Exercise 7.6 In the U.S. companies data set, test the equality of means between the energy and manufacturing sectors, taking the full vector of observations x1 to *x*6. Derive the simultaneous confidence intervals for the differences.

The data set consists of measurements for 79 U.S. companies. The abbreviations in this section are as follows:

*x*1: A Assets (USD),

*x*2: S Sales (USD),

*x*3: MV Market Value (USD),

*x*4: P Profits (USD),

*x*5: CF Cash Flow (USD), and

*x*6: E Employees.

Sol.

根據題目假設檢定 $H_0: \mu_1 - \mu_2 = \delta$; $H_1: \mu_1 - \mu_2 \neq \delta$

首先假設$X_{i1}(energy)$和$X_{j2}(manufacturing)$服從多變量常態，且$\Sigma_1 = \Sigma_2$。

$X_{i1} \sim N_p(\mu_1, \Sigma), i = 1, \ldots, 15 \quad X_{j2} \sim N_p(\mu_2, \Sigma), j = 1, \ldots, 10$ *變數間兩兩互相獨立*。

設$n_1 = 15, n_2 = 10, p = 6$

經由 r 計算得：

$$\overline{x_1} = \begin{bmatrix} 4084 \\ 2580.467 \\ 1299.933 \\ 156.527 \\ 334.893 \\ 7.007 \end{bmatrix}, \qquad \overline{x_2} = \begin{bmatrix} 4307.2 \\ 4925.2 \\ 1710.2 \\ 36.29 \\ 202.99 \\ 48.39 \end{bmatrix}$$

$S_1 = \frac{n_1 - 1}{n_1} * \Sigma =$

$$\begin{bmatrix} 16634749.47 & 12409636.67 & 4146592.53 & 4554655.39 & 1348646.46 & 30291.51 \\ 12409636.67 & 13747417.45 & 2295973.3 & 280118.62 & 1192122.98 & 24615.66 \\ 4146592.53 & 2295973.3 & 1210407 & 125379.91 & 302657.99 & 7295.56 \\ 4554655.39 & 280118.62 & 125379.91 & 20796.574 & 43209.47 & 775.81 \\ 1348646.46 & 1192122.98 & 302657.99 & 43209.47 & 128966.99 & 2511.19 \\ 30291.51 & 24615.66 & 7295.56 & 775.81 & 2511.19 & 57.61 \end{bmatrix}$$

, $S_2 = \frac{n_2 - 1}{n_2} * \Sigma =$

$$\begin{bmatrix} 12247662.8 & 11425397.76 & 3805597.2 & 105022.7 & 375779.5 & 134516.2 \\ 11425397.8 & 15111585.16 & 4725907.26 & 4347.03 & 386676.10 & 183333.67 \\ 3805597.2 & 4725907.26 & 2457676.96 & 229788.45 & 389628.45 & 67564.02 \\ 105022.7 & 4347.03 & 229788.45 & 85696.86 & 85455.00 & 2687.31 \\ 375779.5 & 386676.10 & 389628.45 & 85455.00 & 100311.34 & 7774.78 \\ 134516.2 & 183333.67 & 67564.02 & 2687.31 & 7774.78 & 2423.66 \end{bmatrix}$$

因此，由公式 $S=\frac{n_1S_1+n_2S_2}{n_1+n_2}$ ，得

$$S=\begin{bmatrix}14879914.78 & 12015941.1 & 4010194.38 & 314802.33 & 959499.68 & 71981.39\\ 12015941.1 & 14293084.53 & 3267946.88 & 169809.99 & 869944.23 & 88102.86\\ 4010194.38 & 3267946.88 & 1709314.98 & 167143.43 & 337446.18 & 31402.95\\ 314802.33 & 169809.99 & 167143.43 & 46756.69 & 60107.68 & 1540.407\\ 959499.68 & 869944.23 & 337446.18 & 60107.68 & 117504.73 & 4616.62\\ 71981.39 & 88102.86 & 31402.95 & 1540.407 & 4616.62 & 1004.03\end{bmatrix}$$

$H_0: \delta = 0$

拒絕域：

$$\frac{n_1n_2(n_1+n_2-p-1)}{p(n_1+n_2)^2}(\overline{x_1}-\overline{x_2})^T S^{-1}(\overline{x_1}-\overline{x_2}) \geq F_{1-\alpha;\, p, n_1+n_2-p-1}$$

$$\Rightarrow \frac{15*10(25-6-1)}{6*25^2}\begin{bmatrix}-223.2 & -2344.7 & -410.3 & 120.2 & 131.9 & -41.4\end{bmatrix}$$

$$\begin{bmatrix}14879914.78 & 12015941.1 & 4010194.38 & 314802.33 & 959499.68 & 71981.39\\ 12015941.1 & 14293084.53 & 3267946.88 & 169809.99 & 869944.23 & 88102.86\\ 4010194.38 & 3267946.88 & 1709314.98 & 167143.43 & 337446.18 & 31402.95\\ 314802.33 & 169809.99 & 167143.43 & 46756.69 & 60107.68 & 1540.407\\ 959499.68 & 869944.23 & 337446.18 & 60107.68 & 117504.73 & 4616.62\\ 71981.39 & 88102.86 & 31402.95 & 1540.407 & 4616.62 & 1004.03\end{bmatrix}^{-1}\begin{bmatrix}-223.2\\ -2344.7\\ -410.3\\ 120.2\\ 131.9\\ -41.4\end{bmatrix}$$

$\geq F_{6,18}$

$\Rightarrow F = 2.1526,\ F_{0.95;6,18} = 2.6613$

因為 $F < \ F_{0.95;6,18}$，不拒絕H_0，表示沒有足夠證據證明 $\mu_1 \neq \mu_2$。

但若以 90%的信賴水準評估，則結果為 $F = 2.1526 > F_{0.90;6,18} = 2.1296$，拒絕 H_0，表示有足夠證據證明μ_1 *與*μ_2之間存在不同。

Confidence Region:

$$a^T\delta \in a^T(\overline{x_1}-\overline{x_2}) \pm \sqrt{\frac{p(n_1+n_2)^2}{n_1n_2(n_1+n_2-p-1)}F_{1-\alpha;\, p,n_1+n_2-p-1}a^TSa}, a=\begin{bmatrix}1\\ -1\\ 0\\ 0\\ 0\\ 0\end{bmatrix}$$

$$\Rightarrow -7639.40 \leq \mu_{1a}-\mu_{2a} \leq 7192.997$$
$$\Rightarrow -9613.22 \leq \mu_{1s}-\mu_{2s} \leq 4923.75$$
$$\Rightarrow -2923.84 \leq \mu_{1mv}-\mu_{2mv} \leq 2103.31$$
$$\Rightarrow -295.49 \leq \mu_{1p}-\mu_{2p} \leq 535.96$$
$$\Rightarrow -527.13 \leq \mu_{1cf}-\mu_{2cf} \leq 790.94$$
$$\Rightarrow -102.30 \leq \mu_{1e}-\mu_{2e} \leq 19.54$$

根據信賴區間估計公式得六種變數在 energy 與 manufacturing 兩種產業的公司間皆無足夠證據說明兩者存在差異性，因此不拒絕H_0，μ_1與μ_2可能不獨立。

透過 mvn 套件也可以檢驗出每個變數個別 p-value 與多變量的 p-value 皆小於 0.05，亦即該資料不符合多變量及單變量的常態分佈。

```
$multivariateNormality
             Test        Statistic               p value Result
1 Mardia Skewness 242.233307803641 5.21767710767055e-25     NO
2 Mardia Kurtosis 9.74673862485956                     0     NO
3             MVN             <NA>                  <NA>     NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    V2        0.7720    1e-04     NO
2 Shapiro-Wilk    V3        0.7016   <0.001     NO
3 Shapiro-Wilk    V4        0.8191    5e-04     NO
4 Shapiro-Wilk    V5        0.7827    1e-04     NO
5 Shapiro-Wilk    V6        0.8260    6e-04     NO
6 Shapiro-Wilk    V7        0.5853   <0.001     NO
```

R Code:

```
# clear variables and close windows
rm(list = ls(all = TRUE))
graphics.off()

# Load data
x = read.table("C:/Users/user/Desktop/多變量 11101/MVA-ToDo-master/QID-1659-MVAsimcidif/uscomp2.dat")
y = data.frame(x)

# Create subsets for Energy and Manufacturing
yE = subset(y, y$V8 == "Energy")
yM = subset(y, y$V8 == "Manufacturing")

# Calculate means of groups
exE = cbind(apply(yE[, 2:7], 2, mean)) # 1:by row;2:by column
```

```r
exM = cbind(apply(yM[, 2:7], 2, mean))
# https://kemushi54.github.io/R-basic/apply_family.html

# Estimating variance of the groups observations within the groups and overall
nE    = length(yE[, 1])
nM    = length(yM[, 1])
n     = nE + nM

# number of groups
p     = length(exE)

sE    = ((nE - 1)/nE) * cov(yE[, 2:7]) # S1
sM    = ((nM - 1)/nM) * cov(yM[, 2:7]) # S2

s     = (nE * sE + nM * sM)/(nE + nM)
sinv  = solve(s)
k     = nE * nM * (n - p - 1)/(p * (n^2))

# Computing the test statistic
(f = k * t(exE - exM) %*% sinv %*% (exE - exM))

# Computing the critical value
(critvalue = qf(1 - 0.05, 6, 18))
# ALPHA=0.05 95%CI  右尾  IF f>critvalue  拒絕 HO
(critvalue = qf(1 - 0.1, 6, 18))

# Computes the simultaneous confidence intervals
deltau    = (exE - exM) + sqrt(qf(1 - 0.05, p, n - p - 1) * (1/k) * diag(s))
deltal    = (exE - exM) - sqrt(qf(1 - 0.05, p, n - p - 1) * (1/k) * diag(s))

(confit = cbind(deltal, deltau))

# extrapoints
library(MVN)
result_uni <- mvn(yEM, mvnTest = "mardia", univariateTest = "SW", showOutliers =
TRUE)
result_multi <- mvn(yEM, mvnTest = "mardia", multivariateOutlierMethod = "quan",
showOutliers = TRUE)
```