assignment 5.1

Drew Murray

2022-12-18

```
wine = read.csv("C:/Users/dgmur/Downloads/wineFl2022.csv")

#1

#a

par(mar = c(1, 1, 1, 1))

par(mfrow=c(3,3))

boxplot(wine$fixed_acidity, horizontal = TRUE)

boxplot(wine$volatile_acidity, horizontal = TRUE)

boxplot(wine$citric_acid, horizontal = TRUE)

boxplot(wine$residual_sugar, horizontal = TRUE)

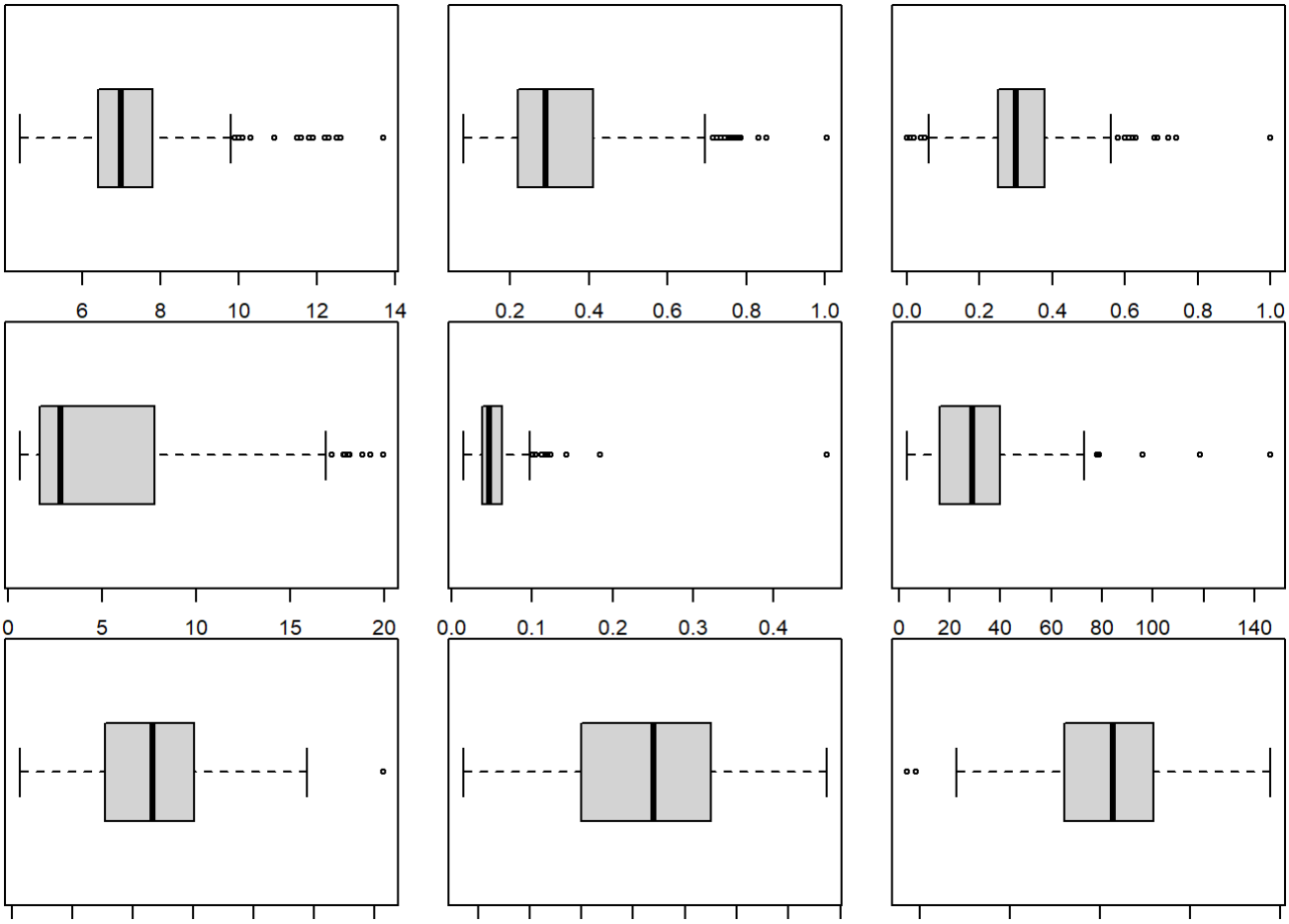
boxplot(wine$chlorides, horizontal = TRUE)

boxplot(wine$free_sulfur_dioxide, horizontal = TRUE)

boxplot(wine$total_sulfur_dioxide, horizontal = TRUE)

boxplot(wine$density, horizontal = TRUE)

boxplot(wine$pH, horizontal = TRUE)
```


6 8 10 12 14
0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20
0.0 0.1 0.2 0.3 0.4
0 20 40 60 80 100 140

```
boxplot(wine$sulphates, horizontal = TRUE)

boxplot(wine$alcohol, horizontal = TRUE)


qq1 = qqnorm(wine$fixed_acidity)
qqline(wine$fixed_acidity)


qq2 = qqnorm(wine$volatile_acidity)
qqline(wine$volatile_acidity)

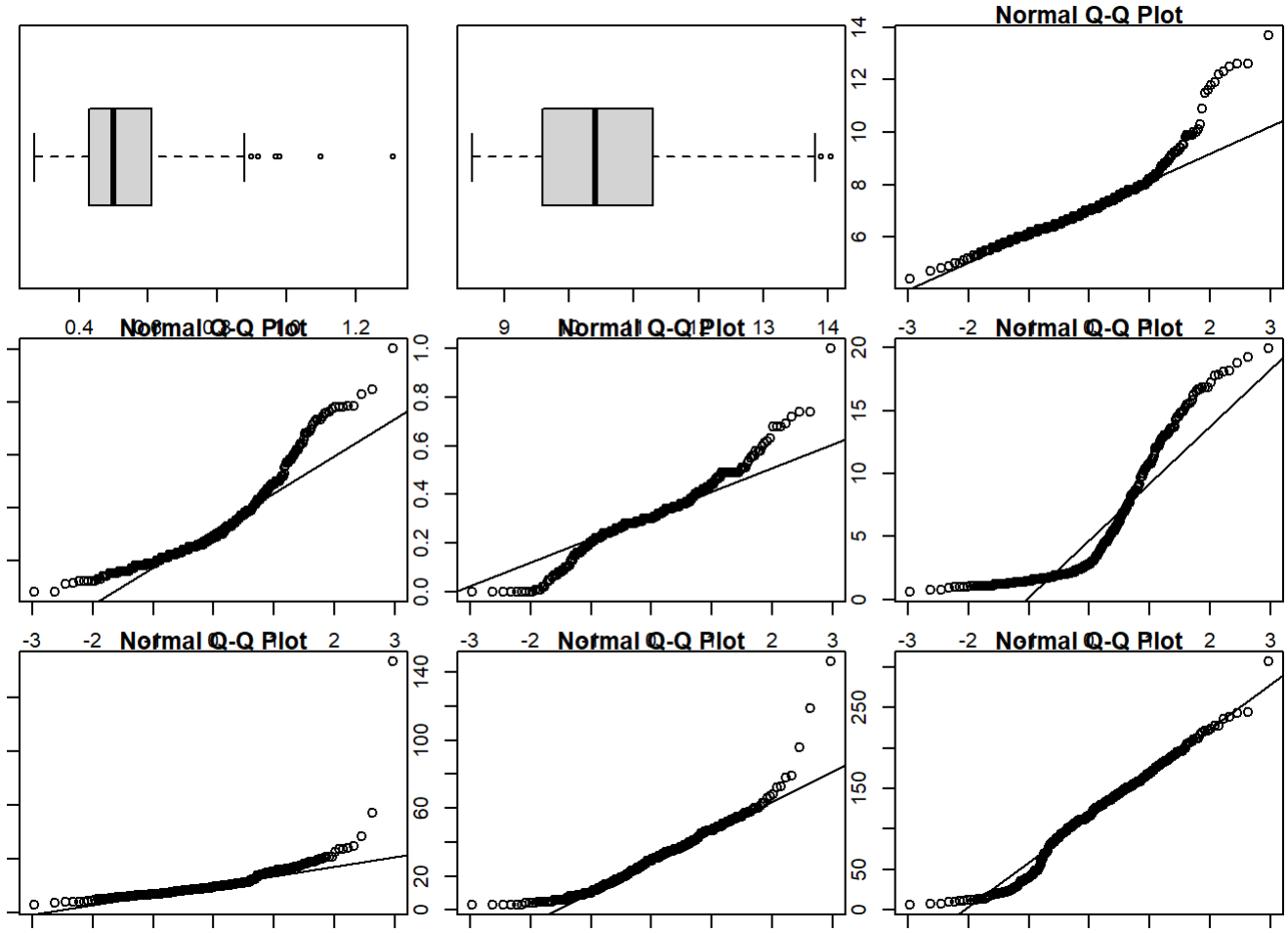

qq3 = qqnorm(wine$citric_acid)
qqline(wine$citric_acid)


qq4 = qqnorm(wine$residual_sugar)
qqline(wine$residual_sugar)


qq5 = qqnorm(wine$chlorides)
qqline(wine$chlorides)

qq6 = qqnorm(wine$free_sulfur_dioxide)
qqline(wine$free_sulfur_dioxide)

qq7 = qqnorm(wine$total_sulfur_dioxide)
qqline(wine$total_sulfur_dioxide)
```




```
qq8 = qqnorm(wine$density)
qqline(wine$density)

qq9 = qqnorm(wine$pH)
qqline(wine$pH)

qq10 = qqnorm(wine$sulphates)
qqline(wine$sulphates)

qq11 = qqnorm(wine$alcohol)
qqline(wine$alcohol)

#all variables besides density show rightly skewed, and according to the q-q plots,
#they lack normality. For density, it shows evidence of being symmetric and normality


apply(wine[1:11],2,shapiro.test)
```

```
## $fixed_acidity
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.88806, p-value = 2.558e-15
##
##
## $volatile_acidity
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.89262, p-value = 5.565e-15
##
##
## $citric_acid
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.96288, p-value = 9.248e-08
##
##
## $residual_sugar
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.8111, p-value < 2.2e-16
##
##
## $chlorides
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.59094, p-value < 2.2e-16
##
##
## $free_sulfur_dioxide
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.91989, p-value = 1.01e-12
##
##
## $total_sulfur_dioxide
##
##  Shapiro-Wilk normality test
##
```

```
## data:  newX[, i]
## W = 0.97692, p-value = 2.153e-05
##
##
## $density
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.98133, p-value = 0.0001657
##
##
## $pH
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.99657, p-value = 0.6624
##
##
## $sulphates
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.93298, p-value = 1.907e-11
##
##
## $alcohol
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.95496, p-value = 7.085e-09
```

```
#using the critical value of 0.995 all variables were signficant in rejecting
#normality assumption besides pH

#Chi-square plot
wine=as.matrix(wine)
n=nrow(wine)
Xbar=colMeans(wine[,1:11])

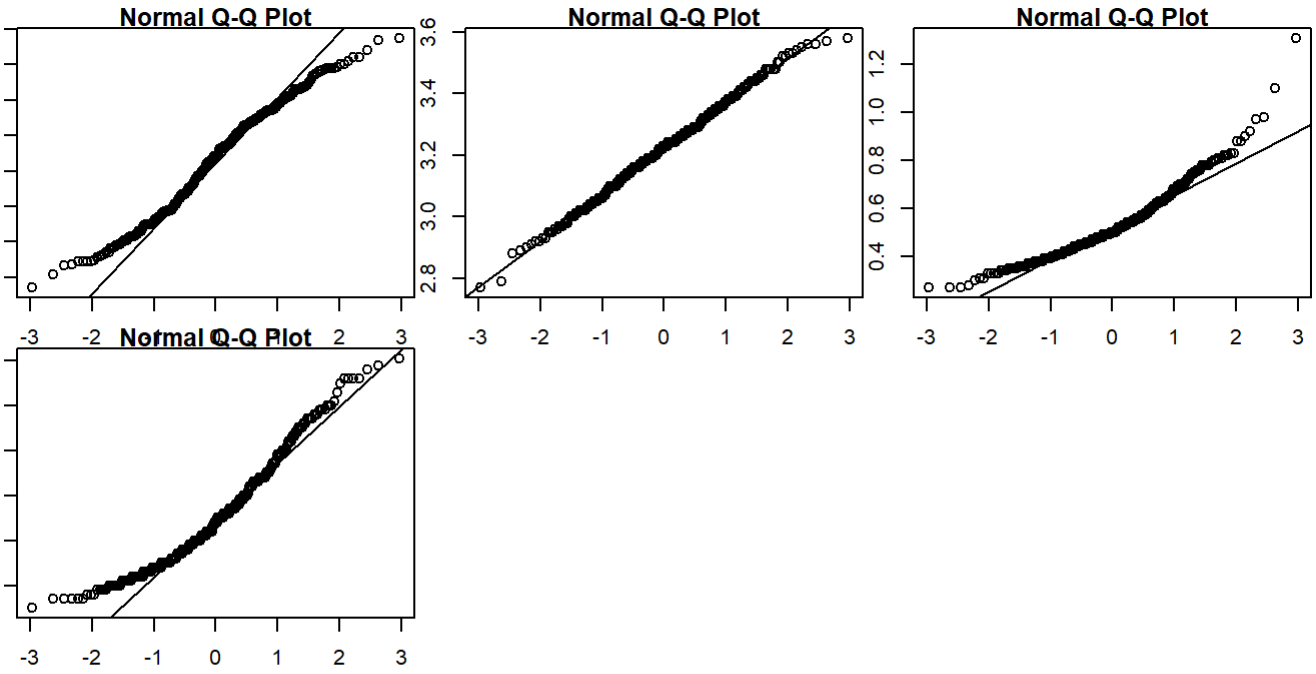
S=var(wine[,1:11])

invS=solve(S)

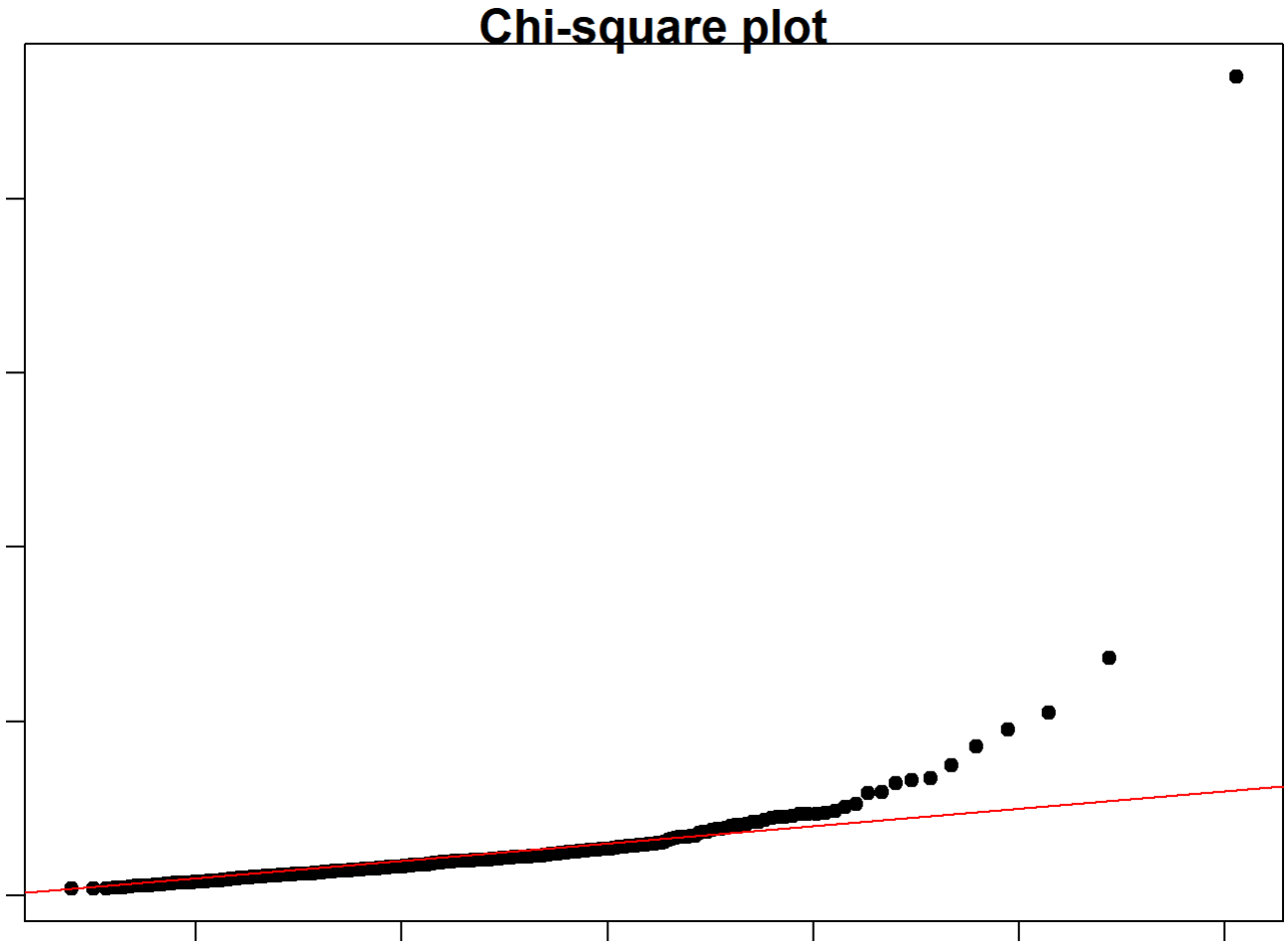
D=rep(0,n)

for (i in 1:n){D[i]=t(wine[i,1:11]-Xbar)%*%invS%*%(wine[i,1:11]-Xbar)}

par(mfrow=c(1,1))
```




```
plot(qchisq((seq(1:n)-0.5)/n,11),sort(D),pch=19,main="Chi-square plot",xlab="Chi-square quantile
s",ylab="generalized distances",cex.axis=1, cex.lab=1.5, cex.main=1.5)
abline(0,1,lty=1,cex=1.5,col="red")
```




```
#B
```

The chi-square plot is used to check normality. Shows deviation from multivariate normality due to outliers skewing the the line upwards.

```
#C
 source("C:/Users/dgmur/Downloads/functions.R")
a = distances(wine)
tail(sort(a))
```

```
## [1]  44.19994  52.20142  53.87997  60.28769  70.66763 237.25010
```

```
tail(order(a))
```

```
## [1]   7  39 141 124 157  63
```

```
round(wine[c(157,63),],2)
```

```
##      fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
## [1,]           7.1             0.49        0.22            2.0      0.05
## [2,]           7.8             0.41        0.68            1.7      0.47
##      free_sulfur_dioxide total_sulfur_dioxide density   pH sulphates alcohol
## [1,]               146.5                307.5    0.99 3.24      0.37    11.0
## [2,]                18.0                 69.0    1.00 3.08      1.31     9.3
##      quality type
## [1,]       4    0
## [2,]       5    1
```

Observation 157 and 63 are the outliers. 157 is a white wine that has an exteme value in total sulfur dioxide, and 63 is a red wine that has an extreme value free sulfur dioxide and chlorides.

```
#D

library(MASS)
source("C:/Users/dgmur/Downloads/functions.R")

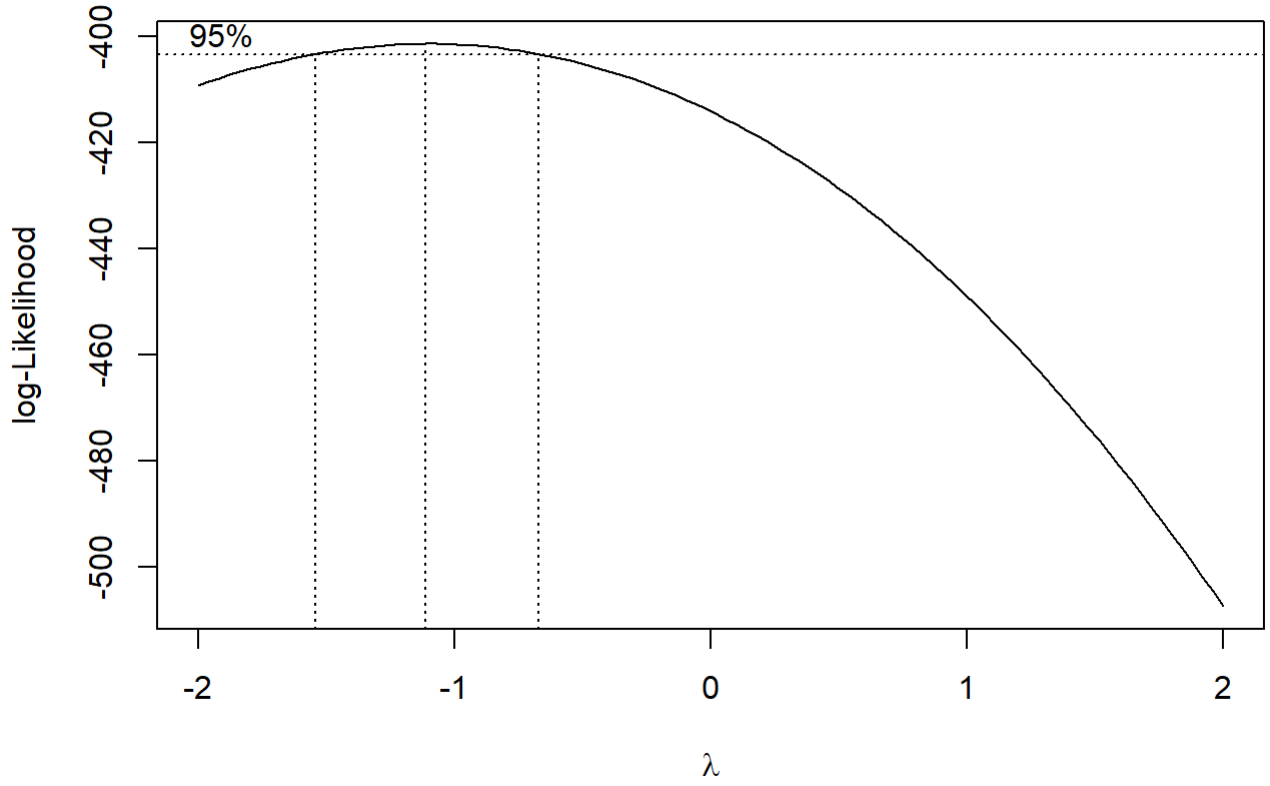

wine2 = wine[,1:11]

wine2[,3]=wine[,3]+.01

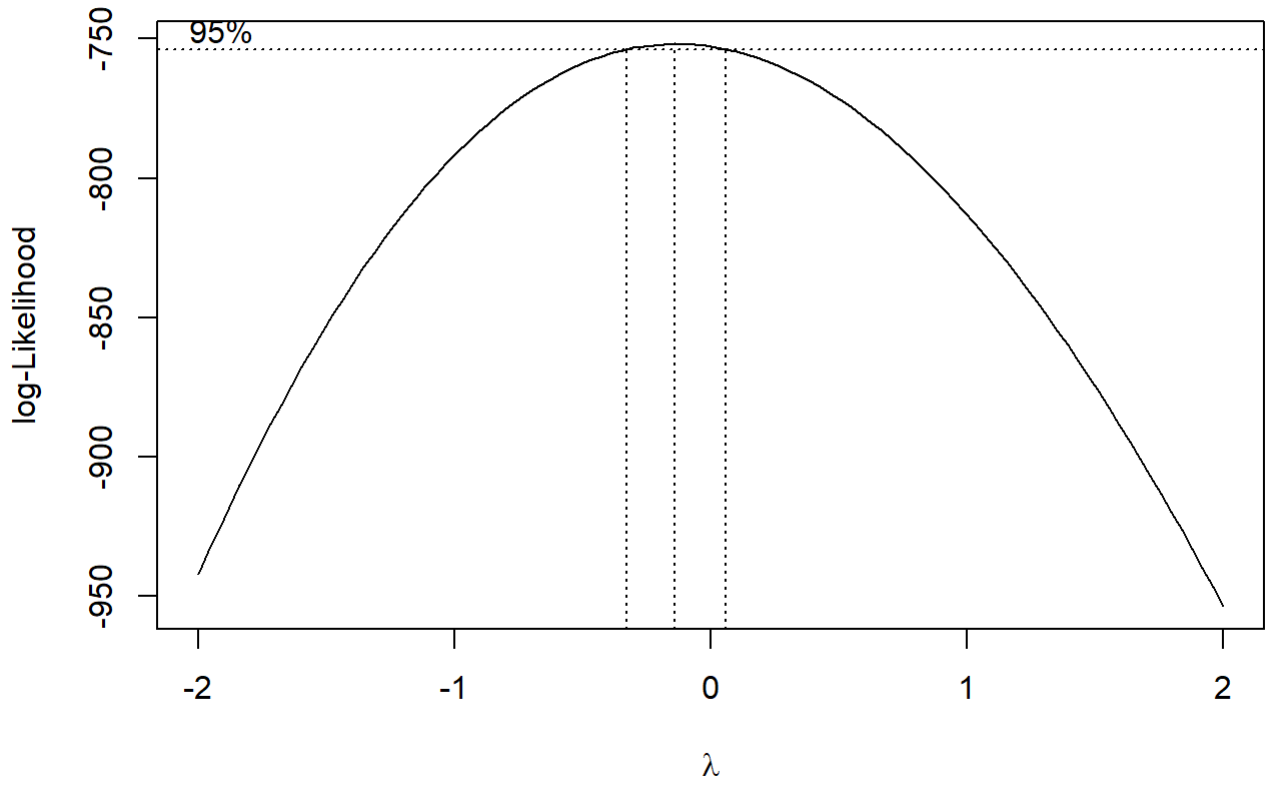
boxcoxestimate(wine2)
```

```
##  [1]  0.4733031  0.3515782  0.7468556  0.2388118 -0.2249394  0.3237410
##  [7]  0.4656051  3.6459611  1.8299778 -0.1602876 -1.3503966
```

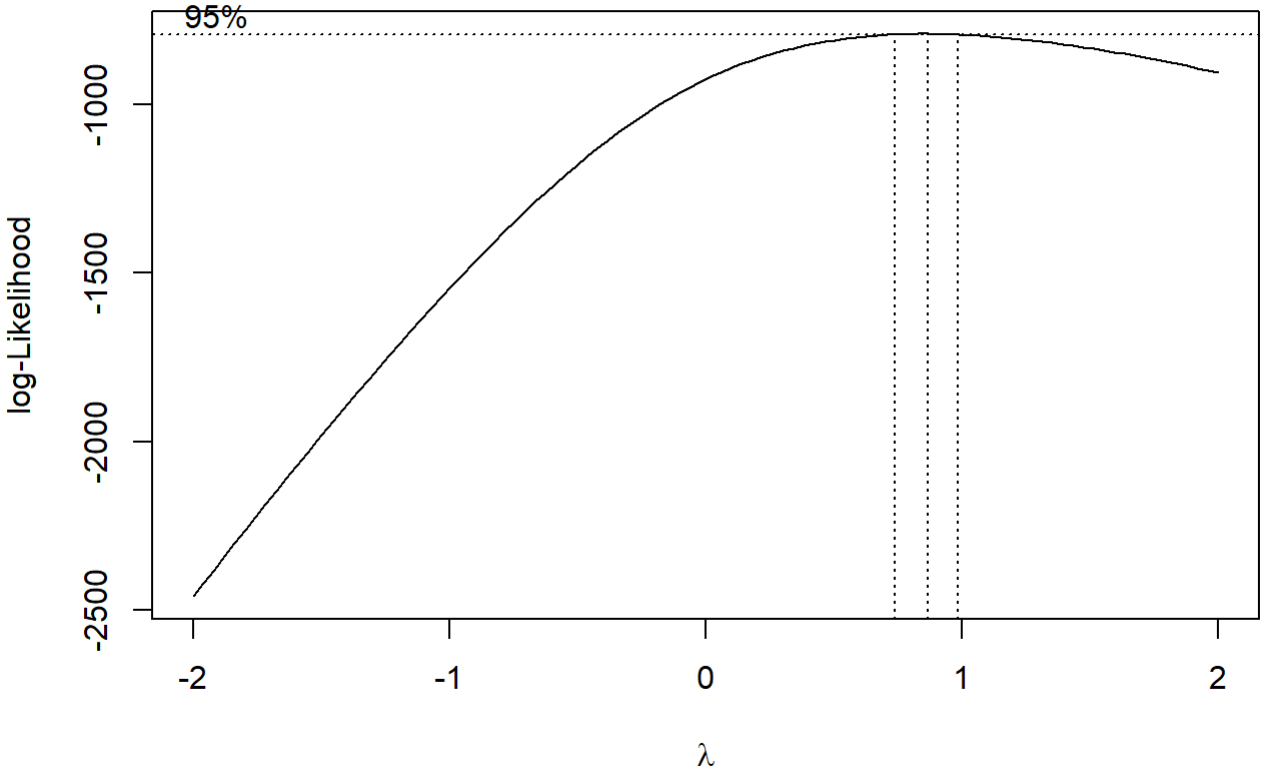
```
boxcoxplot(wine2[,1])
```



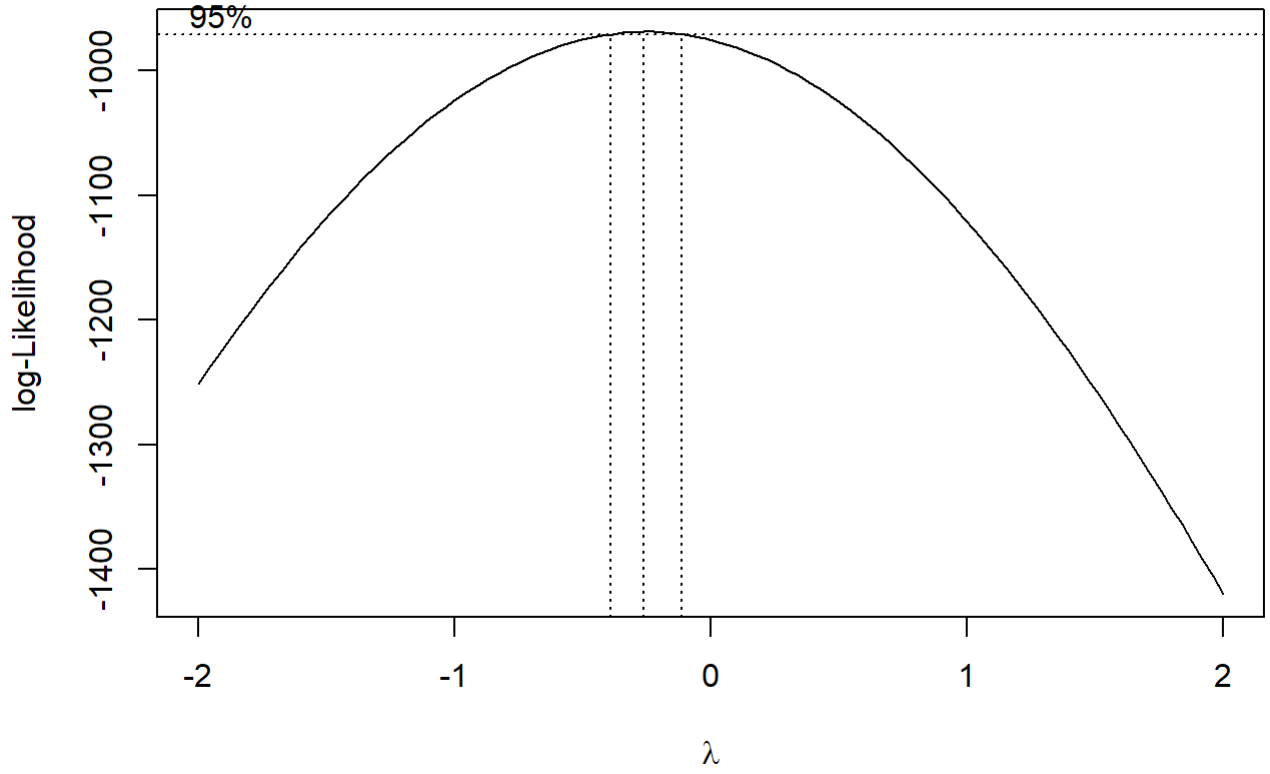

```
boxcoxplot(wine2[,2])
```



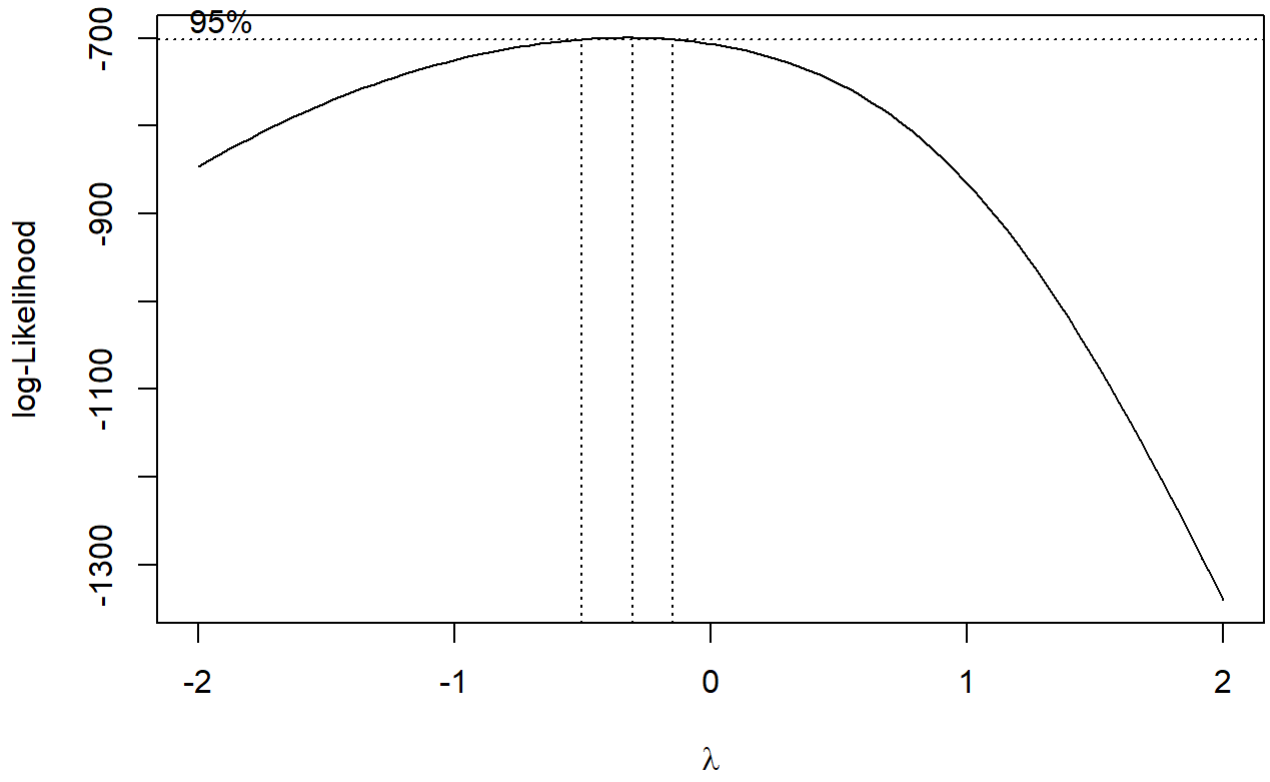

```
boxcoxplot(wine2[,3])
```



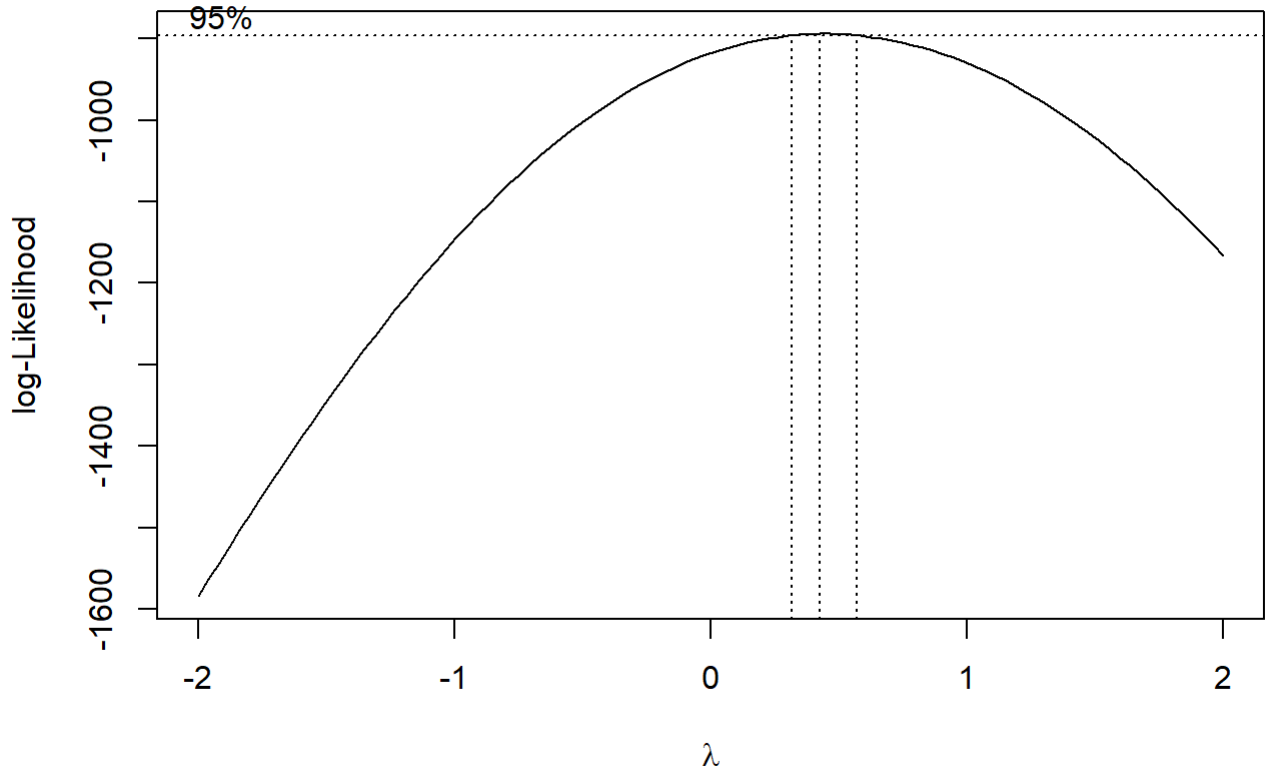

```
boxcoxplot(wine2[,4])
```



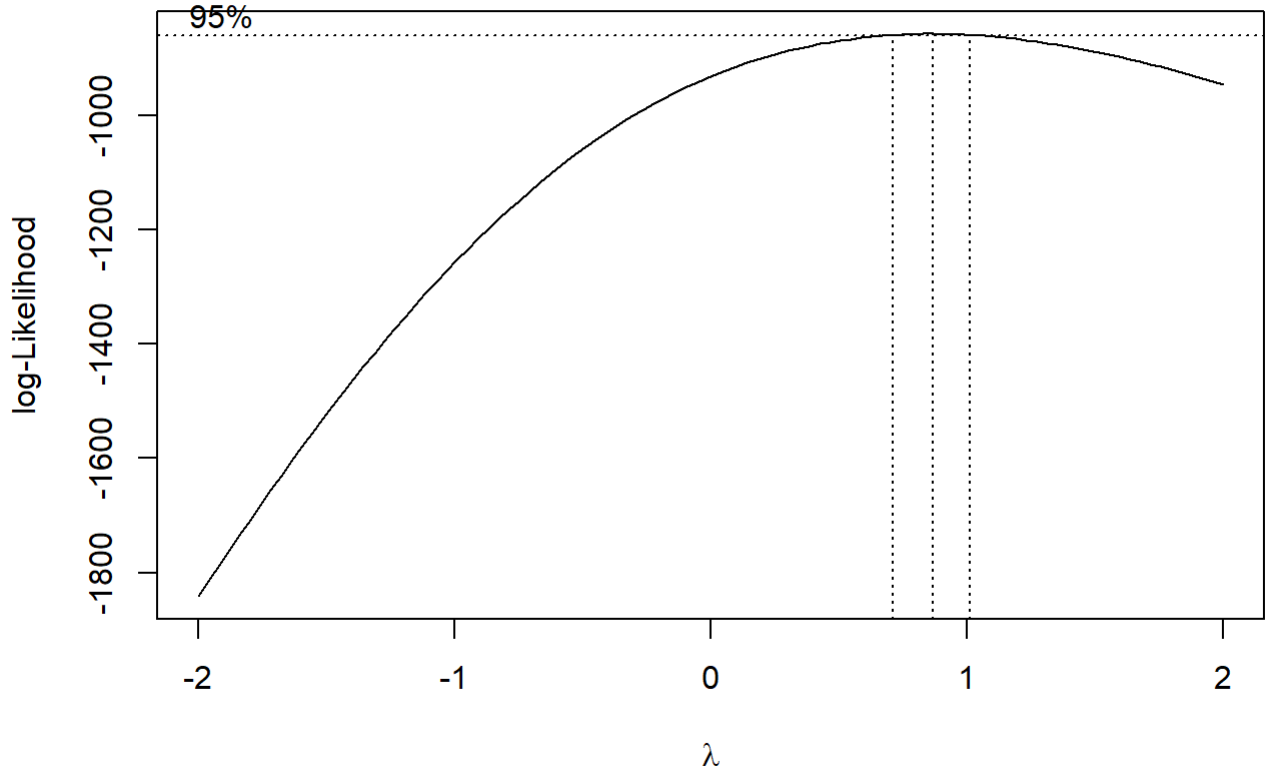

```
boxcoxplot(wine2[,5])
```



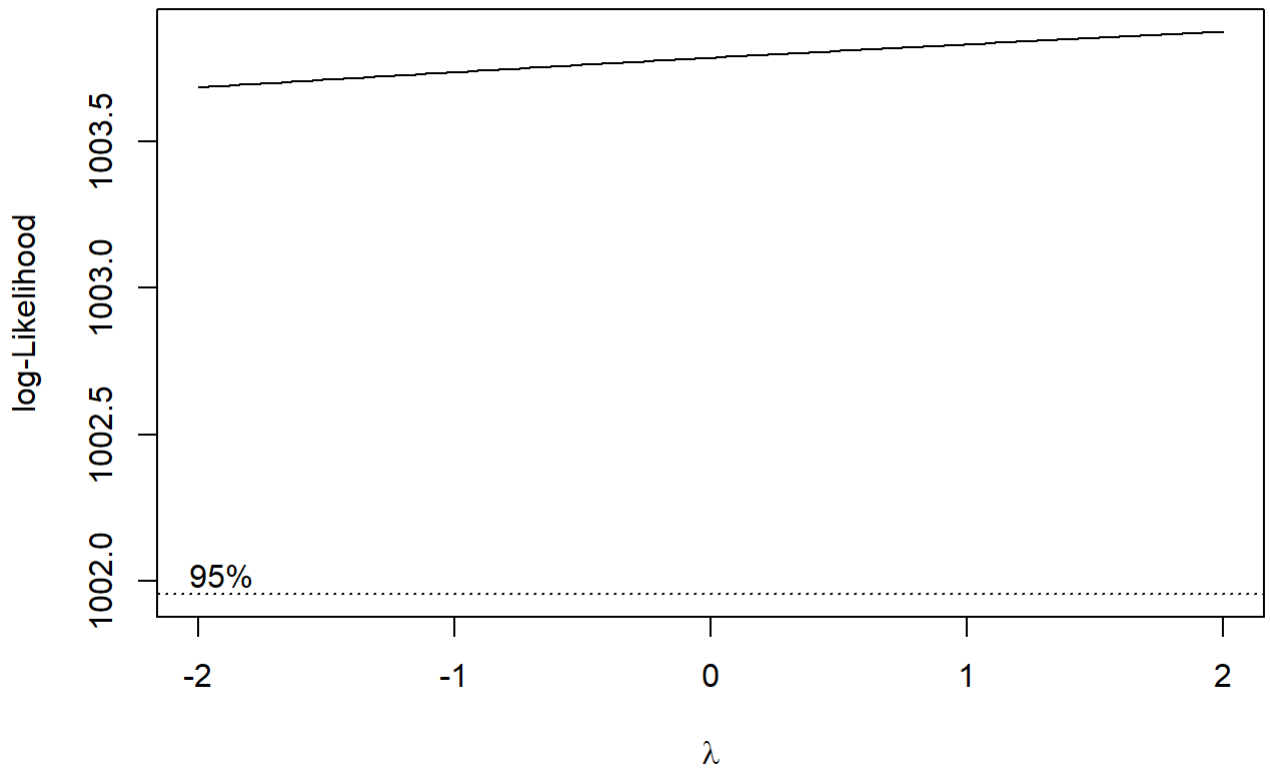

```
boxcoxplot(wine2[,6])
```



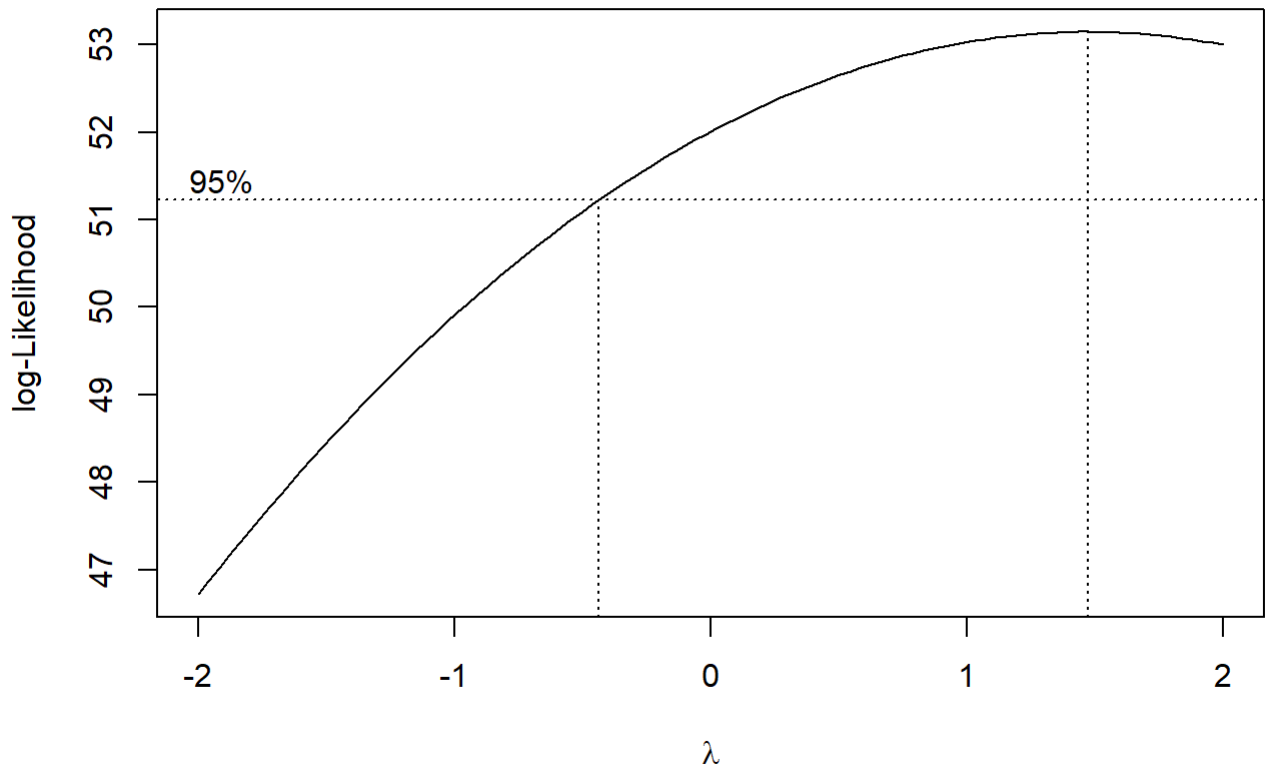

```
boxcoxplot(wine2[,7])
```



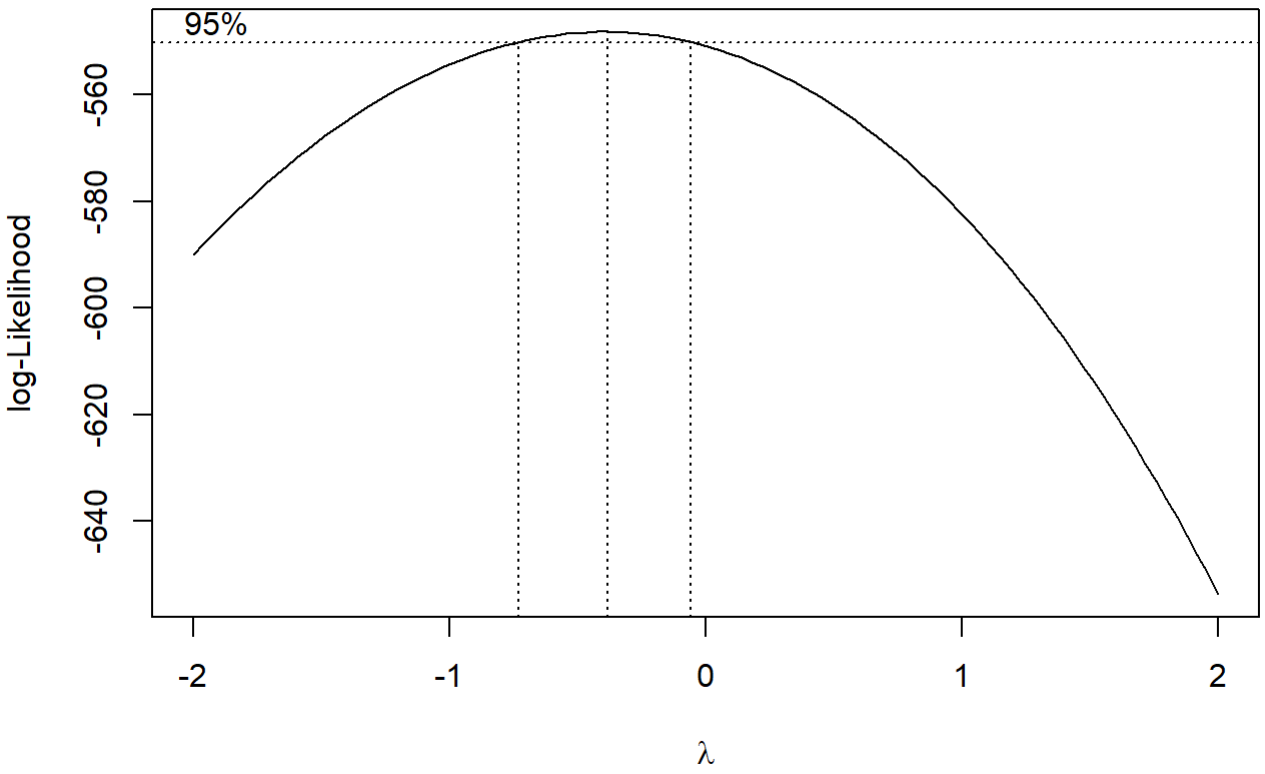

```
boxcoxplot(wine2[,8])
```



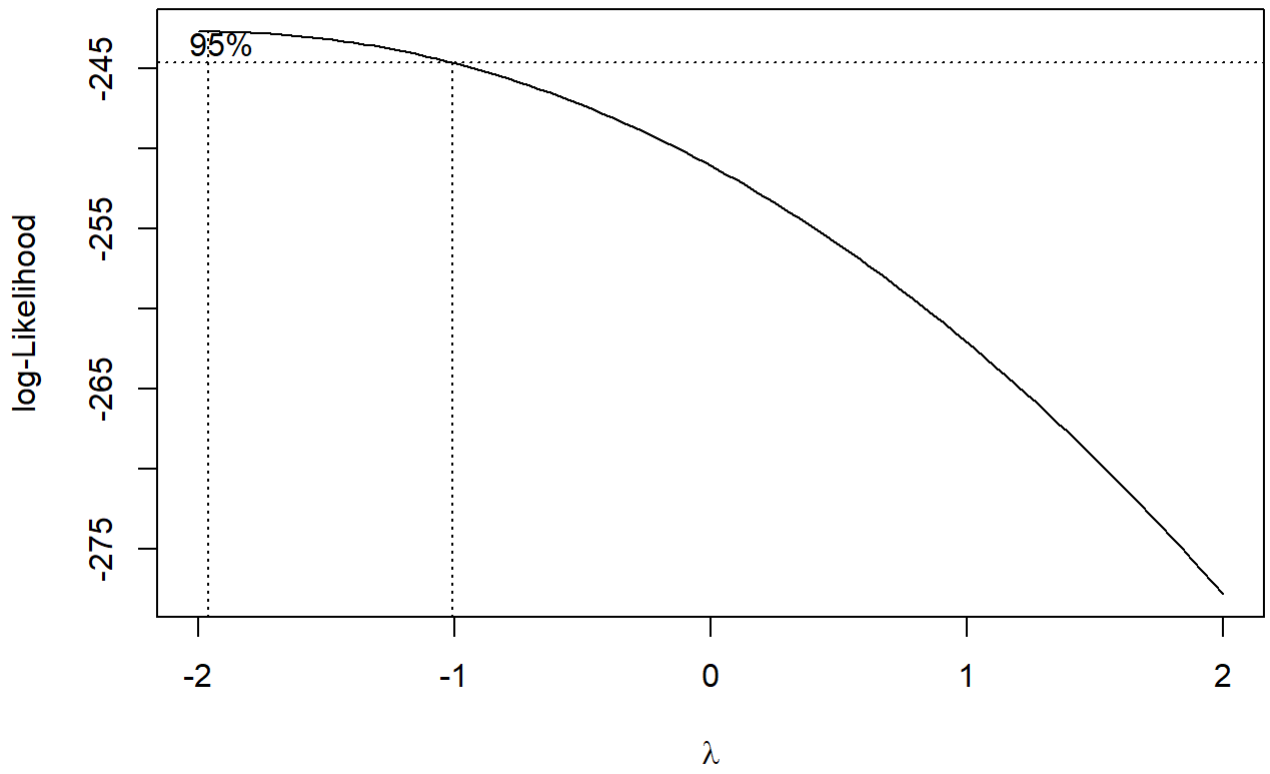

```
boxcoxplot(wine2[,9])
```




```
boxcoxplot(wine2[,10])
```




```
boxcoxplot(wine2[,11])
```




```
par(mfrow=c(4,4))
```

Using the boxcox transformation method. It show that most of the variables have improved after the transformation. #E

if the population mean for both red wines and white wines are different than combining may cause bimodality Bimodality shows a lack of normality in the data

```
```r
#Bonus


e = as.matrix(wine[,c(7,11)])
e[1:3,]
```

```
##      total_sulfur_dioxide alcohol
## [1,]                  144    10.3
## [2,]                  100     9.2
## [3,]                  148    10.7
```

```
barX <- c(mean(e[,1]), mean(e[,2]))
barX
```
```

```
## [1] 113.46286  10.56176
```

```
S <- var(e)

source("ellipseFunctions.R")

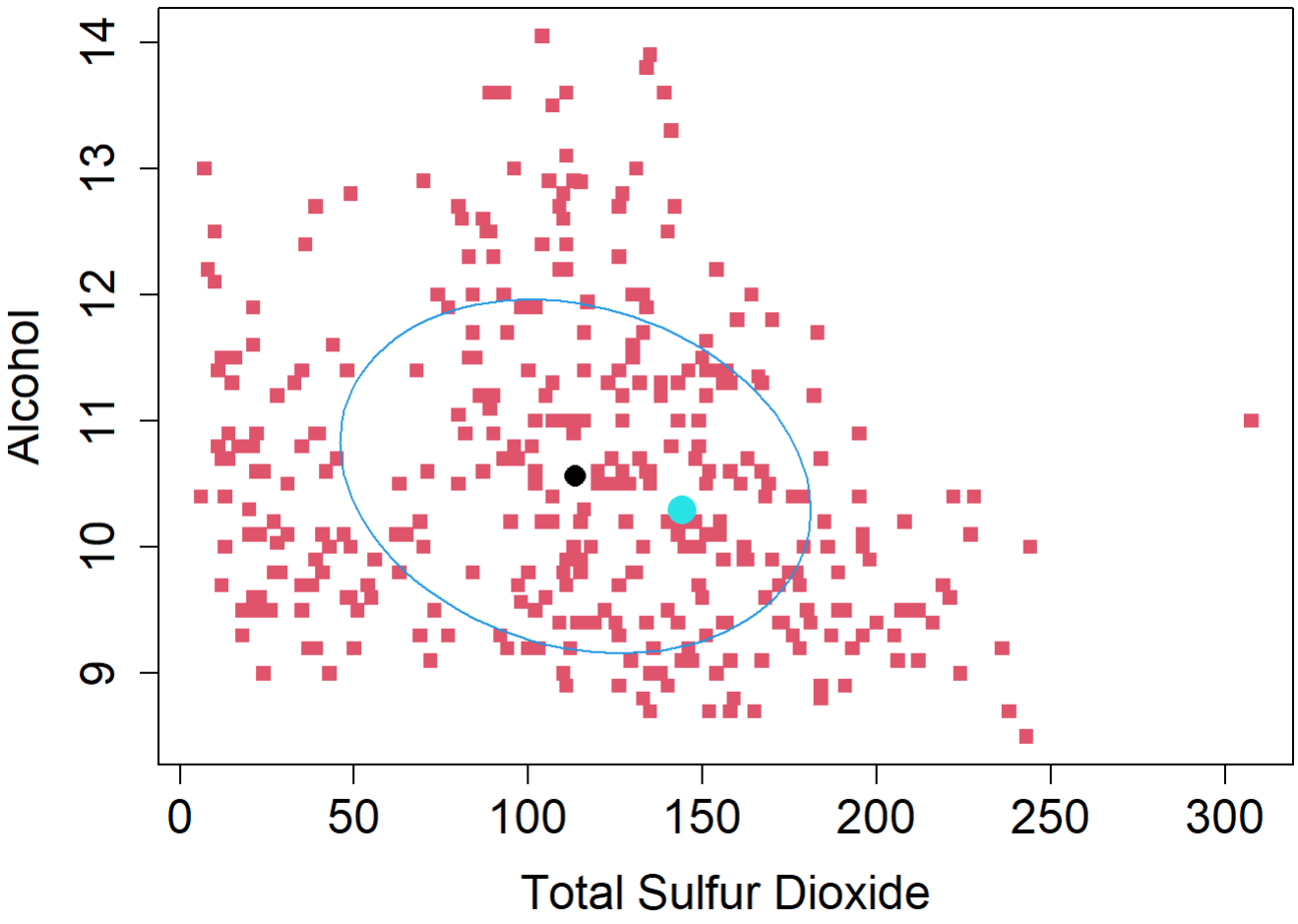
par(mfrow=c(1,1), mar=c(4,4,1,1))

plot(e,xlab="Total Sulfur Dioxide", ylab="Alcohol", pch=15, col=2, cex.axis=1.5,cex.lab=1.5)

points(barX[2]~barX[1], col=1, pch=16,cex=1.5)

points(e[1,2]~e[1,1],pch=16,col=5,cex=2)

ellipsem(barX,solve(S),1.39, col=4)
```




```
i=1

xi <-e[i,]

sc <- (xi - barX) %*% solve(S) %*% (xi - barX)

sc
```

```
##           [,1]
## [1,] 0.2986895
```

```
Inside=0

for( i in 1:nrow(e))
    {
      xi <- e[i,]
      sc <- (xi - barX) %*% solve(S) %*% (xi - barX)
      if(sc < 1.39) Inside=Inside+1
      }

Inside
```

```
## [1] 149
```

```
Inside/nrow(e)
```

```
## [1] 0.4257143
```

```
#2

men = read.table("C:/Users/dgmur/Downloads/T1-9.dat")
women = read.table("C:/Users/dgmur/Downloads/T8-6.dat")

#a

t1 <- men[,c(2:6)]
t2 <- women[,c(2:6)]

d <- t1 - t2

dbar<-colMeans(d)

round(dbar,2)
```

```
##   V2   V3   V4   V5   V6
## 1.14 2.58 6.16 0.25 0.54
```

```
covd = cov(d)



library(ICSNP)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: ICS
```

```
HotellingsT2(t1,t2, level = 0.5)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  t1 and t2
## T.2 = 105.1, df1 = 5, df2 = 102, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0)
```

```
#male mean vector is different from female mean vector
#b
p<-ncol(t1)

alpha = 0.05

m = men[,c(1:6)]

w = women[,c(1:6)]

n1 = nrow(m)

n2 = nrow(w)


for(i in 1:p){

   cat("Variable",i, dbar[i] + c(-1,1)*qt(alpha/(2*p),n-1,lower.tail=FALSE)

       *sqrt(covd[i,i]/n),"\n")
}
```

```
## Variable 1 1.10247 1.179381
## Variable 2 2.486037 2.668037
## Variable 3 5.917326 6.402674
## Variable 4 0.2462381 0.2622805
## Variable 5 0.5131771 0.5590451
```

```
L1 = cov(t1)

L2= cov(t2)

Sp=((n1-1)/(n1+n2-2))*L1 + ((n2-1)/(n1+n2-2))*L2

xbar1 = colMeans(t1)

xbar2 = colMeans(t2)

for(i in 1:p){
   cat("Variable",i,(xbar1-xbar2)[i] + c(-1,1)*qt(alpha/(2*p),n1+n2-2,lower.tail=FALSE)
         *sqrt((1/n1 + 1/n2)*Sp[i,i]),"\n")
}
```

```
## Variable 1 0.9795923 1.30226
## Variable 2 2.191933 2.962141
## Variable 3 5.100199 7.219801
## Variable 4 0.2180364 0.2904821
## Variable 5 0.4248167 0.6474055
```

```
#C
```

Paired sample approach was used because each population has a naturally pairing with the country

```
#Bonus Points 2
wine1 = as.data.frame(wine)

#1
m1 <- manova(cbind(fixed_acidity, density) ~ as.factor(quality), data= wine1)

summary.manova(m1,test = c("Wilks"))
```

```
##                      Df   Wilks approx F num Df den Df    Pr(>F)
## as.factor(quality)   4 0.87444   5.9671      8    688 2.036e-07 ***
## Residuals          345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#There is a signifcant difference

#2
m2 = manova(cbind(volatile_acidity, alcohol) ~ as.factor(quality), data= wine1)

summary.manova(m2,test = c("Wilks"))
```

```
##                      Df   Wilks approx F num Df den Df    Pr(>F)
## as.factor(quality)   4 0.74225   13.822      8    688 < 2.2e-16 ***
## Residuals          345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#There is a significant difference, and the answer did not change

#3
```

Shows that there is not a clear distinction, we cannot conclude that we can differentiate between all of the quality scores