.570508091  
## Oldpeak 0.223268285  
## ST\_SlopeFlat 0.404710887  
## ST\_SlopeUp -0.602558838

lda.pred=predict(lda.fit,heart.test)$class  
table(lda.pred, heart.test$HeartDisease)

##   
## lda.pred 0 1  
## 0 100 15  
## 1 19 141

mean(lda.pred != heart.test$HeartDisease)#test error rate

## [1] 0.1236364

1-mean(lda.pred != heart.test$HeartDisease)#fraction of correct predictions

## [1] 0.8763636

#The test error rate appears to be the same as the logistics regression  
  
#QDA  
qda.heart=qda(HeartDisease~., data = heart.train)  
qda.pred=predict(qda.heart,heart.test)$class  
table(qda.pred, heart.test$HeartDisease)

##   
## qda.pred 0 1  
## 0 100 20  
## 1 19 136

err.qda=mean(qda.pred != heart.test$HeartDisease)  
err.qda#test error rate

## [1] 0.1418182

1-err.qda# %correct predictions

## [1] 0.8581818

#QDA's correct predictions decreased 2% from the LDA  
  
#KNN  
knn.heart=lm(HeartDisease~Age+RestingBP +Cholesterol + MaxHR + Oldpeak, data = heart.train)  
train.x=heart.train[,c("Age","RestingBP", "Cholesterol", "MaxHR", "Oldpeak")]  
test.x=heart.test[,c("Age","RestingBP", "Cholesterol", "MaxHR", "Oldpeak")]  
train.y=heart.train[,"HeartDisease"]  
knn.pred=knn(train.x, test.x, train.y, k=11)  
table(knn.pred, heart.test$HeartDisease)

##   
## knn.pred 0 1  
## 0 82 45  
## 1 37 111

mean(knn.pred != heart.test$HeartDisease)

## [1] 0.2981818

#below is a list of test error rates for different values of k  
#k=2: 0.3672727  
#k=4: 0.32  
#k=6: 0.3054545  
#k=11: 0.2981818

#Which one of the methods provides the better results?  
  
#Based on what I have calculated, both Logistic Regression and LDA methods produce the same test error rate. This could be because we are dealing with a 2 class setting, which makes logistic regression and LDA closely connected. However, based on the confusion matrix, we see that LDA had less false negative predictions compared to logistic regression. Therefore, I'm inclined to say that the LDA is the best method for this data.

#Question 5  
  
#first let's try removing predictors from 2 that were not statistically significant  
  
log.reg2=glm(HeartDisease~Sex+ChestPainType+Cholesterol+FastingBS+ExerciseAngina+Oldpeak+ST\_Slope, data=heart.train, family=binomial)  
  
summary(log.reg2)

##   
## Call:  
## glm(formula = HeartDisease ~ Sex + ChestPainType + Cholesterol +   
## FastingBS + ExerciseAngina + Oldpeak + ST\_Slope, family = binomial,   
## data = heart.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7404 -0.3832 0.1594 0.4535 2.6834   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.614235 0.689444 -0.891 0.372975   
## SexM 1.657996 0.332834 4.981 6.31e-07 \*\*\*  
## ChestPainTypeATA -2.068991 0.373409 -5.541 3.01e-08 \*\*\*  
## ChestPainTypeNAP -1.910160 0.320553 -5.959 2.54e-09 \*\*\*  
## ChestPainTypeTA -1.785460 0.466571 -3.827 0.000130 \*\*\*  
## Cholesterol -0.005324 0.001267 -4.201 2.66e-05 \*\*\*  
## FastingBS 1.489602 0.326055 4.569 4.91e-06 \*\*\*  
## ExerciseAnginaY 0.879206 0.281155 3.127 0.001765 \*\*   
## Oldpeak 0.478847 0.142400 3.363 0.000772 \*\*\*  
## ST\_SlopeFlat 1.478998 0.524561 2.819 0.004810 \*\*   
## ST\_SlopeUp -0.680764 0.550859 -1.236 0.216525   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 885.59 on 642 degrees of freedom  
## Residual deviance: 427.68 on 632 degrees of freedom  
## AIC: 449.68  
##   
## Number of Fisher Scoring iterations: 6

logreg.probs=predict(log.reg2,heart.test)  
logreg.pred= rep(1, length(logreg.pred))  
logreg.pred[logreg.probs < 0.5] = 0  
table(logreg.pred, heart.test$HeartDisease)

##   
## logreg.pred 0 1  
## 0 107 24  
## 1 12 132

mean(logreg.pred != heart.test$HeartDisease)#test error rate

## [1] 0.1309091

mean(logreg.pred == heart.test$HeartDisease)

## [1] 0.8690909

#test error rate up by 1%  
  
#LDA  
lda2=lda(HeartDisease~Sex+ChestPainType+Cholesterol+FastingBS+ExerciseAngina+Oldpeak+ST\_Slope, data = heart.train)  
  
lda2.pred=predict(lda2,heart.test)$class  
table(lda2.pred, heart.test$HeartDisease)

##   
## lda2.pred 0 1  
## 0 102 17  
## 1 17 139

mean(lda2.pred != heart.test$HeartDisease)#test error rate

## [1] 0.1236364

1-mean(lda2.pred != heart.test$HeartDisease)#fraction of correct predictions

## [1] 0.8763636

#Error for LDA didn't change  
  
qda2.heart=qda(HeartDisease~Sex+ChestPainType+Cholesterol+FastingBS+ExerciseAngina+Oldpeak+ST\_Slope, data = heart.train)  
qda2.pred=predict(qda2.heart,heart.test)$class  
table(qda2.pred, heart.test$HeartDisease)

##   
## qda2.pred 0 1  
## 0 101 18  
## 1 18 138

err.qda=mean(qda2.pred != heart.test$HeartDisease)  
err.qda#test error rate

## [1] 0.1309091

1-err.qda# %correct predictions

## [1] 0.8690909

#slightly better than one in #3  
  
#Now let's try a KNN removing the numeric predictors that weren't significant  
knn2.heart=lm(HeartDisease~Cholesterol + MaxHR + Oldpeak, data = heart.train)  
train.x=heart.train[,c("Cholesterol", "MaxHR", "Oldpeak")]  
test.x=heart.test[,c("Cholesterol", "MaxHR", "Oldpeak")]  
train.y=heart.train[,"HeartDisease"]  
knn.pred=knn(train.x, test.x, train.y, k=11)  
table(knn.pred, heart.test$HeartDisease)

##   
## knn.pred 0 1  
## 0 78 44  
## 1 41 112

mean(knn.pred != heart.test$HeartDisease)

## [1] 0.3090909

#Not much of a change from #3  
  
  
#Let's add an interaction (sex and Cholesterol) to the previous model  
  
log.reg2=glm(HeartDisease~Sex+ChestPainType+Cholesterol+FastingBS+ExerciseAngina+Oldpeak+ST\_Slope +Sex\*Cholesterol, data=heart.train, family=binomial)  
  
summary(log.reg2)

##   
## Call:  
## glm(formula = HeartDisease ~ Sex + ChestPainType + Cholesterol +   
## FastingBS + ExerciseAngina + Oldpeak + ST\_Slope + Sex \* Cholesterol,   
## family = binomial, data = heart.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7361 -0.3841 0.1562 0.4575 2.6294   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.964651 0.951902 -1.013 0.310872   
## SexM 2.074071 0.857700 2.418 0.015599 \*   
## ChestPainTypeATA -2.063651 0.373801 -5.521 3.38e-08 \*\*\*  
## ChestPainTypeNAP -1.912264 0.320945 -5.958 2.55e-09 \*\*\*  
## ChestPainTypeTA -1.774068 0.466475 -3.803 0.000143 \*\*\*  
## Cholesterol -0.003776 0.003189 -1.184 0.236413   
## FastingBS 1.507804 0.328760 4.586 4.51e-06 \*\*\*  
## ExerciseAnginaY 0.875882 0.281225 3.115 0.001842 \*\*   
## Oldpeak 0.477737 0.142362 3.356 0.000791 \*\*\*  
## ST\_SlopeFlat 1.459395 0.526166 2.774 0.005543 \*\*   
## ST\_SlopeUp -0.699736 0.552082 -1.267 0.204994   
## SexM:Cholesterol -0.001794 0.003425 -0.524 0.600396   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 885.59 on 642 degrees of freedom  
## Residual deviance: 427.41 on 631 degrees of freedom  
## AIC: 451.41  
##   
## Number of Fisher Scoring iterations: 6

logreg.probs=predict(log.reg2,heart.test)  
logreg.pred= rep(1, length(logreg.pred))  
logreg.pred[logreg.probs < 0.5] = 0  
table(logreg.pred, heart.test$HeartDisease)

##   
## logreg.pred 0 1  
## 0 107 24  
## 1 12 132

mean(logreg.pred != heart.test$HeartDisease)#test error rate

## [1] 0.1309091

mean(logreg.pred == heart.test$HeartDisease)

## [1] 0.8690909

#test error rate up by 1%  
  
qda3=qda(HeartDisease~Sex+ChestPainType+Cholesterol+FastingBS+ExerciseAngina+Oldpeak+ST\_Slope +Sex\*Oldpeak, data = heart.train)  
qda3.pred=predict(qda3,heart.test)$class  
table(qda3.pred, heart.test$HeartDisease)

##   
## qda3.pred 0 1  
## 0 98 18  
## 1 21 138

err.qda=mean(qda3.pred != heart.test$HeartDisease)  
err.qda#test error rate

## [1] 0.1418182

1-err.qda# %correct predictions

## [1] 0.8581818