Martha Foye, Dave Hiltbrand, and Allen Phelps

Dr. Cuixian Chen

STT 592

20 September 2016

**Section 1: Linear Discriminant Analysis**

From logistic regression, exploration continues to alternative classification methods in search of a potentially more accurate model of professional Posttraumatic Stress Disorder (PTSD) diagnosis, the first of which is Linear Discriminant Analysis; thus, the predictors from the logit model—the Trauma Symptom Inventory Anxious Arousal and Defensive Avoidance t-scores and the interaction of the two clinical scales—were fit to a discriminant function by group in attempt to better predict whether or not each participant belongs among the diagnosed. A model of the full data resulted in the following confusion matrix:

ptsd

lclass FALSE TRUE

FALSE 141 45

TRUE 134 295

The overall, specificity, and sensitivity accuracy rates of 70.89431, 51.27273, and 86.76471 percent, respectively, were approximately identical to those observed as a result of logistic regression with only one additional false positive misclassification.

As before, a training model based on eighty percent of the data was fit to test the remaining twenty percent with the following results:

ptest

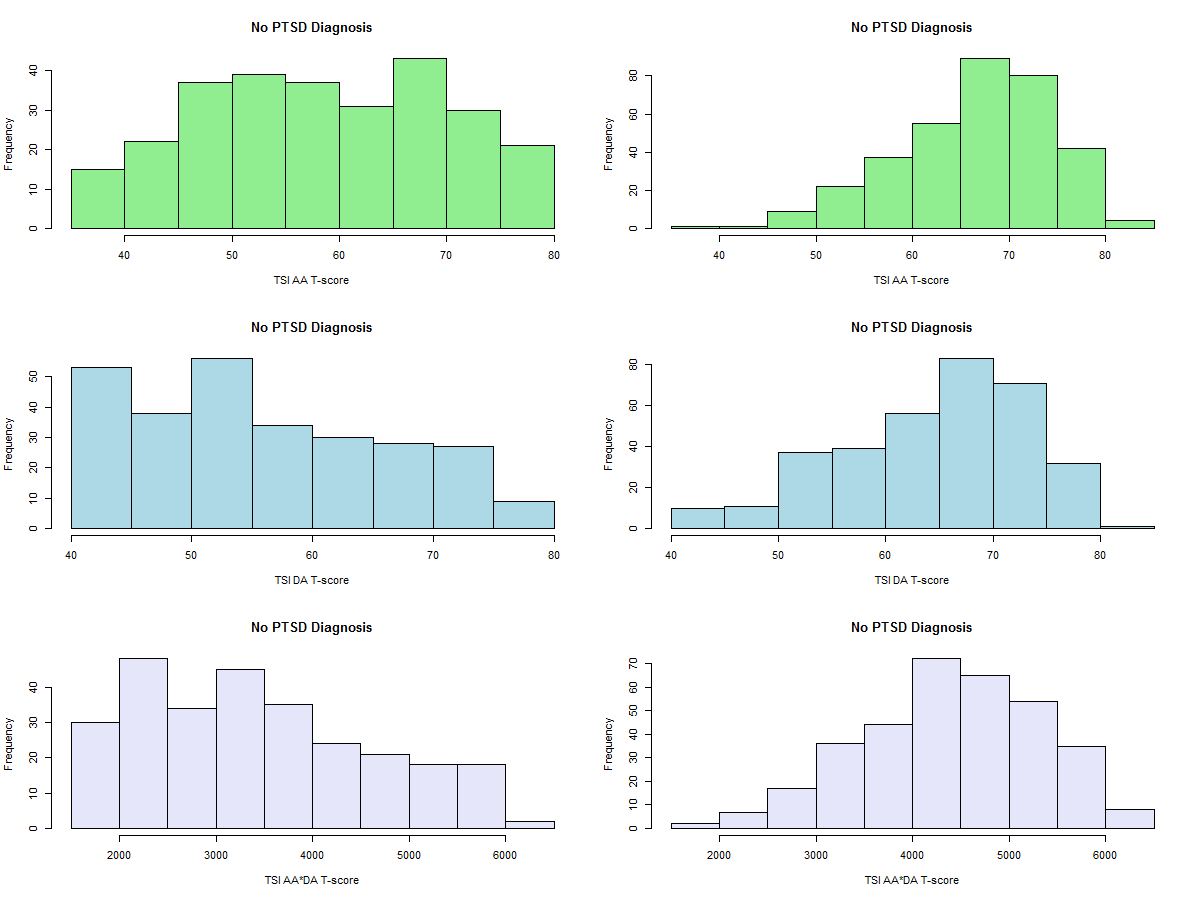
lclass FALSE TRUE

FALSE 27 3

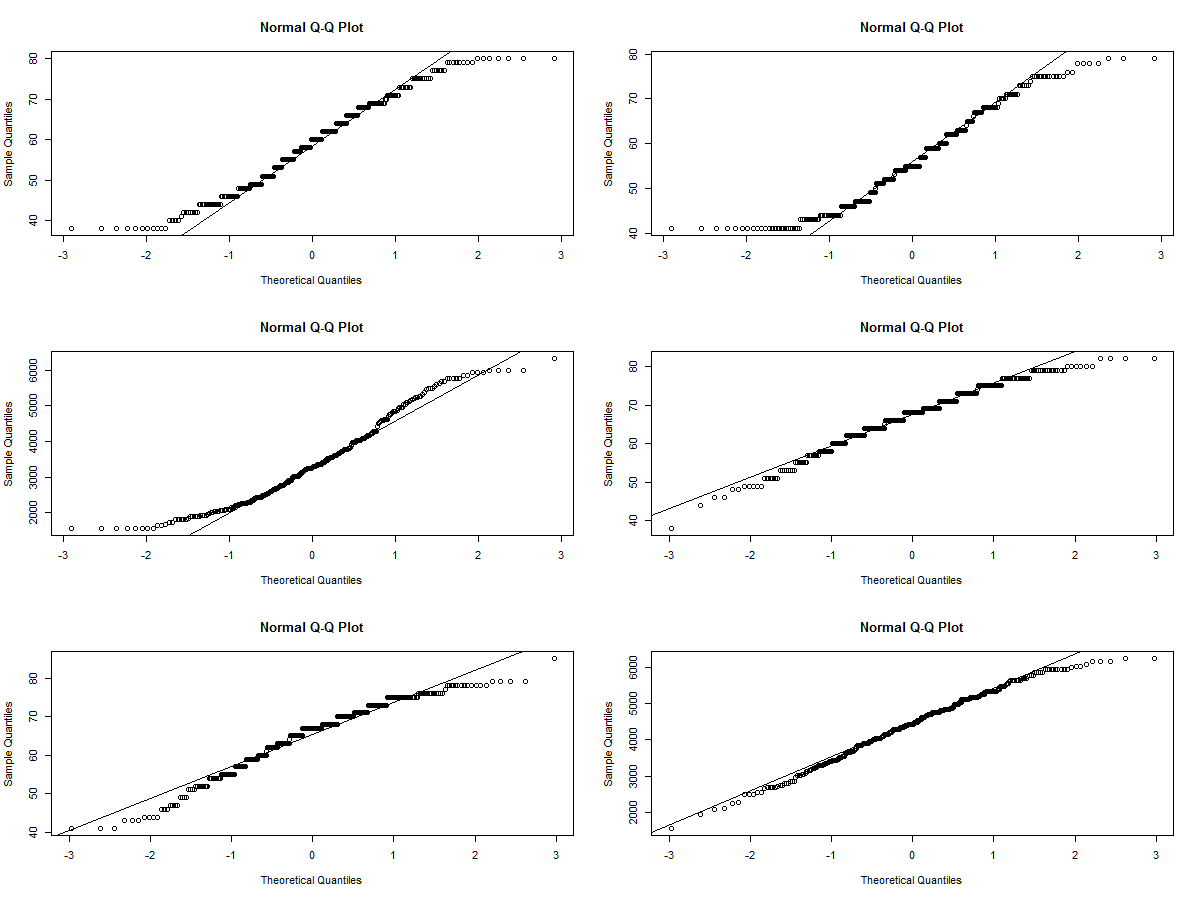
TRUE 32 61

This time, the overall, specificity, and sensitivity accuracy rates of 71.54472, 45.76271, and 95.31250 percent, respectively, were *precisely* identical to the results of logistic regression. The interchangeability between the two methods—in this case—was surprising to say the least; Linear Discriminant Analysis was expected to pale in comparison for its violated assumptions.

Consider the following histograms of each predictor variable by Posttraumatic Stress Disorder diagnosis:



Formal Anderson-Darling tests for normality agree with visual analysis that the predictors are not normally distributed in either group. Indeed, if the three pairs of distributions were overlaid, then the discrimination between groups would appear clearly attributable to their opposing skewness rather than normal distinctiveness. Consider the corresponding normal quantile plots below as an alternative visualization:



A clear s-shaped rather than linear pattern is observed; thus, at least two assumptions of Linear Discriminant Analysis are violated.

**Section 2: Quadratic Discriminant Analysis**

A third violation called for Quadratic Discriminant Analysis (QDA). Formal F tests for homogeneous variance concluded that the predictors do not share the same spread between groups. QDA is intended to remediate the effect of heterogeneous variance by considering a covariance matrix in its modeling of the discriminant function by group; however, the resultant confusion matrix for the full data follows to decreased accuracy:

ptsd

qclass FALSE TRUE

FALSE 109 30

TRUE 166 310

Overall, specificity, and sensitivity rates fell to 68.13008, 39.63636, and 91.17647 percent, respectively. Moreover, the training model based on eighty percent of the data that was fit to test the remaining twenty percent produced even worse results:

ptest

qclass FALSE TRUE

FALSE 21 2

TRUE 38 62

While the overall and sensitivity accuracy rates are fairly consistent between the three classification methods mentioned, the decline is specificity is particularly bothersome for Quadratic Discriminant Analysis (QDA). Testing with the complementary subsets of the data returned only 35.59322 versus the 45.76271 percent specificity rate obtained from Logistic Regression. Even though both rates are undesirable, QDA—in terms of this dataset—grossly misclassified 64.40678% of participants without posttraumatic stress among those who had been professionally diagnosed with Posttraumatic Stress Disorder (PTSD). Overall, Logistic Regression and Linear Discriminant Analysis are unexpectedly preferable and similar in their classification of professional PTSD diagnosis when modeled after the t-scores of the Anxious Arousal and Defensive Avoidance Trauma Symptom Inventory (TSI) scales and their interaction. Although the models are not strong enough to substitute as replacement for professional classification, their strength is moderate enough to support the hypothesis that the TSI is measuring PTSD as it was designed to, which—as concluded previously—further supports a sense of reliability in the dataset.

**Appendix: R Code**

dat=read.csv("stt592dat.csv", header=T)

attach(dat)

names(dat)

dat=data.frame(chartn,ptsd,dat[,19:31])

dat=dat[complete.cases(dat),]

dim(dat)

attach(dat)

names(dat)

install.packages('nortest')

library(MASS)

library(nortest)

#Linear Discriminant Analysis

l=lda(ptsd~aa\*da,data=dat)

lpred=predict(l)

lclass=lpred$class

acctab=table(lclass,ptsd);acctab

accuracy=sum(diag(acctab))/sum(acctab);accuracy

specificity=acctab[1]/sum(acctab[1:2]);specificity

sensitivity=acctab[4]/sum(acctab[3:4]);sensitivity

length(ptsd)\*0.8

dat[493,1]

train=(chartn<31464)

test=dat[!train,]

ptest=ptsd[!train]

ltrain=lda(ptsd~aa\*da,data=dat,subset=train)

lpred=predict(ltrain,test)

lclass=lpred$class

acctab=table(lclass,ptest)

acctab

accuracy=sum(diag(acctab))/sum(acctab);accuracy

specificity=acctab[1]/sum(acctab[1:2]);specificity

sensitivity=acctab[4]/sum(acctab[3:4]);sensitivity

#Quadratic Discriminant Analysis

q=qda(ptsd~aa\*da,data=dat)

qpred=predict(q)

qclass=qpred$class

acctab=table(qclass,ptsd);acctab

accuracy=sum(diag(acctab))/sum(acctab);accuracy

specificity=acctab[1]/sum(acctab[1:2]);specificity

sensitivity=acctab[4]/sum(acctab[3:4]);sensitivity

qtrain=qda(ptsd~aa\*da,data=dat,subset=train)

qpred=predict(qtrain,test)

qclass=qpred$class

acctab=table(qclass,ptest)

acctab

accuracy=sum(diag(acctab))/sum(acctab); accuracy

specificity=acctab[1]/sum(acctab[1:2]);specificity

sensitivity=acctab[4]/sum(acctab[3:4]);sensitivity

#Validating Assumptions

##Normality of Variables

par(mfrow=c(3,2))

aaf=ifelse(ptsd==F,aa,NA);daf=ifelse(ptsd==F,da,NA);aadaf=ifelse(ptsd==F,aa\*da,NA)

aat=ifelse(ptsd==T,aa,NA);dat=ifelse(ptsd==T,da,NA);aadat=ifelse(ptsd==T,aa\*da,NA)

hist(aaf,col='lightgreen',main='No PTSD Diagnosis',xlab='TSI AA T-score');hist(aat,col='lightgreen',main='No PTSD Diagnosis',xlab='TSI AA T-score')

hist(daf,col='lightblue',main='No PTSD Diagnosis',xlab='TSI DA T-score');hist(dat,col='lightblue',main='No PTSD Diagnosis',xlab='TSI DA T-score')

hist(aadaf,col='lavender',main='No PTSD Diagnosis',xlab='TSI AA\*DA T-score');hist(aadat,col='lavender',main='No PTSD Diagnosis',xlab='TSI AA\*DA T-score')

qqnorm(aaf);qqline(aaf);qqnorm(aat);qqline(aat)

qqnorm(daf);qqline(daf);qqnorm(dat);qqline(dat)

qqnorm(aadaf);qqline(aadaf);qqnorm(aadat);qqline(aadat)

dev.off()

ad.test(aaf);ad.test(daf);ad.test(aadaf)

ad.test(aat);ad.test(dat);ad.test(aadat)

##Equal Variance of Variables b/t Groups

var.test(aaf,aat);var.test(daf,dat);var.test(aadaf,aadat)