lab2

Maquieira Mariani

29/11/2019

Lab 2 - Multivariate Statistical Methods.

Question 3

We will look at a data set on Egyptian skull measurements (published in 1905 and now in heplots R package as the object Skulls). Here observations are made from five epochs and on each object the maximum breadth (mb), basibregmatic height (bh), basialiveolar length (bl) and nasal height (nh) were measured.

```
library(heplots)
attach(Skulls)
```

a) Explore the data first

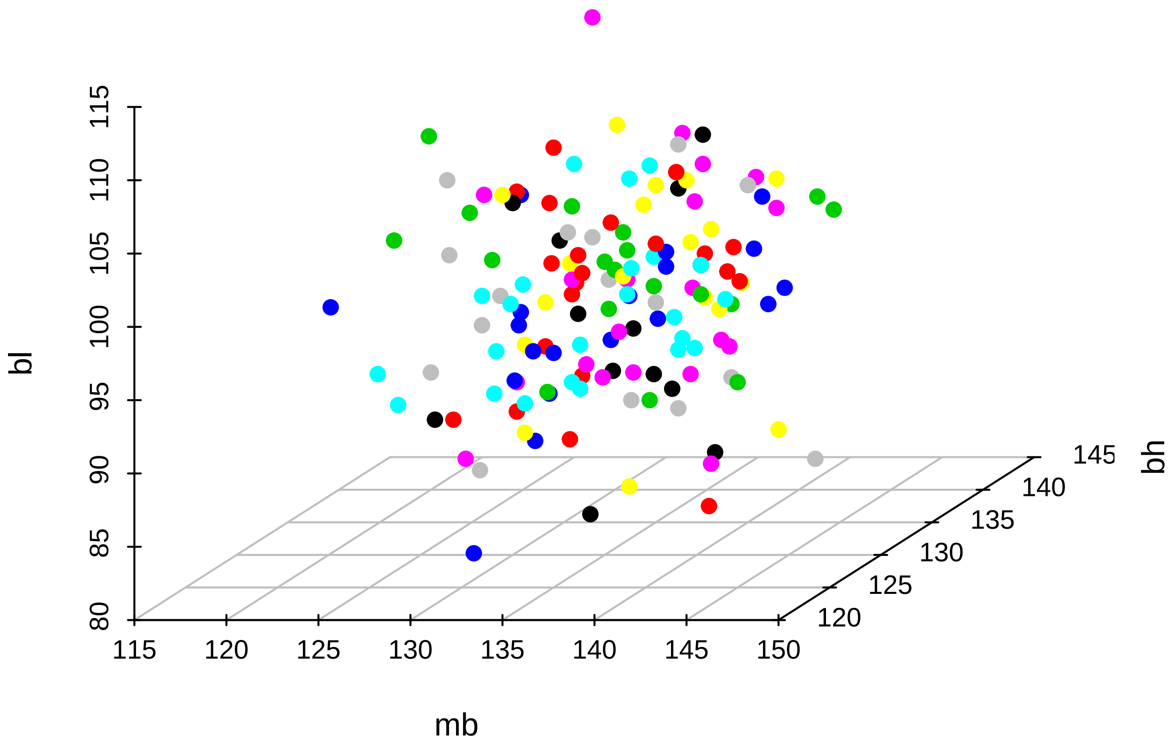
```
means = tapply(1:nrow(Skulls),
               Skulls$epoch,
               function(i) apply(Skulls[i,colnames(Skulls)[-1]], 2, mean)
               )


means = matrix(unlist(means),
               nrow = length(means),
               byrow = TRUE
               )

colnames(means) = colnames(Skulls)[-1]

rownames(means) = levels(Skulls$epoch)

means
```

```
##               mb       bh       bl       nh
## c4000BC 131.3667 133.6000 99.16667 50.53333
## c3300BC 132.3667 132.7000 99.06667 50.23333
## c1850BC 134.4667 133.8000 96.03333 50.56667
## c200BC  135.5000 132.3000 94.53333 51.96667
## cAD150  136.1667 130.3333 93.50000 51.36667
```

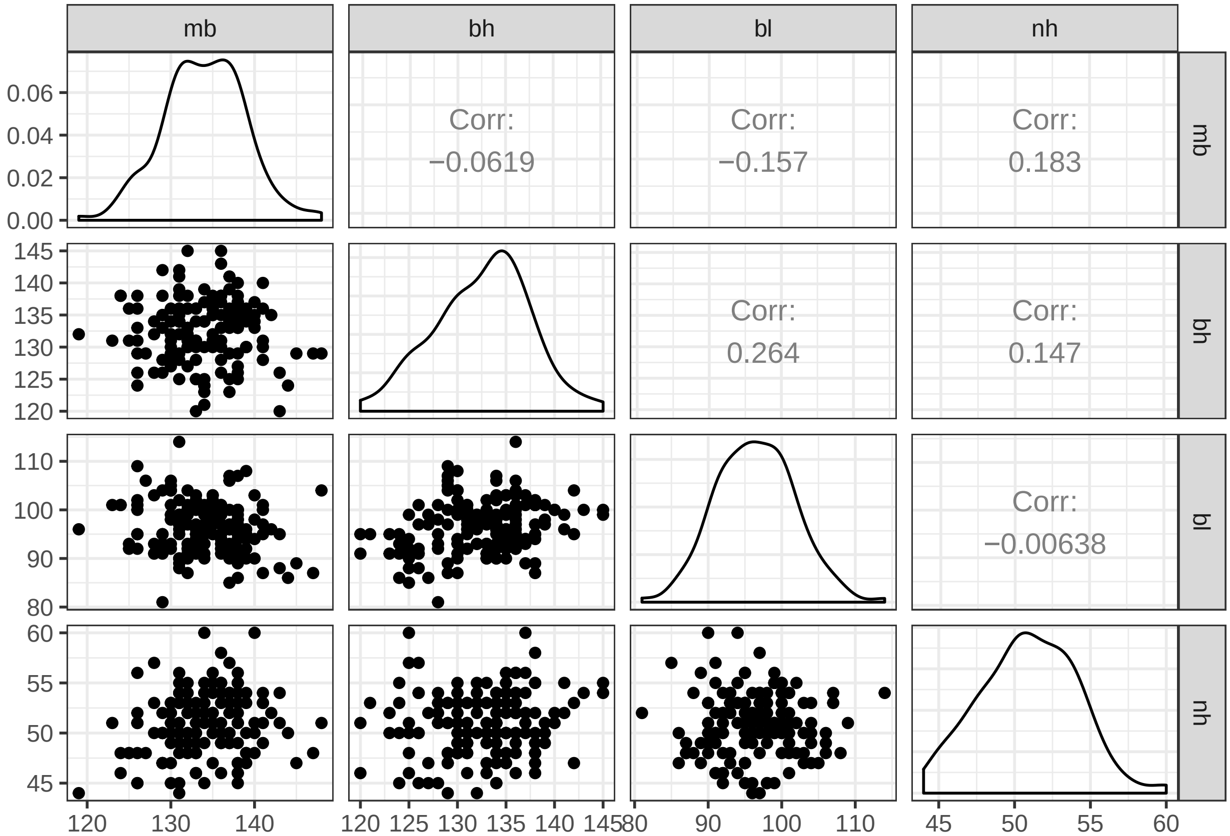
a) and present plots that you find informative.

```
library(scatterplot3d)
scatterplot3d(
  Skulls[,-1], pch = 19, color = "steelblue",
   grid = TRUE, box = FALSE,
```

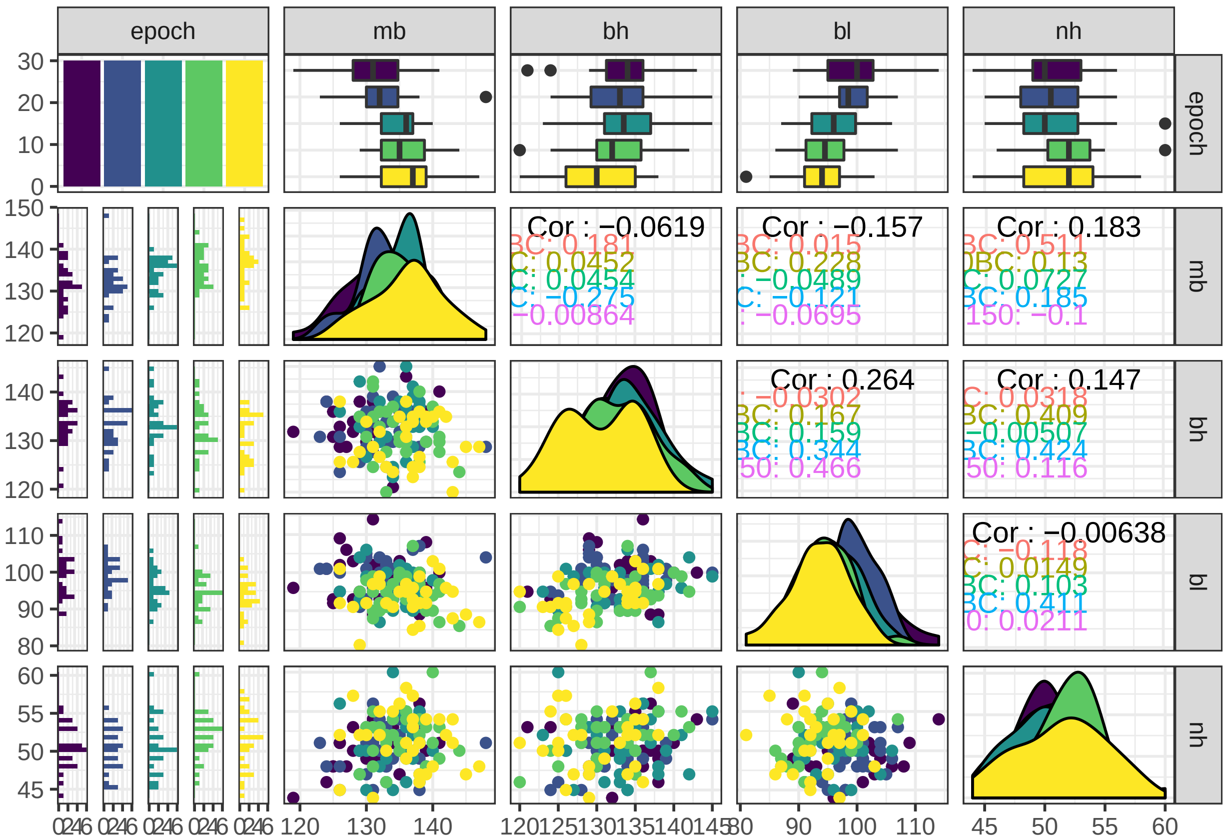
```
  mar = c(3, 3, 0.5, 3)
)
```




```
library(GGally)
library(ggplot2)
ggpairs(Skulls[,-1])+ theme_bw()
```




```
ggpairs(Skulls, aes(color = epoch))+ theme_bw()
```



epoch
mb
bh
bl
nh
Cor : −0.0619
Cor : −0.157
Cor : 0.183
Cor : 0.264
Cor : 0.147
Cor : −0.00638

b) Now we are interested whether there are differences between the epochs.

Before applying MANOVA, we state our 3 assumptions:

1) Independent observations: since a really high correlation would mean that we are measuring almost the same thing; MANOVA works best when the variables have little correlation (which is in fact our case).
2) Normality.
3) Equal variance-covariance matrices between groups.

Now, we are looking if the data are different between ephocs... We look at how much the variance from their means...

```
epoch=factor(epoch)
Y=cbind(mb,bh,bl,nh)
Skulls.manova=manova(Y~epoch)
```

We use the summary function, for calculating Wilks,Pillai, Hotteling-Lawley and Roy tests.

```
summary(Skulls.manova,test = "Wilks")
```

```
##            Df   Wilks approx F num Df den Df   Pr(>F)
## epoch       4 0.66359   3.9009     16 434.45 7.01e-07 ***
## Residuals 145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Skulls.manova,test = "Pillai")
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## epoch       4 0.35331    3.512     16    580 4.675e-06 ***
## Residuals 145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Skulls.manova,test = "Hotelling-Lawley")
```

```
##            Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## epoch       4          0.48182    4.231     16    562 8.278e-08 ***
## Residuals 145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Skulls.manova,test = "Roy")
```

```
##            Df    Roy approx F num Df den Df    Pr(>F)
## epoch       4 0.4251    15.41      4    145 1.588e-10 ***
## Residuals 145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) Do the mean vectors differ? Study this question and justify your conclusions.

We say our null hypothesis is:

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

Per our results above, we can clearly see p-value $<.05$.

Therefore, we can *reject* the null hypothesis and it is concluded there are significant differences in the means.

c) If the means differ between epochs compute and report simultaneous confidence intervals.

```
n=nrow(Y)
p=ncol(Y)
summary(Skulls.manova,test = "Wilks")
```

```
##            Df   Wilks approx F num Df den Df    Pr(>F)
## epoch       4 0.66359   3.9009     16 434.45 7.01e-07 ***
## Residuals 145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
g=5
alpha=0.05

#from here below it is as if we have 5 different datasets...
#mean for each epoch, for each variable
xbar4000BC=colMeans(Skulls[Skulls$epoch=="c4000BC",-1])
xbar3300BC=colMeans(Skulls[Skulls$epoch=="c3300BC",-1])
xbar1850BC=colMeans(Skulls[Skulls$epoch=="c1850BC",-1])
xbar200BC=colMeans(Skulls[Skulls$epoch=="c200BC",-1])
xbarAD150=colMeans(Skulls[Skulls$epoch=="cAD150",-1])

#covariance matrix for each epoch
S4000BC=cov(Skulls[Skulls$epoch=="c4000BC",-1])
S3300BC=cov(Skulls[Skulls$epoch=="c3300BC",-1])
S1850BC=cov(Skulls[Skulls$epoch=="c1850BC",-1])
S200BC=cov(Skulls[Skulls$epoch=="c200BC",-1])
SAD150=cov(Skulls[Skulls$epoch=="cAD150",-1])

W= (30-1)*S4000BC+
  (30-1)*S3300BC+
  (30-1)*S1850BC+
  (30-1)*S200BC+
  (30-1)*SAD150



qtlevel= qt(1-alpha/(p*g*(g-1)),
            df=n-g)
cii=function(x1,x2,W)
{
  list(
  lo=(x1-x2)-qtlevel*sqrt(W/(n-g)*(1/30+1/30)),
  up=(x1-x2)+qtlevel*sqrt(W/(n-g)*(1/30+1/30))
  )
}


#tau represents the ci of each epoch.
for (i in 1:p)
{
  ci=cii(xbar4000BC[i],xbar3300BC[i],W[i,i])
```

```
  cat("tau1[",i,"]-tau2[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar4000BC[i],xbar1850BC[i],W[i,i])
  cat("tau1[",i,"]-tau3[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar4000BC[i],xbar200BC[i],W[i,i])
  cat("tau1[",i,"]-tau4[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar4000BC[i],xbarAD150[i],W[i,i])
  cat("tau1[",i,"]-tau5[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar3300BC[i],xbar1850BC[i],W[i,i])
  cat("tau2[",i,"]-tau3[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar3300BC[i],xbar200BC[i],W[i,i])
  cat("tau2[",i,"]-tau4[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar3300BC[i],xbarAD150[i],W[i,i])
  cat("tau2[",i,"]-tau5[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar1850BC[i],xbar200BC[i],W[i,i])
  cat("tau3[",i,"]-tau4[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar1850BC[i],xbarAD150[i],W[i,i])
  cat("tau3[",i,"]-tau5[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

  ci=cii(xbar200BC[i],xbarAD150[i],W[i,i])
  cat("tau4[",i,"]-tau5[",i,"] belongs to(",ci$lo,",",ci$up,")\n",sep="")

}
```

```
## tau1[1]-tau2[1] belongs to(-4.905276,2.905276)
## tau1[1]-tau3[1] belongs to(-7.005276,0.8052757)
## tau1[1]-tau4[1] belongs to(-8.038609,-0.2280576)
## tau1[1]-tau5[1] belongs to(-8.705276,-0.8947243)
## tau2[1]-tau3[1] belongs to(-6.005276,1.805276)
## tau2[1]-tau4[1] belongs to(-7.038609,0.7719424)
## tau2[1]-tau5[1] belongs to(-7.705276,0.1052757)
## tau3[1]-tau4[1] belongs to(-4.938609,2.871942)
## tau3[1]-tau5[1] belongs to(-5.605276,2.205276)
## tau4[1]-tau5[1] belongs to(-4.571942,3.238609)
## tau1[2]-tau2[2] belongs to(-3.218991,5.018991)
## tau1[2]-tau3[2] belongs to(-4.318991,3.918991)
## tau1[2]-tau4[2] belongs to(-2.818991,5.418991)
## tau1[2]-tau5[2] belongs to(-0.8523246,7.385658)
## tau2[2]-tau3[2] belongs to(-5.218991,3.018991)
## tau2[2]-tau4[2] belongs to(-3.718991,4.518991)
## tau2[2]-tau5[2] belongs to(-1.752325,6.485658)
## tau3[2]-tau4[2] belongs to(-2.618991,5.618991)
## tau3[2]-tau5[2] belongs to(-0.6523246,7.585658)
## tau4[2]-tau5[2] belongs to(-2.152325,6.085658)
## tau1[3]-tau2[3] belongs to(-4.079451,4.279451)
## tau1[3]-tau3[3] belongs to(-1.046117,7.312784)
## tau1[3]-tau4[3] belongs to(0.4538827,8.812784)
```

```
## tau1[3]-tau5[3] belongs to(1.487216,9.846117)
## tau2[3]-tau3[3] belongs to(-1.146117,7.212784)
## tau2[3]-tau4[3] belongs to(0.3538827,8.712784)
## tau2[3]-tau5[3] belongs to(1.387216,9.746117)
## tau3[3]-tau4[3] belongs to(-2.679451,5.679451)
## tau3[3]-tau5[3] belongs to(-1.646117,6.712784)
## tau4[3]-tau5[3] belongs to(-3.146117,5.212784)
## tau1[4]-tau2[4] belongs to(-2.408251,3.008251)
## tau1[4]-tau3[4] belongs to(-2.741584,2.674917)
## tau1[4]-tau4[4] belongs to(-4.141584,1.274917)
## tau1[4]-tau5[4] belongs to(-3.541584,1.874917)
## tau2[4]-tau3[4] belongs to(-3.041584,2.374917)
## tau2[4]-tau4[4] belongs to(-4.441584,0.9749174)
## tau2[4]-tau5[4] belongs to(-3.841584,1.574917)
## tau3[4]-tau4[4] belongs to(-4.108251,1.308251)
## tau3[4]-tau5[4] belongs to(-3.508251,1.908251)
## tau4[4]-tau5[4] belongs to(-2.108251,3.308251)
```

```
#tau represents the ci of each epoch.
```

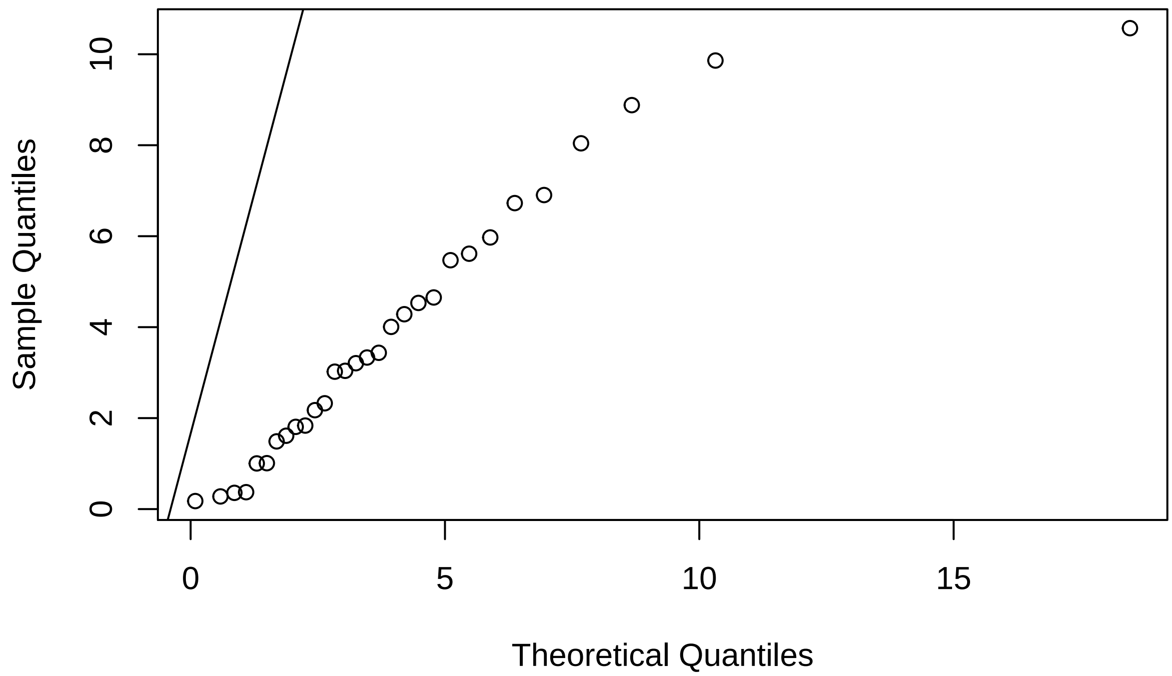
Not all simultaneous confidence intervals cover zero; therefore we can say that there is significant difference between:

tau1[1]-tau4[1] belongs to(-8.038609,-0.2280576) tau1[1]-tau5[1] belongs to(-8.705276,-0.8947243)

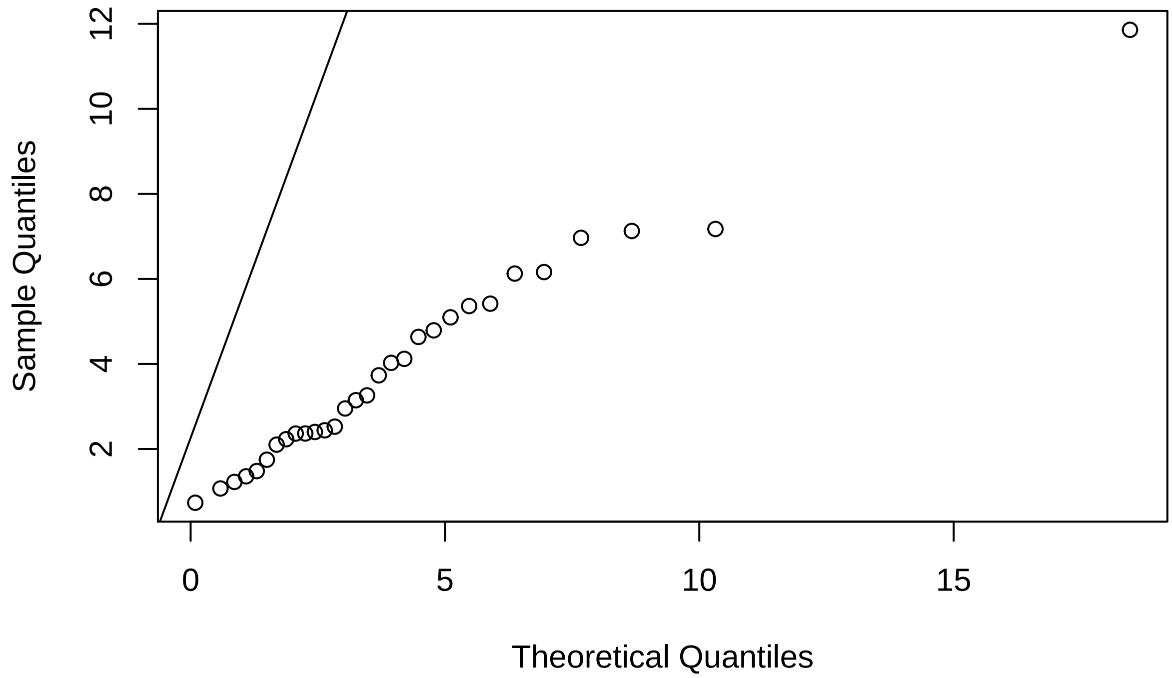
c)Inspect the residuals whether they have mean 0 and if they deviate from normality (graphically).

Now we explore if each epoch is normally distributed.

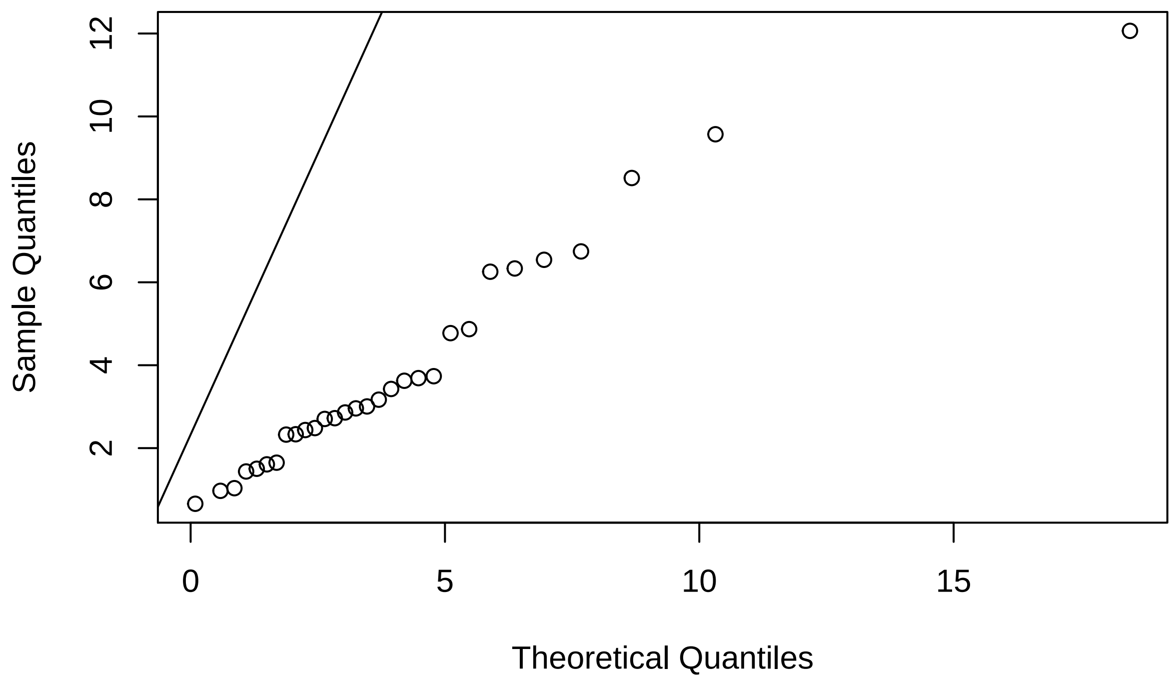
```
plotqq=function(S,epo,xbar)
{
Sinv = solve(S)
Skullsx= Skulls[Skulls$epoch==epo,-1]
datachisq =diag(t(t(Skullsx)-xbar) %*% Sinv %*% (t(Skullsx)-xbar))
qqplot(qchisq(ppoints(500),df=p),
       datachisq, main="",
      xlab="Theoretical Quantiles",ylab="Sample Quantiles")
qqline(datachisq,
       distribution=function(p) qchisq(p, df = p))
}

plotqq(S4000BC,"c4000BC",xbar4000BC)
```



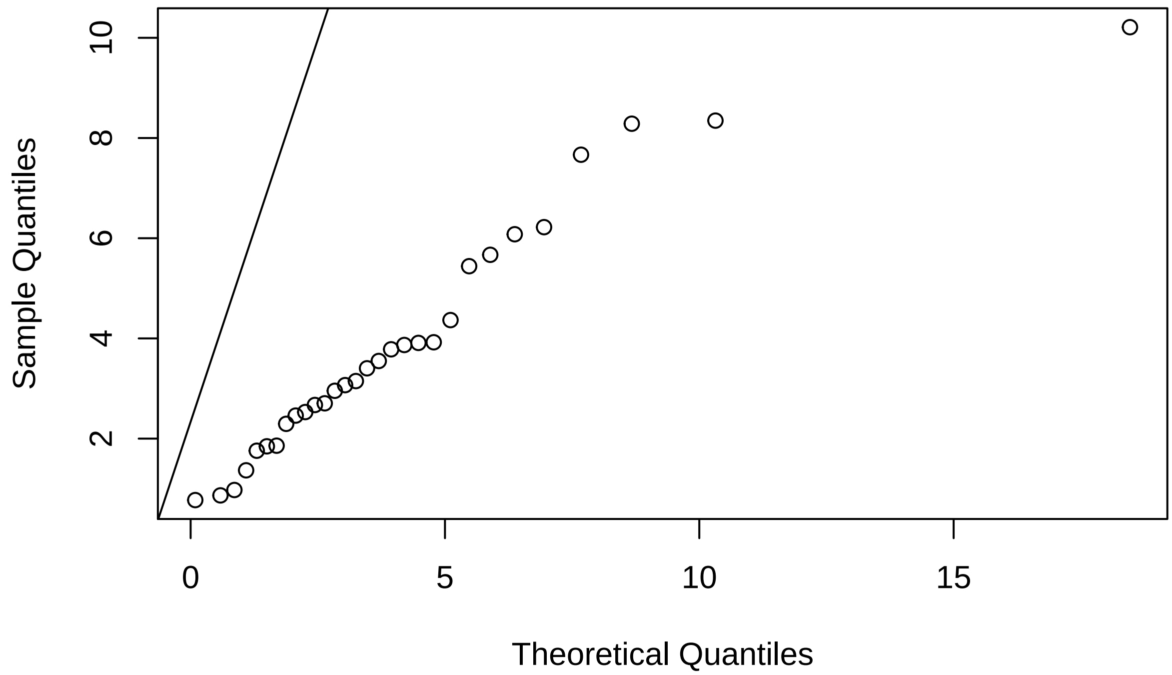

```
plotqq(S3300BC,"c3300BC",xbar3300BC)
```



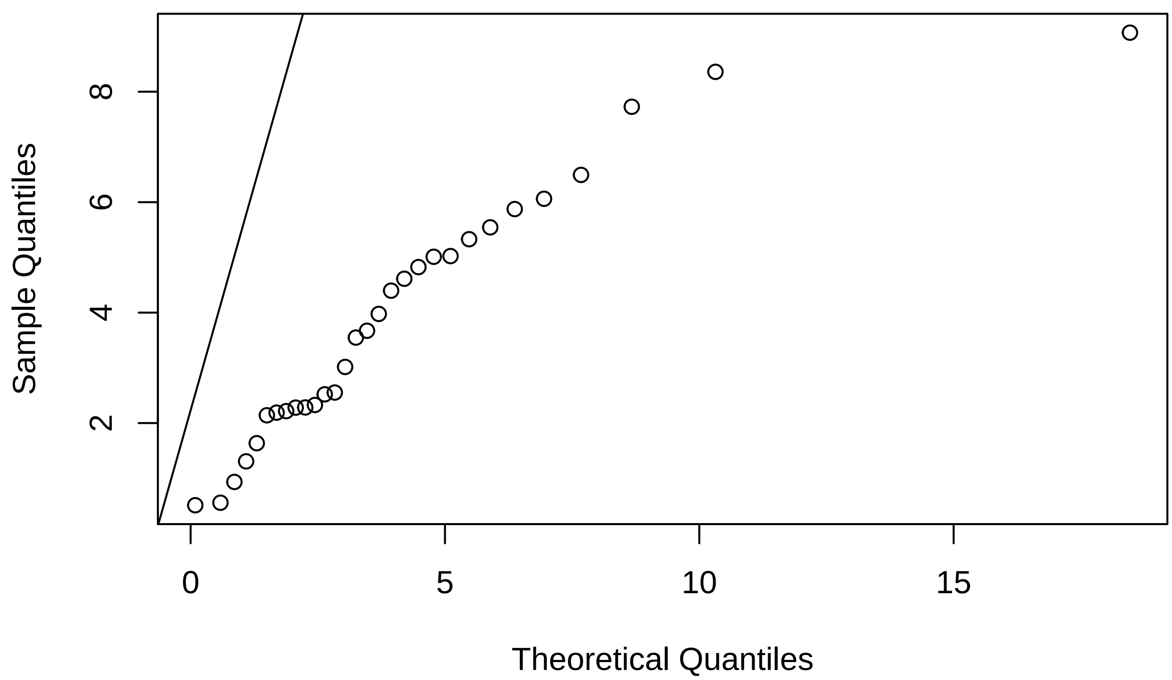

```
plotqq(S1850BC,"c1850BC",xbar1850BC)
```




```
plotqq(S200BC,"c200BC",xbar200BC)
```




```r
plotqq(SAD150,"cAD150",xbarAD150)
```



The normality assumption is violated for each epoch; therefore we conclude that the MANOVA assumptions are not realistic for the data.