**1. Please check the normality of and of the banknotes data.**

**Sol.**

Suppose

**Univariate normality:**

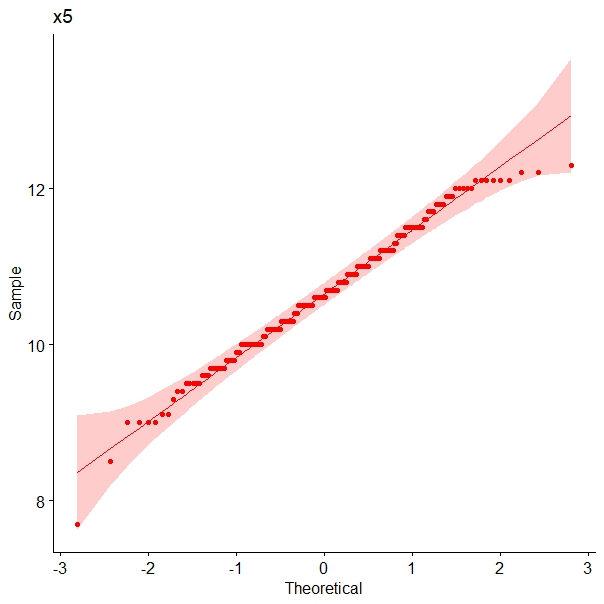
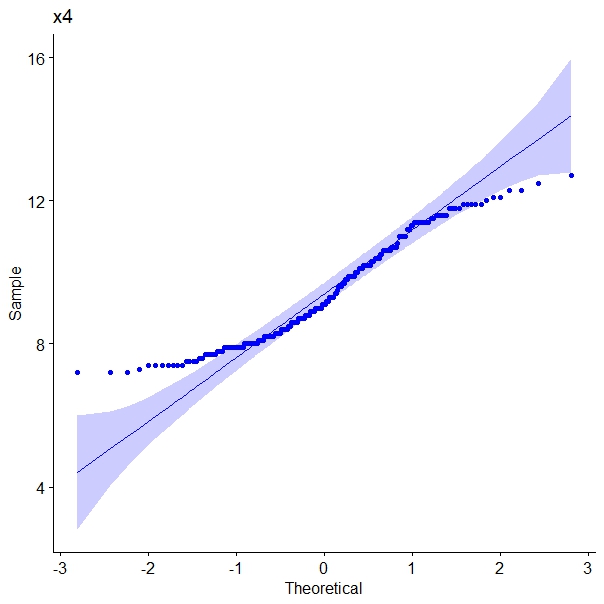


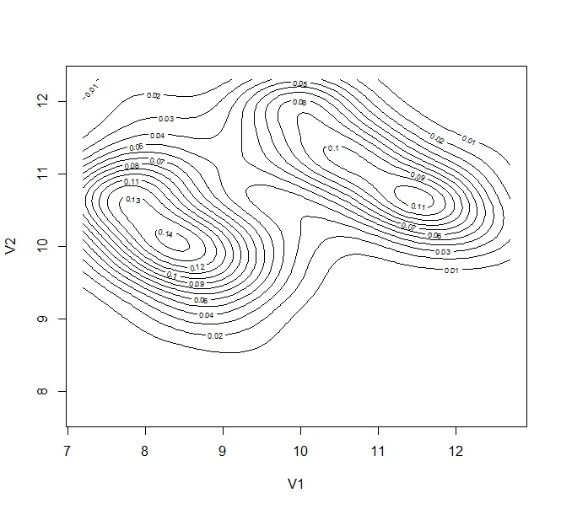
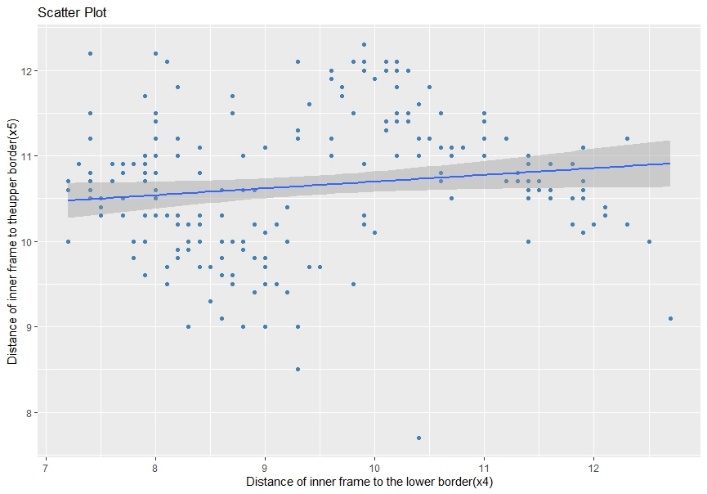
Figure 1: univariate qq-plot for Figure 2: univariate qq-plot for

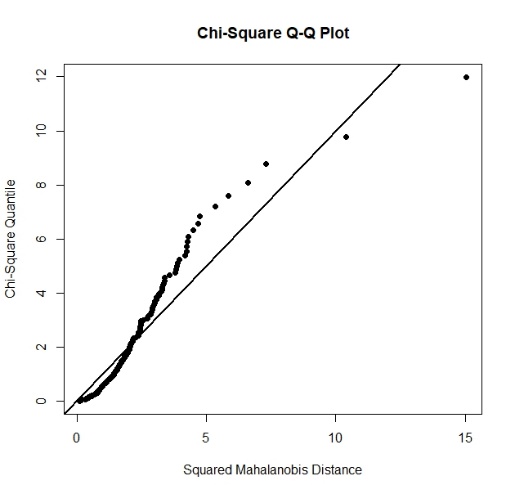
According to the Q-Q plot of above, the sample quantile values are plotted on the x-axis, and the corresponding quantile values of the dataset are plotted on the y-axis. We can see that the points lie nearly along the 45-degrees reference line, it means that the are approximately normal. On the other hand, we can see that the points of on both endpoints are deviate from the reference line, it means that the data of is not normally distributed.

Furthermore, we can also use the Shapiro-Wilk test to detect if each distribution is exact normality. From the results of r code, for the data of , have p-value equal 6.354e-07. When p-value<0.05, we have sufficient evidence to say that the sample data does not come from a population that is normally distributed. For the data of , it has a p-value equal 0.05856, so the distribution of the data are not significantly different from normal distribution.

In conclusion, if we test these two variables separately, is probably not normally distributed in the population, is normally distributed in the population.

**Bivariate normality:**



Figure 3: scatterplot for versus Figure 4: contour of the density of

and

Figure 5: Chi-square Q-Q plot for and

從圖三散佈圖可以看出，和分別散落在兩邊，且從圖四可明顯看出有兩個橢圓，代表兩變數有明顯差異。若將圖三與圖四重疊，則能對應出每筆資料分佈的疏密程度，兩群體中間被明顯地區分。除了用單變量的qq plot與scatter plot，檢查多變量是否為常態還需要經過與chi-squared plot比對，從圖五可以看出全部的點連線並沒有構成一條直線，右上角的兩筆資料也與其他資料有明顯變化，因此無法稱之為多變量常態分配。

**2. Detect the outliers of and of the banknotes data.**

**Sol.**

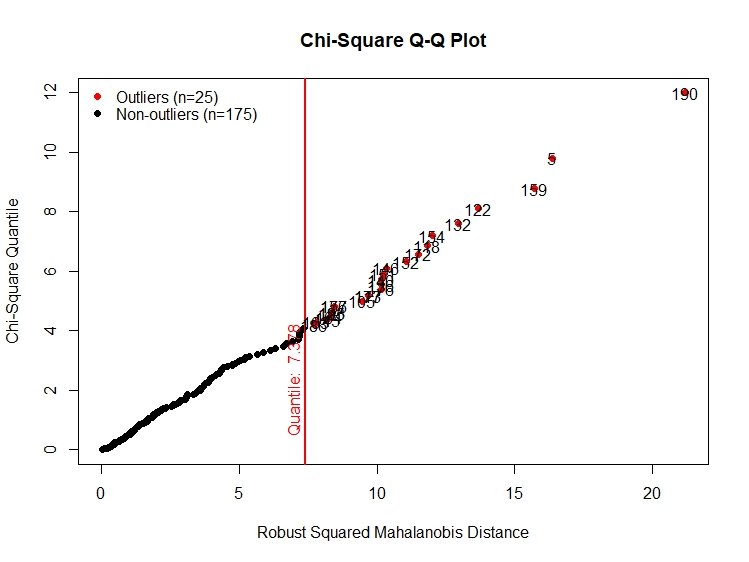


Figure 6: Multivariate outlier detection based on Mahalanobis distance

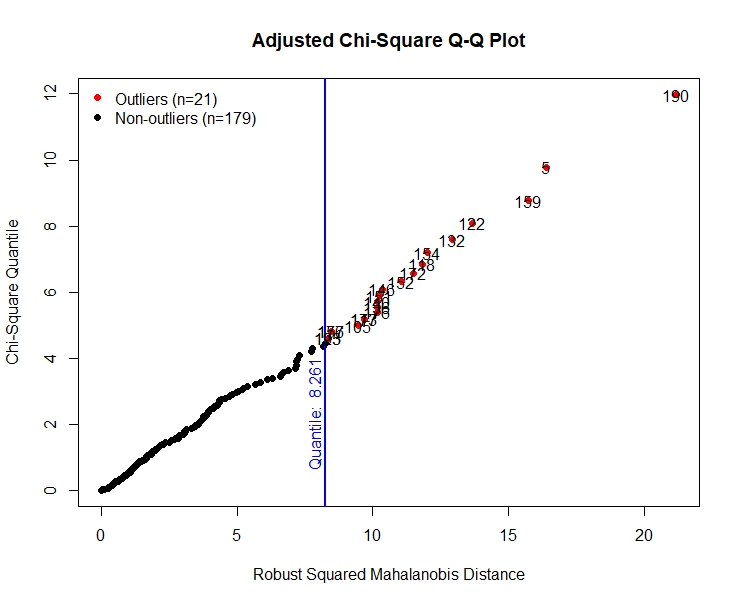


Figure 7: Multivariate outlier detection based on adjusted-Mahalanobis distance

從圖一與圖二中可以看出個別變數中異常大或異常小的離群值，然而在多變量中需要考慮到更多面向，多變量的重點在討論兩兩變數之間的關係，藉由以下幾種方法可以互相驗證找到真實的離群值。

在圖三scatter plot中我們可以看到兩個異常點，一個位於圖表中下方、一個位於圖表右邊中間，兩者都離群體很遠，但圖表中下方的點雖然是的離群值，但對的資料中，他不屬於離群值；另一個右邊中間的點則是反過來，它對來說屬於正常分布範圍內，對而言卻是一筆離群值。

圖五可以看出右上方兩點離其他點都很遠，代表它們的Mahalanobis Distance很大，可以認定為離群值。另外，藉由MVN test可以得到圖六與圖七的離群值結果，我們可以合理推測圖五中距離群體最遠的點，即為圖六、圖七的編號190的點(第190筆資料)，它的Mahalanobis Distance=21.192，表示該點到群體中心的距離為21.192；而第二遙遠的則為第5筆資料，它的Mahalanobis Distance=16.381。觀察全部的離群值，我們可以發現大部分的離群值都出現在假鈔的資料(第101~200筆)。由單變量分析計算z-scores也可以應證離群值的資料，其z分數皆落在個標準差之外。

Observation Mahalanobis Distance Outlier

190 190 21.192 TRUE

5 5 16.381 TRUE

159 159 15.719 TRUE

122 122 13.662 TRUE

132 132 12.941 TRUE

154 154 12.003 TRUE

118 118 11.833 TRUE

172 172 11.503 TRUE

152 152 11.047 TRUE

146 146 10.337 TRUE

151 151 10.233 TRUE

140 140 10.168 TRUE

136 136 10.161 TRUE

176 176 10.160 TRUE

117 117 9.690 TRUE

173 173 9.690 TRUE

105 105 9.467 TRUE

156 156 8.458 TRUE

177 177 8.458 TRUE

123 123 8.365 TRUE

195 195 8.365 TRUE

R code:

# clear all variables

rm(list = ls(all = TRUE))

graphics.off()

# install and load packages

install.packages("dplyr")

install.packages("ggpubr")

# import data

bank <- read.table("C:/Users/user/Desktop/多變量11101/MVA-ToDo-master/QID-948-MVApcabankr/bank2.dat")

x4 <- bank$V4

x5 <- bank$V5

x45 <- as.data.frame(cbind(bank[, 4],bank[, 5]))

# Shapiro-Wilk test

# p-value>0.05:不拒絕虛無假設(符合常態分佈)

# p-value<0.05:拒絕虛無假設(不符合常態分佈)

shapiro.test(x4)

shapiro.test(x5)

# Q-Q plot

library(ggpubr)

ggqqplot(x4,color = 'blue', main = 'x4')

ggqqplot(x5,color = 'red', main = 'x5')

# scatter plot with linear fit line

ggplot(bank,

aes(x = x4, y = x5)) +

geom\_point(color= "steelblue") +

geom\_smooth(formula = y ~ x, method = "lm") +

labs(y = "Distance of inner frame to theupper border(x5)",

x = "Distance of inner frame to the lower border(x4)",

title = "Scatter Plot")

# chi-square plot

# install and load packages

install.packages("mvoutlier")

install.packages("sgeostat")

library(mvoutlier)

# draw the plot

chisq.plot(x45, quan=1/2, ask=TRUE)

# detecting outliers

# z score

install.packages("outliers")

library(outliers)

z.scores <- x45 %>% scores(type = "z")

# MVN

install.packages("MVN")

library(MVN)

result\_uni <- mvn(x45, mvnTest = "mardia", univariateTest = "SW", showOutliers = TRUE)

result\_multi <- mvn(x45, mvnTest = "mardia", multivariateOutlierMethod = "quan", showOutliers = TRUE)

result\_multi <- mvn(x45, mvnTest = "mardia", multivariateOutlierMethod = "adj", showOutliers = TRUE)

result\_uniqq <- mvn(x45, mvnTest = "mardia", univariatePlot = "qqplot")

result\_multiqq <- mvn(x45, mvnTest = "mardia", multivariatePlot = "qq")

result\_muticon <- mvn(x45, mvnTest = "mardia", multivariatePlot = "contour")