Kyle Beitz

STAT 4300

10/17/2019

Homework 3

```
> smlm = summary.aov(mlm)

Response GPA :
            Df  Sum Sq Mean Sq F value    Pr(>F)
G            2 12.5015  6.2508  173.31 < 2.2e-16
Residuals   82  2.9576  0.0361

Response GMAT :
            Df Sum Sq Mean Sq F value    Pr(>F)
G            2 258471  129236   35.35 8.492e-12
Residuals   82 299784    3656

> SA = summary(MA)

Multivariate Tests: G
                 Df test stat  approx F num Df den Df     Pr(>F)
Pillai            2  1.009630  41.79734      4    164 < 2.22e-16
Wilks             2  0.126377  73.42569      4    162 < 2.22e-16
Hotelling-Lawley  2  5.836656 116.73312      4    160 < 2.22e-16
Roy               2  5.646045 231.48783      2     82 < 2.22e-16

> C = c(0,-1,1); linearHypothesis(mlm,C)
```

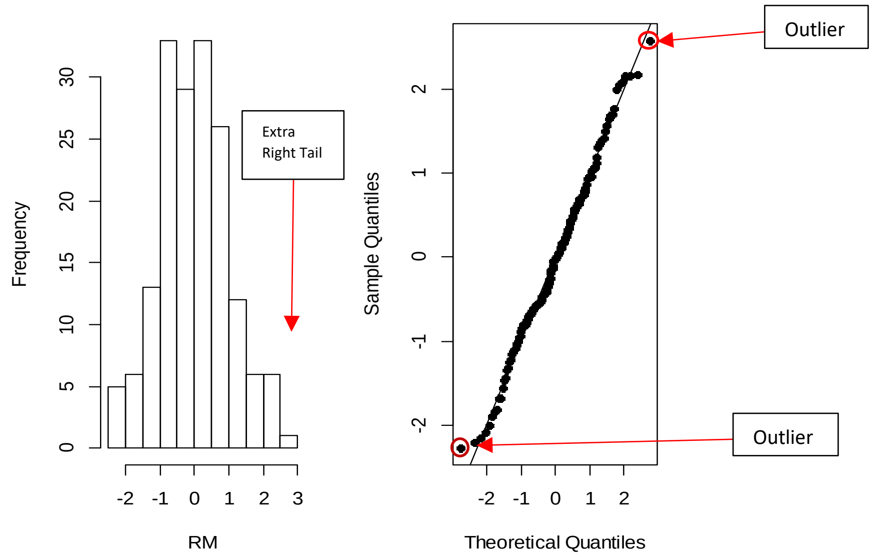
$H_{0,32}$ comparison

```
                   Df test stat approx F num Df den Df     Pr(>F)
Pillai            1  0.848797 227.3514      2     81 < 2.22e-16
Wilks             1  0.151203 227.3514      2     81 < 2.22e-16
Hotelling-Lawley  1  5.613616 227.3514      2     81 < 2.22e-16
Roy               1  5.613616 227.3514      2     81 < 2.22e-16

> shapiro.test(RM) ← Normality

Shapiro-Wilk normality test

      data:  RM
W = 0.99198, p-value = 0.4645
```



Extra Right Tail

Outlier

Outlier

- H.3.1:
  (a)
  - $H_0 : \mu_1 = \mu_2 = \mu_3$, $H_1 : \mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$
  - The wilks test with F(4,162) = 73.43 and a p-value < 0.0001, indicates H_0 should be rejected at the 0.01 level. Thus, at least one of the population means of the 2 variables, GPA and GMAT, are declared to be different across the 3 groups, border, **yes** and **no**.

  (b)
  - Since the p-value < 0.0001 then $H_{0,32}$ is rejected at the 0.01 level. Thus, there is evidence to claim that the population mean of the 2 scores are not the same between these two groups.

  (c)
  - The Cholesky residuals all lie very close to the theoretical normal line except for the smallest and largest Cholesky residual value of -2 and 3 which could be outliers.
  - Since w = 0.99198 and the p-value = 0.4645, fail to reject H_0 at the 0.01 level. Thus, there is not evidence against the assumption that the population distribution of the model error vector is multivariate normal across all observations.

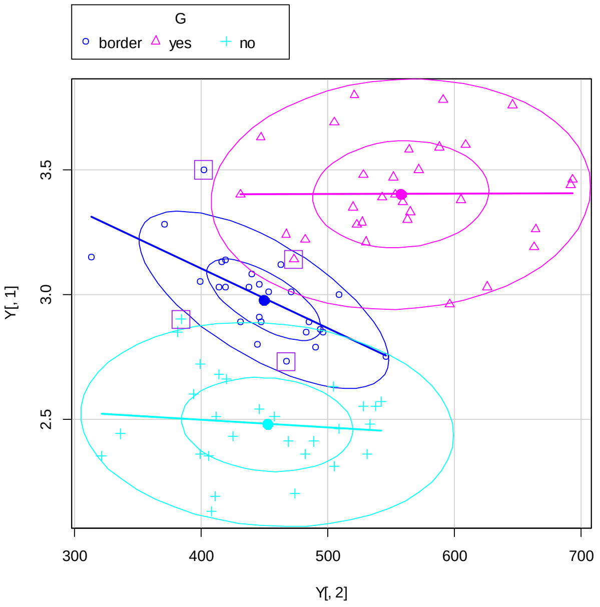
```
> bm.test = boxM(Y,G)

Box's M-test for Homogeneity of Covariance Matrices
data:   Y
Chi-Sq (approx.) = 16.074, df = 6, p-value = 0.01336

> T.class = table(G,G.class)

G.class
G            border      no     yes
Border           24       1       1
yes               1       0      30
no                1      27       0
           correct.rate  error.rate
              0.9529        0.04706

> miss
    GPA GMAT       G G.class
2  3.14  473     yes  border
59 2.90  384      no  border
66 3.50  402 border      yes
75 2.73  467 border       no
```



- H.3.2:

(a)

- H_0 : $\sum_1 = \sum_2 = \sum_3$, H_1 : $\sum_1 \neq \sum_2$ or $\sum_1 \neq \sum_3$ or $\sum_2 \neq \sum_3$
- Since c(6) = 16.01 and the p-value < 0.01, reject H_0 at the 0.01 level. Thus, there is evidence to question the assumption of equal population covariance matrices across the 2 scores.

(b)

- With QDA leave-one-out cross-validation we are nearly perfectly classifying **border**, **yes** and **no** with only 4 misclassifications. **Border** was misclassified twice, once with **yes** and once with **no**. **Yes,** was misclassified one time with **border** and **no** was misclassified once with **border** as well.

(c)

- From the scatterplot above we can determine the following misclassifications:
  - i. A data point in **border** was misclassified with **no** because it has a relatively low GPA score for **border** compared to the other scores in **border** but has a relatively high GMAT score compared to the other scores within **border**. The low GPA caused the data point to be misclassified as **no** instead of **border**.
  - ii. A data point in **yes** was misclassified as **border** because it has a low GPA score and a low GMAT score compared to the other scores within the **yes** group. These low scores caused the misclassification into the **border** group.

**Note: Bold font in answers represents the three groups: "border", "yes" and "no".

R-code

```
> library(car)
> library(MASS)
> library(biotools)

dat = read.csv("C:/Users/Kyle/Desktop/admission.csv")
> Y = as.matrix(dat[,1:2]); n = nrow(Y); p = ncol(Y);
> Gn = dat[,3];
> G = as.factor(Gn);
> levels(G)[3] = "border"
> levels(G)[2] = "no"
> levels(G)[1] = "yes"
> G = relevel(G,ref="border")
> t.G = table(G)
> n1 = t.G[1]; n2 = t.G[2]; n3 = t.G[3]
> ybar = apply(Y,2,mean)
> ybark=by(Y,G,function(x) apply(x,2,mean))
> Sk = by(Y,G,cov)
> mlm = lm(Y~G)
> smlm = summary.aov(mlm)
> MA = Manova(mlm)
> SA = summary(MA)
> E = MA$SSPE
> B = mlm$coefficients
> C = c(0,-1,1);
> bm.test = boxM(Y,G)
> Sp = E/mlm$df.residual
> Uh = mlm$residuals
> Spih = chol(solve(Sp))
> RM = Uh%*%t(Spih)
> par(mfrow=c(1,2))
> hist(RM)
> qqnorm(RM,pch=16,main=NULL);abline(a=0,b=1);
> shapiro.test(RM)
> Tp = ((nk[1]-1)*Sk[[1]]+(nk[2]-1)*Sk[[2]]+(nk[3]-1)*Sk[[3]])
> Sp = Tp/(sum(nk)-3)
> CVS = T
> prior0 = c(1,1,1)/3;
> da2 = qda(G~Y,prior=prior0,CV=CVS)
> if (CVS==F) post.prob = predict(da,as.data.frame(Y))$posterior
> if (CVS==T) post.prob = da$posterior
> if (CVS==F) G.class = predict(da,as.data.frame(Y))$class
> if (CVS==T) G.class = da$class
> t.class = table(G,G.class)
> correct.rate = sum(diag(t.class))/n
> error.rate = 1-sum(diag(t.class))/n
> miss = data.frame(Y,G,G.class)[G!=G.class,]
> scatterplot(Y[,1]~Y[,2]|G,smooth=
> FALSE,ellipse=TRUE,by.groups=TRUE)
> points(miss[,2],miss[,1],pch=22,cex=3,lwd=1.5,col="purple")
> dat = data.frame(Y,G); G.class = NULL
> for(i in 1:n) { # leave-one-out #
> dat.i = dat[i,]; dat.xi = dat[-i,];
> fit = qda(G~.,data=dat.xi,CV=F)
> G.lab = predict(fit,newdata=dat.i,type="class")[[1]]
> G.class = c(G.class,as.character(G.lab)) }
> ct = table(G,G.class)
```