Week 5. Linear discriminant analysis and ROC curve for logistic regression

Linear discriminant analysis (LDA) and its application to identification of high-risk heart attack patients, misclassification error. Linear and nonlinear classification via logistic regression. Connection between the ROC curve, logistic regression, LDA. Optimal bone measurement predictor for identification of female in `Goldman.csv` data set.

R codes: `mah, parsROC`

Data set: `Goldman.csv`

Linear discriminant analysis

Theorem 5.4

Generalization of the total error minimization using the ROC curve. ROC curve and previously solution for the optimal threshold work for one dimension/variable/predictor ($m = 1$) such as blood pressure to identify normal patient of family income to identify mortgage defaulter. LDA solves this problem when m is any ($m > 1$).

Supervised classification of two multivariate normal distributions: find a plane that optimally separates the two multivariate normal distributions

$K = 2$ under equal covariance matrix assumption: $\mathbf{y}^{m\times 1}{\sim}N(\boldsymbol{\mu}_k, \boldsymbol{\Omega})$, $k = 1, 2 = K$. *Normal distribution assumption is required.*

Problem 1 *There are two Gaussian populations (the $m$dimensional vectors) mixed up ($m \geq 1$). Develop an optimal linear discrimination rule (function) to decide what population an observation belongs to. We know what cluster each point belongs to (supervised learning).*

Note: LDA does not tell "who is who" like who is defaulter and who's not defaulter but simply says that observation belongs to generic cluster/group 1 or 2.

How to solve this problem when $m = 1$?

Theorem 2 *The optimal linear discrimination rule is as follows: points* $\mathbf{y}$ *belong to cluster 1 if*

$$(\mathbf{y} - \boldsymbol{\mu})'\mathbf{a} > 0.$$

Otherwise, points belong to cluster 2 (linear discrimination rule), where

$$\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2), \quad \mathbf{a} = \boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{1}$$

The discrimination plane is defined as

$$P = \left\{\mathbf{z} \in R^m : (\mathbf{z} - \boldsymbol{\mu})'\mathbf{a} = 0\right\} = \left\{\mathbf{z} \in R^m : (\mathbf{z} - \boldsymbol{\mu}) \perp \mathbf{a} = 0\right\}.$$

Case 3 *(a) $m = 1$. The discrimination point is $\mu = (\mu_1 + \mu_2)/2$ where the two densities intersect. (b) $m = 2$ and $\mathbf{\Omega} = \sigma^2\mathbf{I}^{2\times 2}$. Contours of the bivariate normal pdfs are circles with centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Vector* **a** *is prallel to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. The discrimination line is orthogonal to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.*

Theorem 4 *The classification rule (1) minimizes the total misclassification error.*

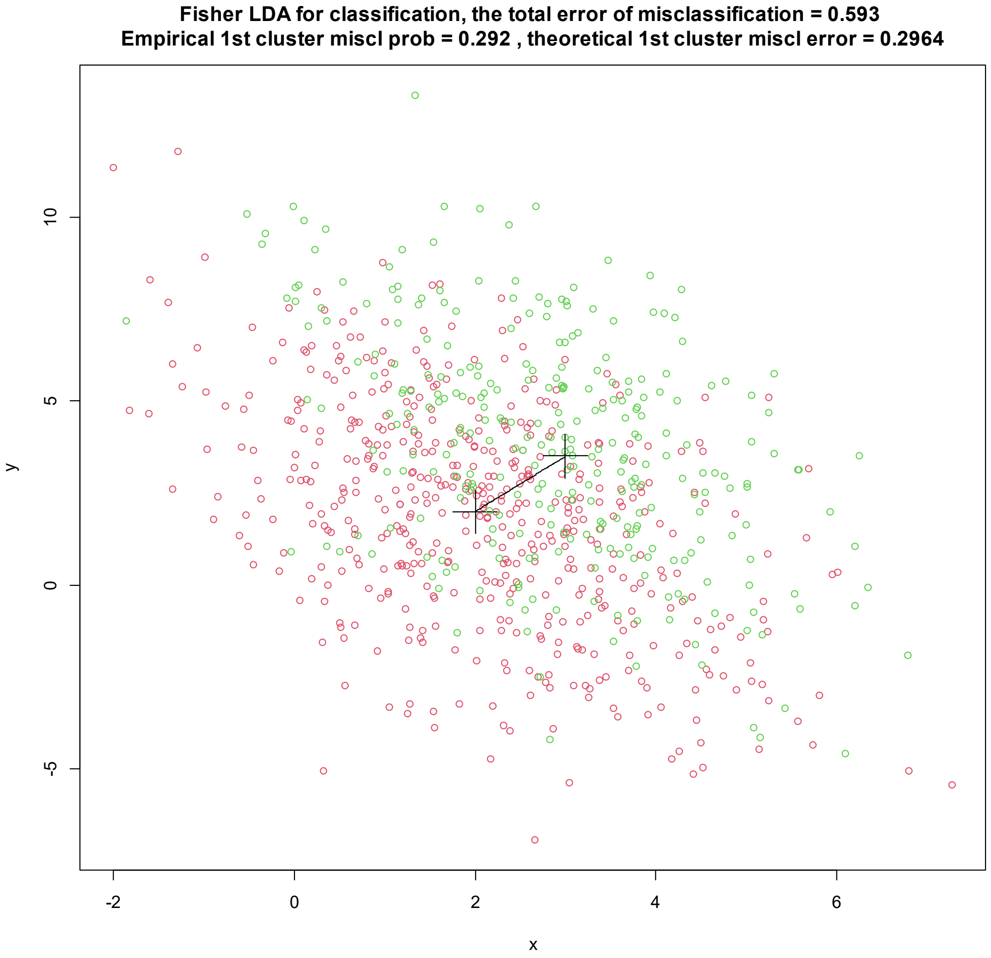
Derive the optimal discrimination rule for $m = 1$ and $m = 2$ (start with $\mathbf{\Omega} = \mathbf{I}$).
How to find the common covariance matrix

$$\widehat{\mathbf{\Omega}} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\text{var}(\mathbf{X}_1) + (n_2 - 1)\text{var}(\mathbf{X}_2)]$$

where matrix $\mathbf{X}_1^{n_1\times m}$ contains data from the first distribution and matrix $\mathbf{X}_2^{n_2\times m}$ contains data from the second distribution. In addition,

$$\widehat{\boldsymbol{\mu}}_1 = \overline{\mathbf{x}}_1, \quad \widehat{\boldsymbol{\mu}}_2 = \overline{\mathbf{x}}_2.$$

See R function `mah`



Theoretical probability of misclassification

Misclassification: assign $\mathbf{y}$ to cluster 1, i.e. apply the rule $(\mathbf{y}-\boldsymbol{\mu})'\mathbf{a}>0$ but in fact $\mathbf{y}$ belongs to cluster 2. The probability is found as

$$\Pr((\mathbf{y}-\boldsymbol{\mu})'\mathbf{a}>0)$$

under condition

$$\mathbf{y}\sim N(\boldsymbol{\mu}_2,\boldsymbol{\Omega}).$$

But

$$(\mathbf{y}-\boldsymbol{\mu})'\mathbf{a}\sim N\left(-\frac{1}{2}\delta^2,\delta^2\right),$$

where

$$\delta^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

and therefore the probability of misclassification of points from cluster 1 to cluster 2 is given by

$$\Phi\left(-\frac{1}{2}\delta\right).$$

The same probability of misclassification of points from cluster 2 to cluster 1 is the same, $\Phi\left(-\frac{1}{2}\delta\right)$. Thus the total misclassification probability is

$$2\Phi\left(-\frac{1}{2}\delta\right).$$

Definition 5 *The Mahalanobis distance between normal populations is defined as*

$$\delta=\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}.$$

and the rule

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Omega}^{-1}\left[\mathbf{y}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right]>0$$

is called the Mahalanobis (linear) discrimination rule.

ROC for logistic regression

Code classes as $Y=0$ (control) and $Y=1$ (cases). We assume that

$$\Pr(Y_i=1|\mathbf{x}_i)=\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)},i=1,2,...,n$$

or using logit notation

$$\text{logit}(Y_i)=\boldsymbol{\beta}'\mathbf{x}_i.$$

An obvious rule: a subject with the set of predictors $\mathbf{u}$ is control if

$$\Pr(Y=1|\mathbf{u})<\frac{1}{2}$$

or if

$$L_i=\widehat{\boldsymbol{\beta}}'x_i<0.$$

The threshold probability may be not necessarily 0.5. Then the rule is:

$$\text{if } L_i<c \text{ then } Y=0 \text{ otherwise } Y=1,$$

where c is the threshold.

We call

$$L_i=\widehat{\boldsymbol{\beta}}'\mathbf{x}_i$$

a linear predictor and

$$\frac{\exp(\widehat{\boldsymbol{\beta}}'\mathbf{x}_i)}{1+\exp(\widehat{\boldsymbol{\beta}}'\mathbf{x}_i)}$$

fitted values.

To compute the ROC curve and AUC we run c from $-\infty$ to ∞ and compute sensitivity and specificity. For two predictors and intercept term we have

$$\beta_1+\beta_2x_1+\beta_3x_2=c$$

and

$$x_2=\frac{c-\beta_1-\beta_2x_1}{\beta_3}.$$

We can use total misclassifcation error as the criterion for choosing optimal c.

LDA versus logistic regression: pros and cons

1. Logistic regression does not require multivariate normal distribution.

2. LDA minimizes the theoretical total misclassification error, but logistic regression is estimated by ML (parameters are of interest).

3. Nonlinear classification is easy to apply using logistic regression by introducing variables, e.g. quadratic variables .

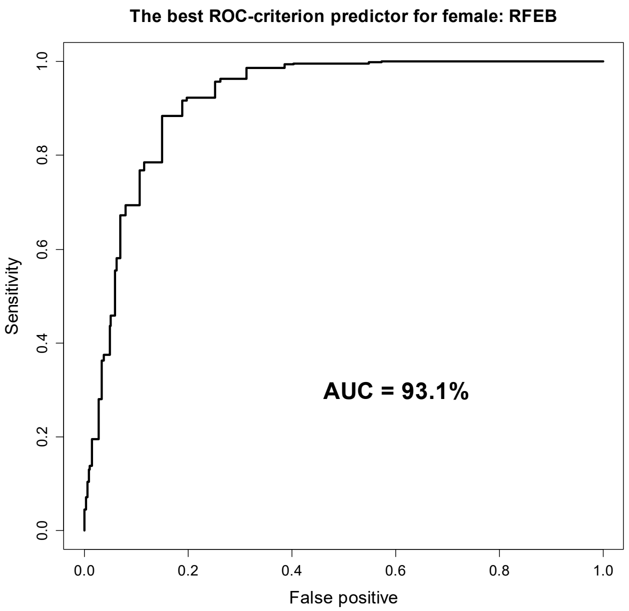
4. Logistic regression allows building a parsimonious model through regression coefficient testing (p-value).

5. Both LA and LDA can use AUC for building parsimonious model by forward variable selection (see below).

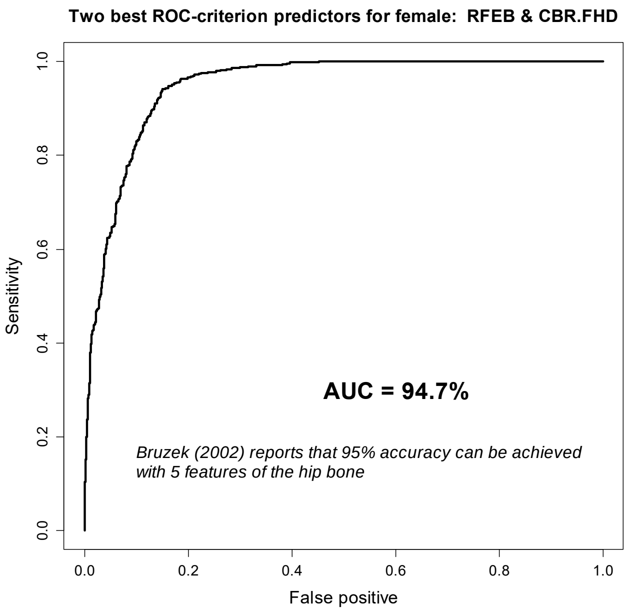
The ROC-criterion for finding the best parsimonious logistic regression: application to female classification in Goldman set

What human bone is the predictor for sex? The answer: the width of the right knee.

See R function `parsROC`

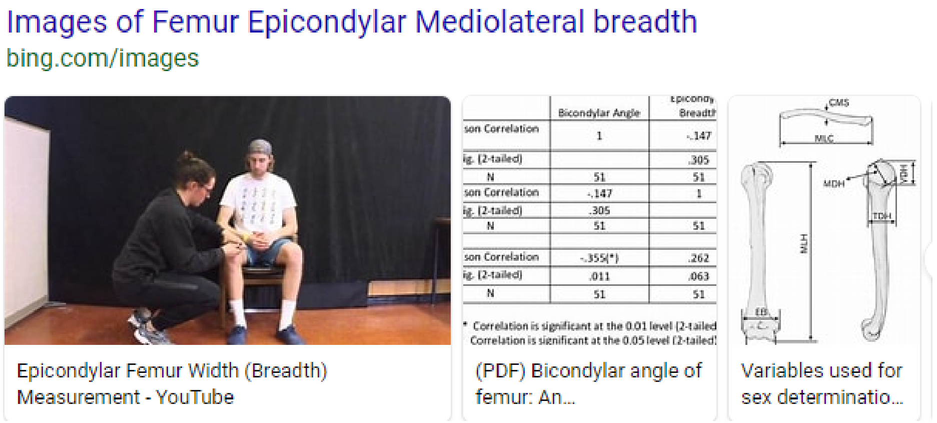

The best ROC-criterion predictor for female: RFEB

AUC = 93.1%


Two best ROC-criterion predictors for female: RFEB & CBR.FHD

AUC = 94.7%

Bruzek (2002) reports that 95% accuracy can be achieved with 5 features of the hip bone


Images of Femur Epicondylar Mediolateral breadth

bing.com/images

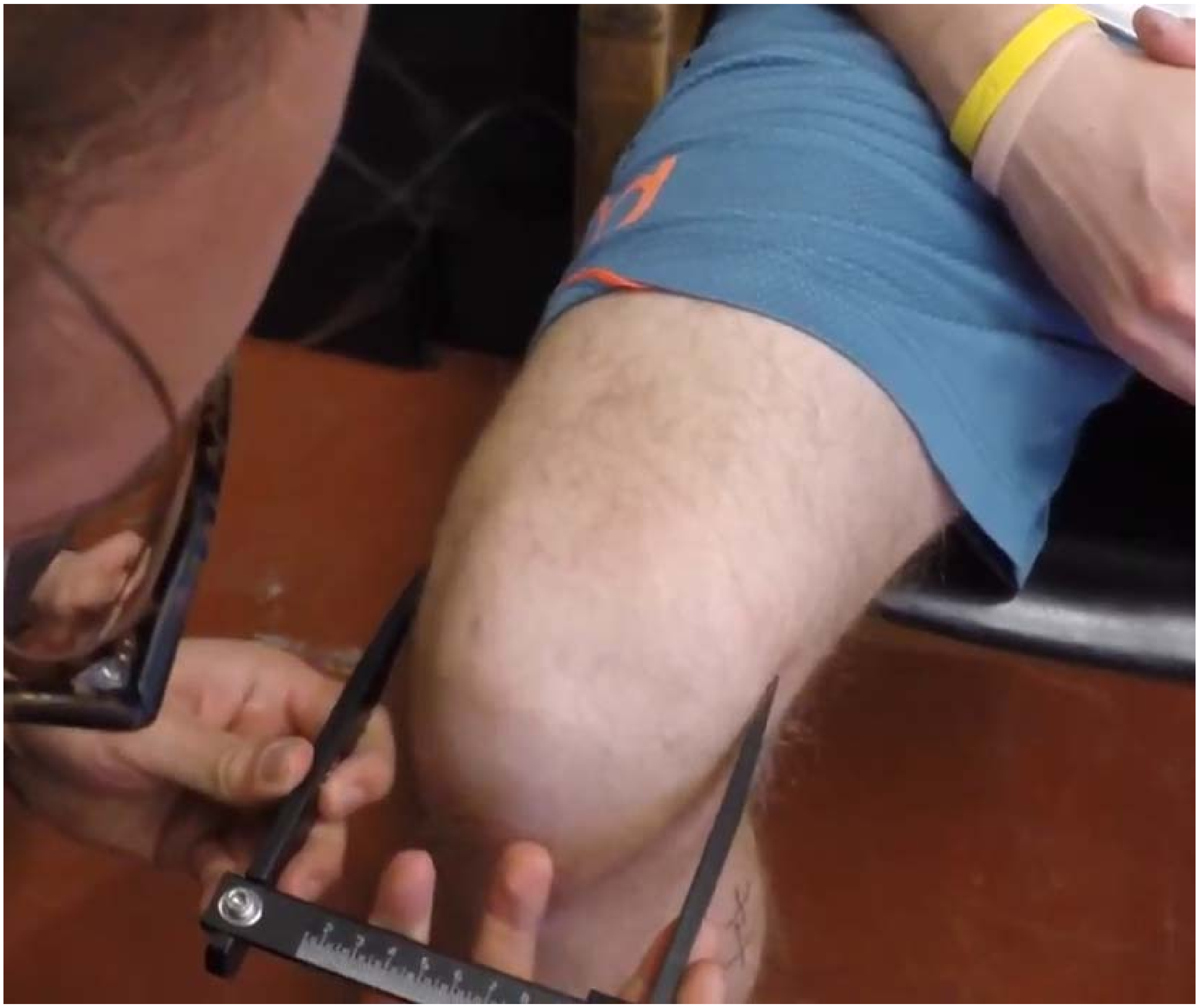
Epicondylar Femur Width (Breadth) Measurement - YouTube

	Bicondylar Angle	Epicondy Breadth
son Correlation	1	-.147
ig. (2-tailed)		.305
N	51	51
son Correlation	-.147	1
ig. (2-tailed)	.305	
N	51	51
son Correlation	-.355(*)	.262
ig. (2-tailed)	.011	.063
N	51	51

* Correlation is significant at the 0.01 level (2-tailed
Correlation is significant at the 0.05 level (2-tailed)

(PDF) Bicondylar angle of femur: An...

Variables used for sex determinatio...



x=RFEB

```
Call:
glm(formula = y ˜x, family = binomial)
Deviance Residuals:
 Min 1Q Median 3Q Max
-3.2634 -0.4493 -0.1688 0.5077 2.5199
Coefficients:
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) 32.67795 1.77624 18.40 <2e-16 ***
x -0.44485 0.02398 -18.55 <2e-16 ***
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
 Null deviance:  1796.36 on 1376 degrees of freedom
Residual deviance:  938.97 on 1375 degrees of freedom
AIC: 942.97
Number of Fisher Scoring iterations:  6
```

x1=RFEB, x2=CBR.FHD

```
Call:
glm(formula = y ~x1i + x2i, family = binomial)
Deviance Residuals:
 Min 1Q Median 3Q Max
-3.4975 -0.3600 -0.1048 0.3502 2.4889
Coefficients:
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) 42.28457 2.30240 18.36 <2e-16 ***
x1i -0.78242 0.04450 -17.58 <2e-16 ***
x2i 0.26482 0.02364 11.20 <2e-16 ***
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
 Null deviance:  1795.47 on 1375 degrees of freedom
Residual deviance:  774.81 on 1373 degrees of freedom
AIC: 780.81
Number of Fisher Scoring iterations:  6
```

Homework

This homework relies of R functions `mah` and `parsROC`. Presentation matters.

Use the following clean version of the data set

```
d=read.csv("c:\\QBS124\\Goldman.csv",stringsAsFactors=F)
nm=names(d[18:(ncol(d)-9)])
sex=as.numeric(as.vector(d[,3]))
d=as.matrix(d[,18:(ncol(d)-9)])
d=d[!is.na(sex),]
sex=sex[!is.na(sex)]
nr=nrow(d);nc=ncol(d)
```

1. (15 points). Apply the LDA to construct the rule for identification of female. (a) Using the double for loop over all predictors find the best two bone predictors that minimize the total theoretical misclassification error. (b) Display the points and the classification line (use red color for female and grean for male). (c) Compute and display the minimum theoretical and empirical misclassification error.

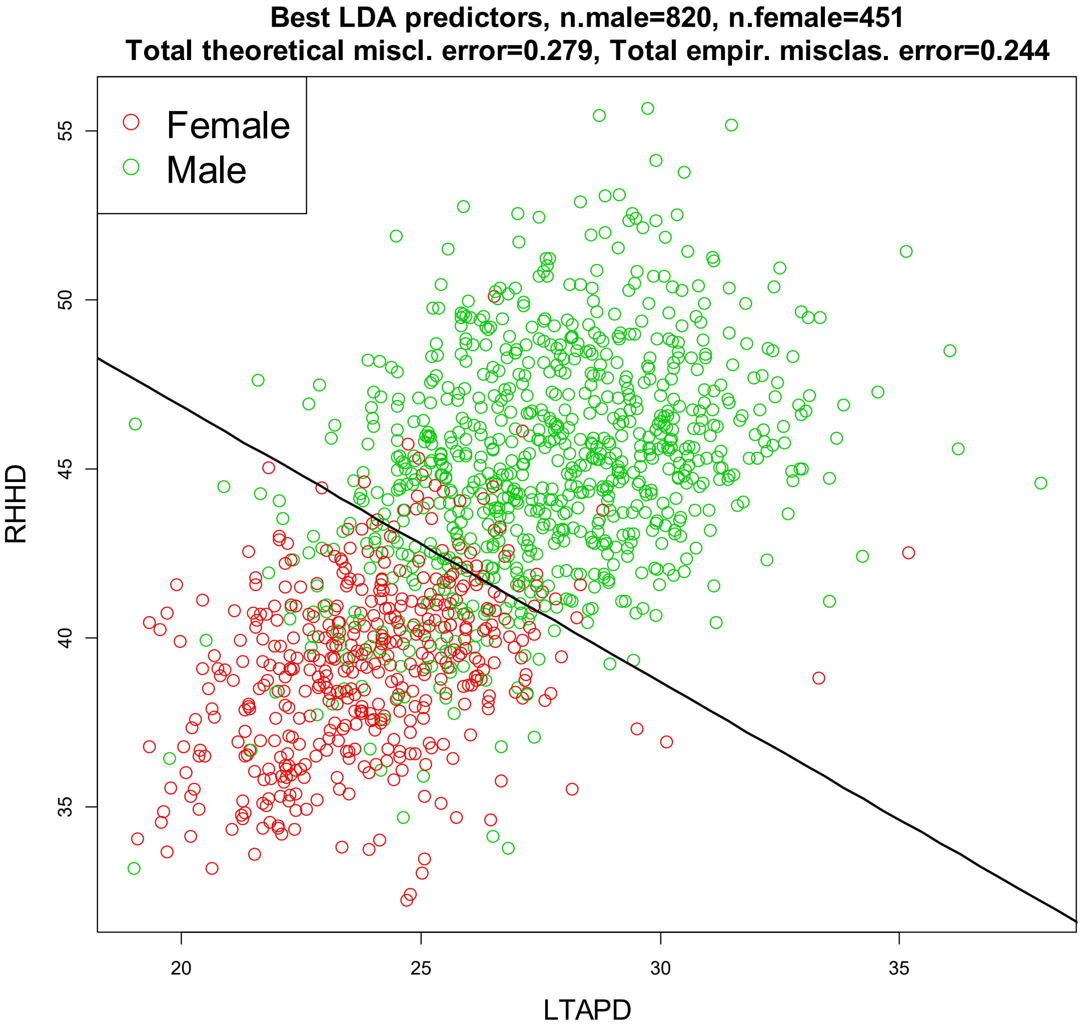
2. (15 points). (a) Add to bone measurements quadratic and cross-product terms and repeat the analysis for the best predictor using logistic regression for identification of female. (b) Display the resulted ROC curve, AUC and optimal threshold that minimizes the total misclassification error (display the value, the point on the curve and two segments parallel to sensitivity and false positive).

Solutions

1. See function `femLDA`

```
 femLDA=function()
{
dump("femLDA","c:\\QBS124\\femLDA.r")
#Application of LDA to identification of female in Goldman.csv file
t0=Sys.time()
d=read.csv("c:\\QBS124\\Goldman.csv",stringsAsFactors=F)
nm=names(d[18:(ncol(d)-9)])
sex=as.numeric(as.vector(d[,3]))
d=as.matrix(d[,18:(ncol(d)-9)])
d=d[!is.na(sex),]
sex=sex[!is.na(sex)]
nr=nrow(d);nc=ncol(d)
#Double loop for search of the two best predictors that maximize delta2
delta2.max=0
for(i1 in 1:nc)
for(i2 in 1:nc)
if(i1>i2)
{
 x1=d[,i1];x2=d[,i2]
 ina=is.na(x1) | is.na(x2)
 x1=x1[!ina];x2=x2[!ina]
 sexi=sex[!ina]

 X=cbind(x1,x2)
 X1=X[sexi==0,];X2=X[sexi==1,]
 n1=nrow(X1);n2=nrow(X2)
 mu1=colMeans(X1);mu2=colMeans(X2)
 Omega=1/(n1+n2-2)*(var(X1)*(n1-1)+var(X2)*(n2-1))
 iOmega=solve(Omega)
 delta2=t(mu1-mu2)%*%iOmega%*%(mu1-mu2)
 if(delta2>delta2.max)
 {
 delta2.max=delta2
 i1.best=i1
 i2.best=i2
 }
}
totmisl=2*pnorm(-.5*sqrt(delta2))
print(c(i1.best,i2.best,delta2.max,totmisl))
x1=d[,i1.best];x2=d[,i2.best]
ina=is.na(x1) | is.na(x2)
x1=x1[!ina];x2=x2[!ina]
sexi=sex[!ina]
X=cbind(x1,x2)
X1=X[sexi==0,];X2=X[sexi==1,]
n1=nrow(X1);n2=nrow(X2)
mu1=colMeans(X1);mu2=colMeans(X2)
mu=0.5*(mu1+mu2)
Omega=1/(n1+n2-2)*(var(X1)*(n1-1)+var(X2)*(n2-1))
iOmega=solve(Omega)
a=iOmega%*%(mu1-mu2)
par(mfrow=c(1,1),mar=c(4.5,4.5,4,1),cex.lab=1.5,cex.main=1.5)
plot(x1,x2,col=3-sexi,cex=1.5,xlab=nm[i1.best],ylab=nm[i2.best])
#Total theoretical misclass error
thms.er=round(2*pnorm(-sqrt(delta2.max)/2),3)
#Total empirical misclass error
classR=as.vector((X-rep(1,n1+n2)%*%t(mu))%*%a)
#print(sex)
#return(classR)
empms.er=round(sum(classR<0 & sexi==0)/n1+sum(classR>0 & sexi==1)/n2,3)
title(paste("Best LDA predictors, n.male=",n1,", n.female=",n2,"\nTotal theoretical miscl. error=",thms.er,
          ", Total empir. misclas. error=",empms.er, sep=""))
legend("topleft",c("Female","Male"),col=2:3,pch=1,cex=2)
y1=seq(from=10,to=50,length=100)
y2=mu[2]-(y1-mu[1])*a[1]/a[2]
lines(y1,y2,lwd=2)
print(Sys.time()-t0)
}
```

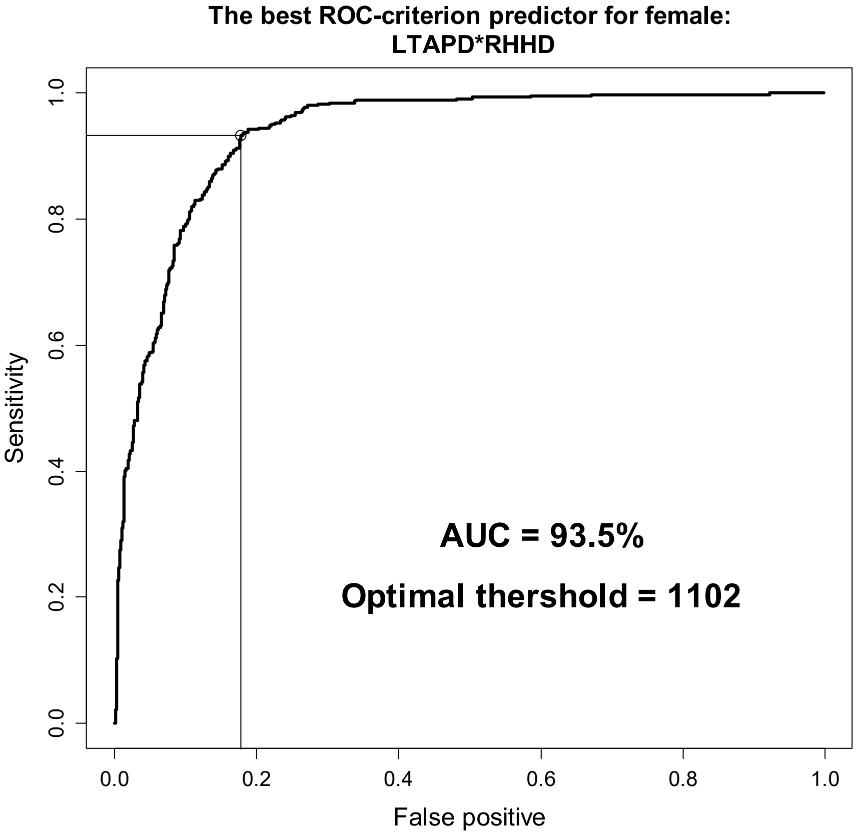


2. See function `parsROC_nl`

```
 parsROC_nl=function()
{
dump("parsROC_nl","c:\\QBS124\\parsROC_nl.r")
#Application of the ROC-criterion for finding a parsimonious logistic regression
t0=Sys.time()
d=read.csv("c:\\QBS124\\Goldman.csv",stringsAsFactors=F)
nm=names(d[18:(ncol(d)-9)])
sex=as.numeric(as.vector(d[,3]))
d=as.matrix(d[,18:(ncol(d)-9)])
nc=ncol(d);nr=nrow(d)
#for(i in 1:nc)
#if(sum(is.na(d[,i]))==nr) {alln=i;break}
#d=d[,-c(alln,5,50,51)]
#nm=nm[-c(alln,5,50,51)]
d=d[!is.na(sex),]
sex=sex[!is.na(sex)]
```

```
nr=nrow(d);nc=ncol(d)
#Expanding d to attach squared and cross-product terms
ntot.ad=nc*(nc+1)/2
d2cross=matrix(ncol=ntot.ad,nrow=nr)
nmd2=rep("",ntot.ad)
k=0
for(i in 1:nc)
for(j in 1:nc)
if(i>=j)
{
 k=k+1
 d2cross[,k]=d[,i]*d[,j]
 nmd2[k]=paste(nm[i],"*",nm[j],sep="")
}
d.new=cbind(d,d2cross)
nm.new=c(nm,nmd2)
AUC=rep(0,ntot.ad)
for(ivar in 1:ntot.ad)
{
 x=d.new[,ivar]
 y=sex[!is.na(x)]
 x=x[!is.na(x)]
 ni=length(x)
 n0=sum(1-y);n1=sum(y)
 o=glm(y~x,family=binomial)
 sod=sort(x)
 fp0=0
 for(i in 1:ni)
  {
 sens=sum(x<sod[i]&y==1)/n1
 fp=sum(x<sod[i]&y==0)/n0
 if(i>1) AUC[ivar]=AUC[ivar]+sens*(fp-fp0)
 fp0=fp
  }
}
i=1:ntot.ad
ibest=i[AUC==max(AUC)]
v1.best=d.new[,ibest];nm.best=nm.new[ibest]
y=sex[!is.na(v1.best)]
x=v1.best[!is.na(v1.best)]
n0=sum(1-y);n1=sum(y)
ni=length(x)
o=glm(y~x,family=binomial)
print(paste("Best single predictor:",nm.best))
print(summary(o))
sod=sort(x)
```

```
AUC.best=fp0=0
sens=fp=toter=rep(0,ni)
for(i in 1:ni)
{
 sens[i]=sum(x<sod[i]&y==1)/n1
 fp[i]=sum(x<sod[i]&y==0)/n0
 if(i>1) AUC.best=AUC.best+sens[i]*(fp[i]-fp0)
 fp0=fp[i]
 toter[i]=(1-sens[i])+fp[i]
}
par(mfrow=c(1,1),mar=c(4.5,4.5,4,1),cex.lab=1.5,cex.main=1.5,cex.axis=1.25)
plot(fp,sens,type="s",lwd=3,xlab="False positive",ylab="Sensitivity",main=paste("The best ROC-criterion predictor for female:\n",nm.best))
opthresh=sod[toter==min(toter)]
optsens=sens[toter==min(toter)]
optfp=fp[toter==min(toter)]
points(optfp,optsens,cex=1.5)
lines(x=c(-1,optfp,optfp),y=c(optsens,optsens,-1))
text(.6,.3,paste("AUC = ",round(AUC.best*100,1),"%",sep=""),cex=2,font=2)
text(.6,.2,paste("Optimal threshold = ",round(opthresh),sep=""),cex=2,font=2)
print(Sys.time()-t0)
}
```


The best ROC-criterion predictor for female:
LTAPD*RHHD
Sensitivity
False positive
AUC = 93.5%
Optimal thershold = 1102