Detecting Outliers

Tanweer Shapla

2023-10-10

## Detecting Multivariate outliers using Mahalanobis distance statistic

## Example 1 : trees data

We use trees data from R. First we find the variable names, dimension, mean of each variable, and covariance matrix for variables in trees data.

names(trees)

[1] "Girth" "Height" "Volume"

dim(trees)

[1] 31 3

smean=colMeans(trees)   
scov=cov(trees)  
print(list(smean, scov))

[[1]]  
 Girth Height Volume   
13.24839 76.00000 30.17097   
  
[[2]]  
 Girth Height Volume  
Girth 9.847914 10.38333 49.88812  
Height 10.383333 40.60000 62.66000  
Volume 49.888118 62.66000 270.20280

The Mahalanobis distance for each multivariate observation is calculated as

for , where is the observation vector, is the sample mean vector, and is the sample covariance matrix.

For large n, approximately follows a chi-squared distribution with degrees of freedom, is the number of variables in the data.

We find Mahalanobis distance for each observation using mahalanobis{stats} function in R. Note that mahalanobis() function has the following arguments: data matrix, mean vector, and covariance matrix. We compare each Mahalanobis distance against the critical value obtained from Table A.6 (Methods of Multivariate Analysis by Rencher, page 557). For and , the critical value is found to be for . Any observation whose Mahalanobis distance is greater than the critical value is considered an outlier.

n=nrow(trees)  
mhd.trees=mahalanobis(trees, smean, scov)  
index=which(mhd.trees>12.24)#gives the observation number for which the mahalanobis distance is greater than 12.24  
index

integer(0)

mhd.trees[index]

numeric(0)

From the output, we find that there is no outlier in the trees data.

If we use the upper 5th percentile chi-squared critical value with p=3 degrees of freedom, then we find that the 31st observation with a value of 10.96274 appears to be an outlier. This is the point at the far right side of the qqplot.

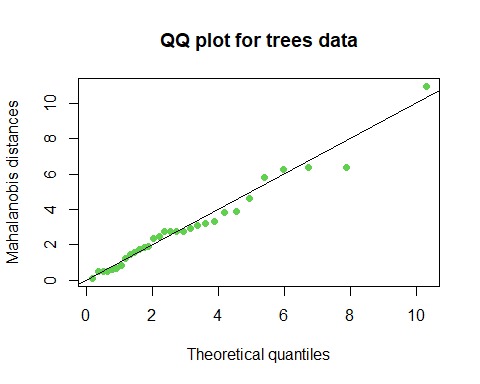
chisq.c=qchisq(0.95, 3, lower.tail=T)  
out=which(mhd.trees>chisq.c)  
out

[1] 31

mhd.trees[out]

[1] 10.96274

theo.quan=qchisq(ppoints(n),3) #df=number of variables = 3  
qqplot(theo.quan, mhd.trees, xlab="Theoretical quantiles", ylab="Mahalanobis distances", pch=16, col=3, main="QQ plot for trees data")  
abline(0,1) #drawing reference line



## Example 2: iris data

We compare each Mahalanobis distance against the critical value obtained from Table A.6 (Methods of Multivariate Analysis by Rencher, page 557). For $p=4 and , the critical value is found to be for . There appears to be no outlier in the iris data.

dim(iris)

[1] 150 5

mean.iris=colMeans(iris[, 1:4])  
cov.iris=cov(iris[,1:4])  
mhd.iris=mahalanobis(iris[,1:4],mean.iris, cov.iris)  
which(mhd.iris>19.51)

integer(0)

If we use an upper 2.5th percentile chi-squared critical value , we find observations 42, 115, 118, 132, 135, and 142 appeared to be outliers.

A more extreme percentile can be used to detect observations that seriously depart from the overall pattern of the datapoints.

which(mhd.iris>qchisq(0.975, 4, lower.tail = T))

[1] 42 115 118 132 135 142

t.q=qchisq(ppoints(nrow(iris)), 4)  
qqplot(t.q, mhd.iris, xlab="Theoretical quantiles", ylab="Mahalanobis distance", main="QQ plot for iris data", ylim=c(0, 14))

