GR5223 - HW1

MJS2364

7 février 2018

Problem 4

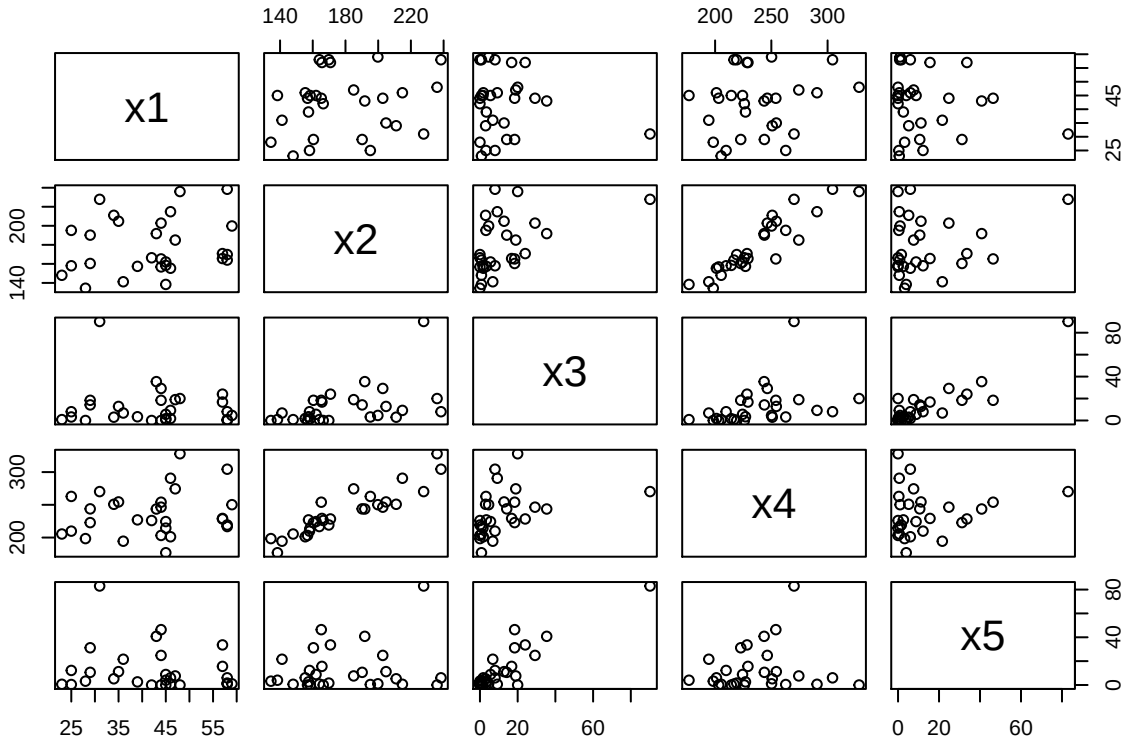
The file MS_Data.csv, in the Data folder on Courseworks, contains data related to response stimuli for people suffering from a multiple-sclerosis-caused visual pathology. Two different visual stimuli (S1 and S2) produced responses in both the left eye (L) and the right eye (R) of each of the n = 29 subjects. The values recorded in the data file include:

Variable Definition Description x1 Age Subject's Age x2 (S1L +S1R) Total response of both eyes to stimulus 1 x3 jS1L S1Rj Difference between responses of eyes to stimulus 1 x4 (S2L + S2R) Total response of both eyes to stimulus 2 x5 jS2L S2Rj Difference between responses of eyes to stimulus 2

a) Construct the scatterplot matrix, and comment on the relationship between variables x2 and x4.

```
data <- read.csv("MS_Data.csv", as.is = TRUE)
data <- as.matrix(data)

library(graphics)
pairs(data)
```



We clearly see that X2 and X4 have a linear association so collinearity issues may arise.

b) Compute the sample mean vector, covariance matrix, and correlation matrix using statistical software.

```
# Calculates the sample mean vector
colMeans(data)
```

```
##        x1        x2        x3        x4        x5
##  42.06897 178.26897  12.27586 236.93103  13.08276
```

```
# Calculates the covariance matrix
cov(data)
```

```
##           x1        x2       x3         x4        x5
## x1 121.13793  52.79507 -20.2197   68.13350 -29.82020
## x2  52.79507 844.68079 244.4632  912.41493 106.76409
## x3 -20.21970 244.46315 317.2640  232.36542 297.31921
## x4  68.13350 912.41493 232.3654 1180.03222  81.09734
## x5 -29.82020 106.76409 297.3192   81.09734 351.04719
```

```
# Calculates the correlation matrix
cor(data)
```

```
##            x1        x2         x3        x4         x5
## x1  1.0000000 0.1650468 -0.1031393 0.1802078 -0.1446065
## x2  0.1650468 1.0000000  0.4722334 0.9139010  0.1960632
```

```
## x3 -0.1031393 0.4722334  1.0000000 0.3797643  0.8909017
## x4  0.1802078 0.9139010  0.3797643 1.0000000  0.1260019
## x5 -0.1446065 0.1960632  0.8909017 0.1260019  1.0000000
```

c) Compute the sample mean vector, covariance matrix, and correlation matrix using matrix algebra.

```
n <- nrow(data)

# Calculates the sample mean vector
X_bar <- (1/n)*t(data) %*% rep(1, time=n)
print(as.vector(X_bar))
```

```
## [1]  42.06897 178.26897  12.27586 236.93103  13.08276
```

```
# Calculates the covariance matrix
S <- 1/(n-1)*t(data) %*% (diag(n) - (1/n)*rep(1,time=n) %*% t(rep(1,time=n))) %*% data
print(S)
```

```
##           x1        x2       x3         x4        x5
## x1 121.13793  52.79507 -20.2197   68.13350 -29.82020
## x2  52.79507 844.68079 244.4632  912.41493 106.76409
## x3 -20.21970 244.46315 317.2640  232.36542 297.31921
## x4  68.13350 912.41493 232.3654 1180.03222  81.09734
## x5 -29.82020 106.76409 297.3192   81.09734 351.04719
```

```
# Calculates the correlation matrix
R <- diag(1/sqrt(diag(S))) %*% (S) %*% diag(1/sqrt(diag(S)))
print(R)
```

```
##            [,1]      [,2]       [,3]      [,4]       [,5]
## [1,]  1.0000000 0.1650468 -0.1031393 0.1802078 -0.1446065
## [2,]  0.1650468 1.0000000  0.4722334 0.9139010  0.1960632
## [3,] -0.1031393 0.4722334  1.0000000 0.3797643  0.8909017
## [4,]  0.1802078 0.9139010  0.3797643 1.0000000  0.1260019
## [5,] -0.1446065 0.1960632  0.8909017 0.1260019  1.0000000
```

Problem 5

The experiment described in the previous exercise was also conducted on a control group of n = 69 subjects not suffering from the disease. Results are contained in the data file Non_MS.csv, also available on Courseworks. For this exercise we will use variables 2 through 5 only.

a) Suppose researchers are only interested in two variables: total response to both stimuli combined, and total absolute difference between left/right responses. For the ith subject define yi1 = xi2 + xi4 and yi2 = xi3 + xi5 for i = 1, . . . , n=69

i. Find the matrix A for which Y = XA' defines the new data matrix.

```
# Reads the new data
data2 <- read.csv("Non_MS.csv", as.is = TRUE)
data2 <- as.matrix(data2[,2:5])

# Creates the relevant matrix A and computes the new data matrix Y
A <- matrix(c(1,0,1,0,0,1,0,1), ncol = 4, byrow = TRUE)
Y <- data2 %*% t(A)

print(A)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    1    0
## [2,]    0    1    0    1
```

```
head(Y)
```

```
##       [,1] [,2]
## [1,] 350.4  1.6
## [2,] 318.8  2.0
## [3,] 330.4  0.8
## [4,] 338.4  3.2
## [5,] 366.4  0.0
## [6,] 323.2  1.6
```

ii. Compute the sample mean vector Y_bar and verify that Y_bar = A*X_bar.

```
# Calculates Y_bar and A*X_bar
Y_bar <- colMeans(Y)
AX_bar <- as.vector(A %*% colMeans(data2))

print(Y_bar)
```

```
## [1] 342.892754   3.182609
```

```
print(AX_bar)
```

```
## [1] 342.892754   3.182609
```

Since both results are the same, then we verified that Y_bar = A*X_bar

iii. Compute the sample covariance matrix Sy and verify that Sy = A*(Sx)*A'.

```
# Calculates the sample covariance matrix of Y
cov(Y)
```

```
##            [,1]     [,2]
## [1,] 511.093623 8.394578
```

```
## [2,]   8.394578 5.163223
```

```
# Calculates Sy = A*(Sx)*A'
Sy <- A %*% cov(data2) %*% t(A)
print(Sy)
```

```
##            [,1]     [,2]
## [1,] 511.093623 8.394578
## [2,]   8.394578 5.163223
```

Since both results are the same, then we verified that Sy = A*(Sx)*A'

b) Working again with the original 69 x 4 data matrix X (4 columns as we are using variables 2 through 5 only):

i. Find the square root matrix S^(1/2) of the sample covariance matrix.

```
# Calculates the eigenvalues and eigenvectors of X
evalues <- eigen(cov(data2))$values
evectors <- eigen(cov(data2))$vectors

# Calculates the square root diagonal matrix
L <- diag(sqrt(evalues))

# Calculates the square root matrix S^(1/2) of the sample covariance matrix
S_05 <- evectors %*% L %*% t(evectors)
print(S_05)
```

```
##           [,1]      [,2]       [,3]      [,4]
## [1,] 9.4376328 0.1081300  4.8162819 0.1325607
## [2,] 0.1081300 1.3232790  0.1205752 0.1665295
## [3,] 4.8162819 0.1205752 12.7280628 0.1188337
## [4,] 0.1325607 0.1665295  0.1188337 1.5152685
```

ii. Define a new data matrix by Z=(X-1n*X__bar')*S^(-0.5). Find the mean vector and covariance matrix of Z and explain why they take that form.

```
# Calculates the inverse square root covariance matrix
S_n05 <- solve(S_05)


# Calculates the multivariate standardized data matrix Z
Z <- (data2 - rep(1,nrow(data2)) %*% t(colMeans(data2))) %*% S_n05


colMeans(Z)
```

```
## [1] -1.844941e-15  4.409717e-17  1.083410e-15  3.597666e-17
```

```
cov(Z)
```

```
##               [,1]          [,2]          [,3]          [,4]
## [1,]  1.000000e+00 -2.410618e-16 -2.865359e-16  1.562447e-17
## [2,] -2.410618e-16  1.000000e+00  2.056348e-15  4.502090e-17
```

```
## [3,] -2.865359e-16  2.056348e-15  1.000000e+00 -2.585750e-16
## [4,]  1.562447e-17  4.502090e-17 -2.585750e-16  1.000000e+00
```

This Mahalanobis transformation gives a standardized (centered and scaled), uncorrelated data matrix Z. Theoretically, we know that the average of standardized data along each explanatory variable will be equal to 0. Here, the means that we computed along each variable are indeed very close to 0 (between e-15 and e-17). Then, the covariance matrix of our standardized data is very close to the identity matrix (since e-15 and e-17 are approximately null). Thus the covariance between all pairs of distinct variables is approximately zero. This means that all explanatory variables are uncorrelated, which result is expected from the Mahalanobis transformation.