S&DS 563 / F&ES 758b - Multivariate Statistics Homework #4 Cluster Analysis

Lanxin Jiang (lj345), Grace Sun (ys544), Chenglin Lu (cl939)

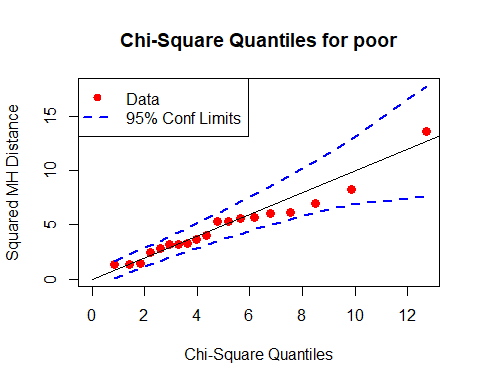
2018-03-08

1. Evaluate the assumptions implicit to Discriminant Analysis for your data - multivariate normality WITHIN each group (i.e. chi-square quantile plots) and similarity of covariances matrices (look at Box’s M or just look at raw standard deviations/covariance matrices). Comment on what you find. Comment on whether you think transformations might help your data to meet the assumptions of DA. If you think they might, make some transformations and find out! You may also want to make a matrix plot (or a pairs plot) to get a sense of what your data looks like two variables at a time (use different symbols for each group).

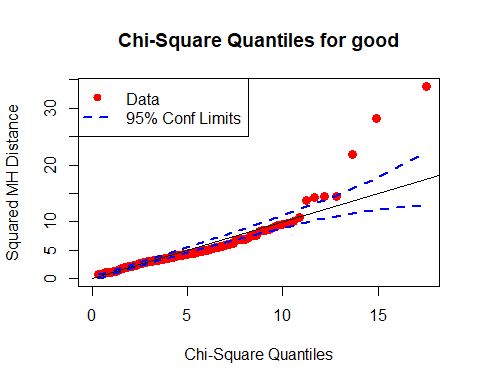
* ## Loading required package: rpanel
* ## Loading required package: tcltk
* ## Package `rpanel', version 1.1-3: type help(rpanel) for summary information
* ## Loading required package: tkrplot
* ## Loading required package: lattice
* ## Loading required package: SpatialEpi
* ## Loading required package: sp
* ## ---  
  ## biotools version 3.1
* ##
* Sys.setenv(JAVA\_HOME="C:\\Program Files\\Java\\jre-9.0.4\\")  
  library(rJava)  
  library(RWeka)  
  CKD <- read.arff("C:/Users/lanxin/Documents/GitHub/Chronic\_Kidney\_Disease/Chronic\_Kidney\_Disease/chronic\_kidney\_disease\_full.arff")  
  CKD.cc<-CKD[complete.cases(CKD),]

use binary variables: appet use continupus variables: age, bmi

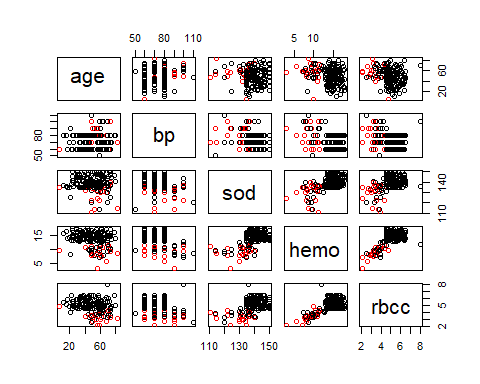
#get online function  
source("http://www.reuningscherer.net/STAT660/R/CSQPlot.r.txt")  
#examine multivariate normality within each group  
 CSQPlot(CKD.cc[CKD.cc$appet=="poor",c(1,2,13,15,18)],label="poor")



CSQPlot(CKD.cc[CKD.cc$appet=="good",c(1,2,13,15,18)],label="good")

 Multinormal assumption is achieved

#make scatter plot to look at differences between groups  
plot(CKD.cc[,c(1,2,13,15,18)],col=as.factor(CKD.cc$appet))



#visually compare sample covariance matrices  
print("Covariance Matrix for patients with good appetite")

## [1] "Covariance Matrix for patients with good appetite"

cov(CKD.cc[CKD.cc$appet=="good",c(1,2,13,15,18)])

## age bp sod hemo rbcc  
## age 233.403086 7.150975 -2.074914 -6.557663 -2.0924565  
## bp 7.150975 112.626421 -13.911480 -6.089928 -1.5822125  
## sod -2.074914 -13.911480 39.211135 6.882901 1.7520384  
## hemo -6.557663 -6.089928 6.882901 5.251040 1.1764170  
## rbcc -2.092456 -1.582212 1.752038 1.176417 0.7483328

print("Covariance Matrix for patients with good appetite")

## [1] "Covariance Matrix for patients with good appetite"

cov(CKD.cc[CKD.cc$appet=="poor",c(1,2,13,15,18)])

## age bp sod hemo rbcc  
## age 248.561404 28.362573 -4.7046784 -4.4347953 -4.3067251  
## bp 28.362573 202.923977 -6.0233918 -7.4239766 -2.5614035  
## sod -4.704678 -6.023392 73.4970760 0.4845029 -0.4257310  
## hemo -4.434795 -7.423977 0.4845029 4.4573099 1.2275146  
## rbcc -4.306725 -2.561404 -0.4257310 1.2275146 0.5224561

#calculate Box's M statistic  
boxM(CKD.cc[,c(1,2,13,15,18)],CKD.cc[,22])

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: CKD.cc[, c(1, 2, 13, 15, 18)]  
## Chi-Sq (approx.) = 23.508, df = 15, p-value = 0.07394

covariance is not different between groups DA assumes -Data from both groups comes from a multivariate normal distribution (test with Q-Q Chi-square) - Covariance matrices in each group are identical (i.e. same shapes, different locations / centroids) However, the second assumption is violated. Thus, scaling might be essenctial. Both assumptions are met

1. Perform stepwise discriminant analysis on your data. Comment on which model seems the best. Use quadratic discriminant analysis if appropriate.

#STEPWISE DA  
step1=stepclass(appet~age+bp+sod+hemo+rbcc, data=CKD.cc, method="lda", direction='both')

## `stepwise classification', using 10-fold cross-validated correctness rate of method lda'.

## 158 observations of 5 variables in 2 classes; direction: both

## stop criterion: improvement less than 5%.

## correctness rate: 0.9125; in: "rbcc"; variables (1): rbcc   
##   
## hr.elapsed min.elapsed sec.elapsed   
## 0.0 0.0 0.3

step2=stepclass(appet~age+bp+sod+hemo+rbcc, data=CKD.cc, method="lda", direction='both', criteria="AS")

## `stepwise classification', using 10-fold cross-validated correctness rate of method lda'.

## 158 observations of 5 variables in 2 classes; direction: both

## stop criterion: improvement less than 5%.

## correctness rate: 0.91208; in: "rbcc"; variables (1): rbcc   
##   
## hr.elapsed min.elapsed sec.elapsed   
## 0.00 0.00 0.32

#Do stepwise quadratic DA  
step3=stepclass(appet~age+bp+sod+hemo+rbcc, data=CKD.cc, method="qda", direction='both', criteria="AS")

## `stepwise classification', using 10-fold cross-validated correctness rate of method qda'.

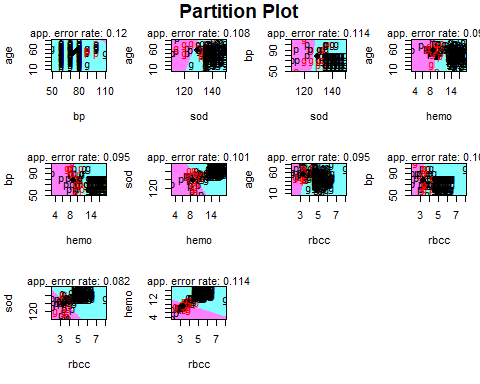
## 158 observations of 5 variables in 2 classes; direction: both

## stop criterion: improvement less than 5%.

## correctness rate: 0.89875; in: "hemo"; variables (1): hemo   
##   
## hr.elapsed min.elapsed sec.elapsed   
## 0.00 0.00 0.24

The above stepwise selection shows that the final model should be “appet~rbcc”

#plot results in space spanned by choosen variables  
partimat(appet~age+bp+sod+hemo+rbcc,data=CKD.cc,method="lda")



1. Comment on whether there is statistical evidence that the multivariate group means are different. could use manova to comapare multivarite group means

manova\_ckd <- manova(cbind(age,bp,sod,hemo,rbcc) ~ appet, data = CKD.cc)  
summary.manova(manova\_ckd,test="Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## appet 1 0.5732 22.636 5 152 < 2.2e-16 \*\*\*  
## Residuals 156   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary.aov(manova\_ckd)

## Response age :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## appet 1 1095 1095.14 4.6571 0.03245 \*  
## Residuals 156 36684 235.15   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response bp :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## appet 1 412.5 412.52 3.3526 0.06901 .  
## Residuals 156 19195.1 123.05   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response sod :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## appet 1 2072.3 2072.27 48.006 1.052e-10 \*\*\*  
## Residuals 156 6734.1 43.17   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response hemo :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## appet 1 499.34 499.34 96.781 < 2.2e-16 \*\*\*  
## Residuals 156 804.88 5.16   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response rbcc :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## appet 1 50.465 50.465 69.87 3.283e-14 \*\*\*  
## Residuals 156 112.674 0.722   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

All the variables are significantly different among different appetite.

1. How many discriminant functions are significant? What is the relative discriminating power of each function?

lda\_CKD<- lda(appet~age+bp+sod+hemo+rbcc, data=CKD.cc)  
lda\_CKD

## Call:  
## lda(appet ~ age + bp + sod + hemo + rbcc, data = CKD.cc)  
##   
## Prior probabilities of groups:  
## good poor   
## 0.8797468 0.1202532   
##   
## Group means:  
## age bp sod hemo rbcc  
## good 48.58993 73.45324 140.1871 14.344604 5.100719  
## poor 56.68421 78.42105 129.0526 8.878947 3.363158  
##   
## Coefficients of linear discriminants:  
## LD1  
## age 0.001961479  
## bp -0.008742255  
## sod -0.049548565  
## hemo -0.259654268  
## rbcc -0.398731340

#normalized coefficient  
sum\_coef<-sum(lda\_CKD$scaling^2)  
norm\_coef<-lda\_CKD$scaling/sum\_coef  
scale\_coef<-scale(norm\_coef)

We have only one significant function.

1. Use classification, both regular and leave-one-out to evaluate the discriminating ability of your functions.

pred <- table(CKD.cc$appet, predict(lda\_CKD)$class)  
# total percent correct  
round(sum(diag(prop.table(pred))),2)

## [1] 0.91

91% percentage correct by classification.

#cross validated results  
lda\_CKDCV<- lda(appet~age+bp+sod+hemo+rbcc, data=CKD.cc,CV=TRUE)  
CKD\_CV <-table(CKD.cc$appet, lda\_CKDCV$class)  
# total percent correct  
round(sum(diag(prop.table(CKD\_CV))),2)

## [1] 0.89

only 89% correct by leave-one-out cross validation

1. Provide some evidence as to which of your original variables are the ‘best’ discriminators amongst your groups (look at standardized discriminant coefficients).

We only have one function. It’s hard to compare. We decided to compare by the correction rate of prediction.

lda\_CKD3<-lda\_CKD2<-lda\_CKD  
lda\_CKD2$scaling<-norm\_coef  
lda\_CKD3$scaling<-scale\_coef  
p1 <- predict(lda\_CKD,newdata = CKD.cc[,c(1,2,13,15,18)])  
t1<-table(p1$class, CKD.cc$appet)  
print(paste("The correctness rate for the original variables",round(sum(diag(prop.table(t1))),2)))

## [1] "The correctness rate for the original variables 0.91"

p2 <- predict(lda\_CKD2,newdata = CKD.cc[,c(1,2,13,15,18)])  
t2<-table(p2$class, CKD.cc$appet)  
print(paste("The correctness rate for the normalized coefficients",round(sum(diag(prop.table(t2))),2)))

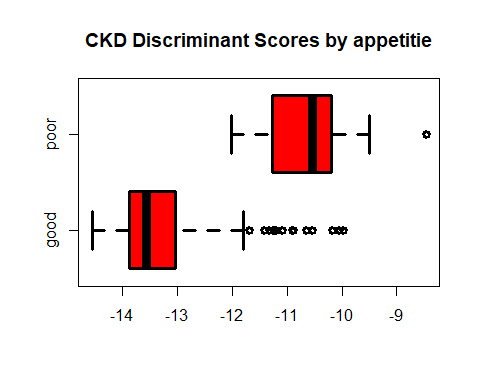
## [1] "The correctness rate for the normalized coefficients 0.89"

p3 <- predict(lda\_CKD3,newdata = CKD.cc[,c(1,2,13,15,18)])  
t3<-table(p3$class, CKD.cc$appet)  
print(paste("The correctness rate for the standardized coefficients",round(sum(diag(prop.table(t3))),2)))

## [1] "The correctness rate for the standardized coefficients 0.61"

1. Make a score plot for the first two or three DA function scores (be sure to use different symbols/colors for each group). Comment on what you see.

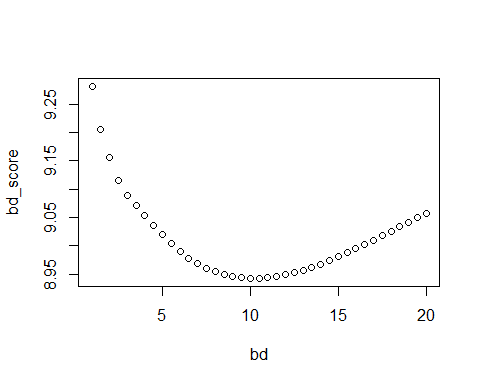
#SCORE PLOTS for linear DA  
ckdlda=lda(CKD.cc[,c(1,2,13,15,18)],grouping=CKD.cc$appet)  
#Calculate scores  
scores=as.matrix(CKD.cc[,c(1,2,13,15,18)])%\*%matrix(ckdlda$scaling,ncol=1)  
#only one discriminant function  
boxplot(scores~CKD.cc$appet,lwd=3, col='red', horizontal=T, main="CKD Discriminant Scores by appetitie")

 The discriminant analysis of one significant function discriminates appetite between variables well.

1. Bonus (and optional)- try kernel smoothing or k-nearest neighbors and get the admiration of your professor and TA (and some extra credit)! You’ll have to use SAS or R for this. kernel smoothing by leave-one-out cross validation to choose h Here I use the Nadaraya-Watson kernel estimator,illustrate by two continous variables: age, sc
2. where the kernel function is Gaussian
3. The leave-one-out cross-validation score can be written as
4. where is the th diagonal element of the smoothing matrix .

locreg<-function(x,y,bandwidth){  
 n<-length(x)  
 L1<-h\_matrix<-matrix(0,ncol=n,nrow=n)  
 for(i in 1:n){  
 L1[i,]<-dnorm((x-x[i])/bandwidth)  
 }  
 for (i in 1:n){  
 for(j in 1:n){  
 h\_matrix[i,j]<-L1[i,j]/sum(L1[,j])  
 }   
 }  
r\_nx<- h\_matrix %\*% y  
score<-1/n\*sum((y-r\_nx)^2/(1-diag(h\_matrix))^2)  
return(list(r\_nx,score,h\_matrix))  
}

#choose bd score  
bd\_score<-bd<-c(seq(1,20,0.5))  
for (i in 1 : length(bd)){  
 bd\_score[i]<-locreg(CKD.cc$age,CKD.cc$sc,bd[i])[[2]]  
}  
plot(bd,bd\_score)



bd\_test<-as.data.frame(cbind(bd,bd\_score))  
#the best bandwidth of the kernel to minimize the risk  
bd\_testorder<-bd\_test[with(bd\_test,order(bd\_score)),]  
#best bandwidth=10  
optimal\_bd<-locreg(CKD.cc$age,CKD.cc$sc,10)  
  
plot(CKD.cc$age,CKD.cc$sc,col="red")  
points(CKD.cc$age,optimal\_bd[[1]])

