Checking Multivariate Normality

T Shapla

09/27/2023

## Checking MV normality using Chisquared QQ plot (also known as Gamma plot) for 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.

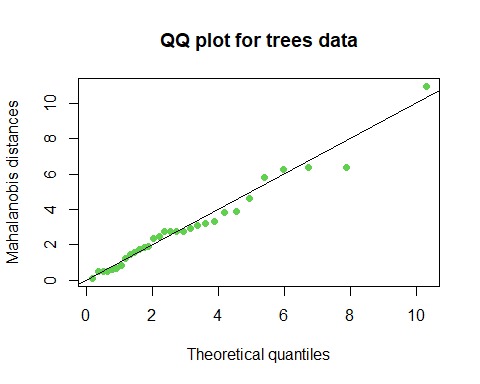
For large n, approximately follows a chi-squared distribution with degrees of freedom, is the number of variables in the data.

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 plot ordered Mahalanobis distances against the quantiles from a chi-squared distribution with p degrees of freedom. Scattering of points around a straight line suggests multivariate normality.

n=nrow(trees)  
mhd.trees=mahalanobis(trees, smean, scov)  
mhd.trees

[1] 4.6286434 5.7962500 6.3984665 0.8276614 2.7334540 3.8283678  
 [7] 2.5013138 0.6544747 1.8926343 0.4773097 1.5863497 0.4906240  
[13] 0.4753530 1.2598371 1.8369900 2.3737864 3.3207072 6.2543970  
[19] 2.7399221 6.3944875 0.1074275 1.7145888 0.5902163 3.2306668  
[25] 1.4803313 3.0973267 2.7724045 3.9121375 2.7363755 2.9247515  
[31] 10.9627439

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



The ppoints argument (stands for probability points) in qchisq() function is used to generate n probability values based on which the chi-squared quantile values can be evaluated. ppoints() generate a vector of length of n of values between 0 and 1 which is calculated using formula (rank-0.5)/n where rank is the order of each observation.

The QQ plot plots the chi-squared quantiles along x axis and the ordered Mahalanobis distances along y-axis. The QQ plot for trees data shows reasonably a linear fit (except one or two points). Since linearity is subjective particularly with graphics, let us check the normality of the data using a formal test.

# Checking multivariate normality using Multivariate Shapiro test

Below we perform Multivariate Shapiro test using mshapiro.test() function from “mvnormtest” package. Note that data needs to be transposed in order to use in mshapiro.test() function.

#install.packages("mvnormtest")  
library(mvnormtest)  
mshapiro.test(t(trees))

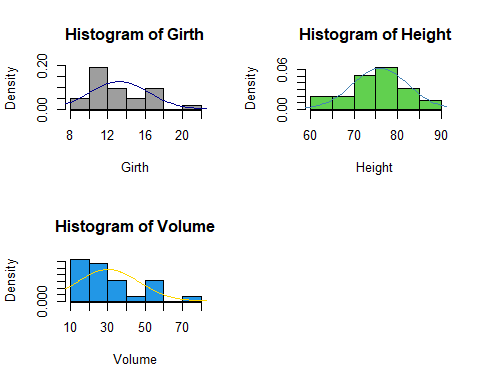
Shapiro-Wilk normality test  
  
data: Z  
W = 0.84615, p-value = 0.00042

The multivariate Shapiro-Wilk p-value is significant. Therefore, we reject the null hypothesis and conclude that the data is not coming from a p-variate normal distribution. Next, we may want to use univariate analysis to see which variables fail to be normally distributed.

## Checking univariate normality: Graphical <tool:Drawing> histograms

We draw histogram for each variable and superimpose a normal curve with parameters estimated as , and .

attach(trees)  
par(mfrow=c(2,2)) #arrange multiple plots in the same plotting space  
  
hist(Girth, prob = T, col=8) #drawing a relative freq. histogram  
x=seq(1, 30, by=.5)  
lines(x, dnorm(x,mean(Girth), sd(Girth)), col="darkblue",lwd=0.8)  
  
  
hist(Height, prob = T, col=11)  
x=seq(55, 95, 0.5)  
lines(x, dnorm(x, mean(Height), sd(Height)), col="steelblue")  
  
  
hist(Volume, prob = T, col=12)  
x=seq(5, 85, 0.5)  
lines(x, dnorm(x, mean(Volume), sd(Volume)), col="gold")



From the histograms, we see that the distribution of Girth and Height variables seem to fit reasonably well a normal curve, however, the distribution of the Volume variable appears to depart from normality and shows a right skewed distribution.

Below we apply Shapiro-Wilk test for checking normality of Girth, Height and Volume variables. What is our observation based on the p-value?

## Formal test for checking normality: Shapiro-Wilk test

shapiro.test(Girth)

Shapiro-Wilk normality test  
  
data: Girth  
W = 0.94117, p-value = 0.08893

shapiro.test(Height)

Shapiro-Wilk normality test  
  
data: Height  
W = 0.96545, p-value = 0.4034

shapiro.test(Volume)

Shapiro-Wilk normality test  
  
data: Volume  
W = 0.88757, p-value = 0.003579

detach(trees)

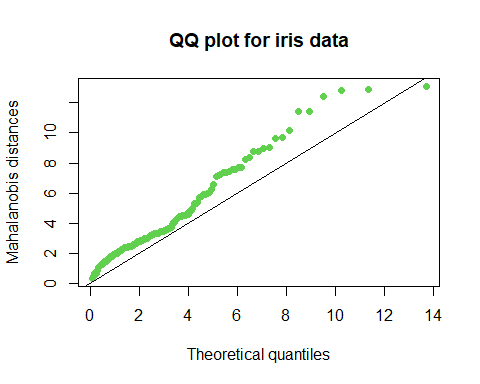
## Example 2: Checking Multivariate Normality for Iris data

## Drawing chi-squared QQ plot

attach(iris)  
head(iris)

Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
1 5.1 3.5 1.4 0.2 setosa  
2 4.9 3.0 1.4 0.2 setosa  
3 4.7 3.2 1.3 0.2 setosa  
4 4.6 3.1 1.5 0.2 setosa  
5 5.0 3.6 1.4 0.2 setosa  
6 5.4 3.9 1.7 0.4 setosa

new.iris=iris[, 1:4]  
  
s.mean=colMeans(new.iris)  
s.cov=cov(new.iris)  
  
  
n=nrow(iris)  
mhd.iris=mahalanobis(new.iris, s.mean, s.cov)  
  
theo.quan=qchisq(ppoints(n),3) #df=number of variables = 4  
qqplot(theo.quan, mhd.iris, xlab="Theoretical quantiles", ylab="Mahalanobis distances", pch=16, col=3, main="QQ plot for iris data")  
abline(0,1) #drawing reference line



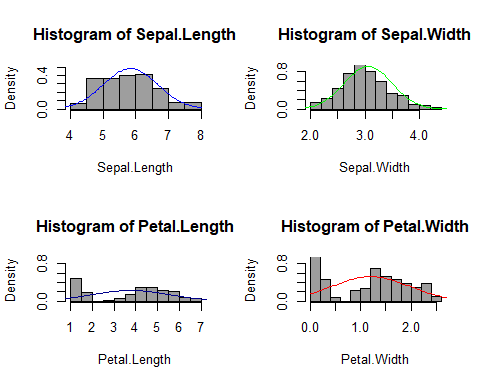
## Checking Multivariate normality using Shapiro-Wilk test

library(mvnormtest)  
mshapiro.test(t(iris[,1:4]))

Shapiro-Wilk normality test  
  
data: Z  
W = 0.97935, p-value = 0.02342

## Checking Univariate normality with histograms

par(mfrow=c(2,2))  
hist(Sepal.Length, prob = T, col=8, ylim=c(0,0.5)) #drawing a relative freq. histogram  
x=seq(2, 8, by=.05)  
lines(x, dnorm(x,mean(Sepal.Length), sd(Sepal.Length)), col="blue",lwd=0.8)  
  
hist(Sepal.Width, prob = T, col=8, ylim=c(0,0.9)) #drawing a relative freq. histogram  
x=seq(1, 5, by=.05)  
lines(x, dnorm(x,mean(Sepal.Width), sd(Sepal.Width)), col="green",lwd=0.8)  
  
hist(Petal.Length, prob = T, col=8, ylim=c(0,0.9)) #drawing a relative freq. histogram  
x=seq(0, 8, by=.05)  
lines(x, dnorm(x,mean(Petal.Length), sd(Petal.Length)), col="darkblue",lwd=0.8)  
  
  
hist(Petal.Width, prob = T, col=8, ylim=c(0,0.9)) #drawing a relative freq. histogram  
x=seq(0, 3, by=.05)  
lines(x, dnorm(x,mean(Petal.Width), sd(Petal.Width)), col="red",lwd=0.8)



## Checking univariate normality using Shapiro-Wilk test

shapiro.test(Sepal.Length)

Shapiro-Wilk normality test  
  
data: Sepal.Length  
W = 0.97609, p-value = 0.01018

shapiro.test(Sepal.Width)

Shapiro-Wilk normality test  
  
data: Sepal.Width  
W = 0.98492, p-value = 0.1012

shapiro.test(Petal.Length)

Shapiro-Wilk normality test  
  
data: Petal.Length  
W = 0.87627, p-value = 7.412e-10

shapiro.test(Petal.Width)

Shapiro-Wilk normality test  
  
data: Petal.Width  
W = 0.90183, p-value = 1.68e-08

detach(iris)