**MATH 4310** **Name:** Solution

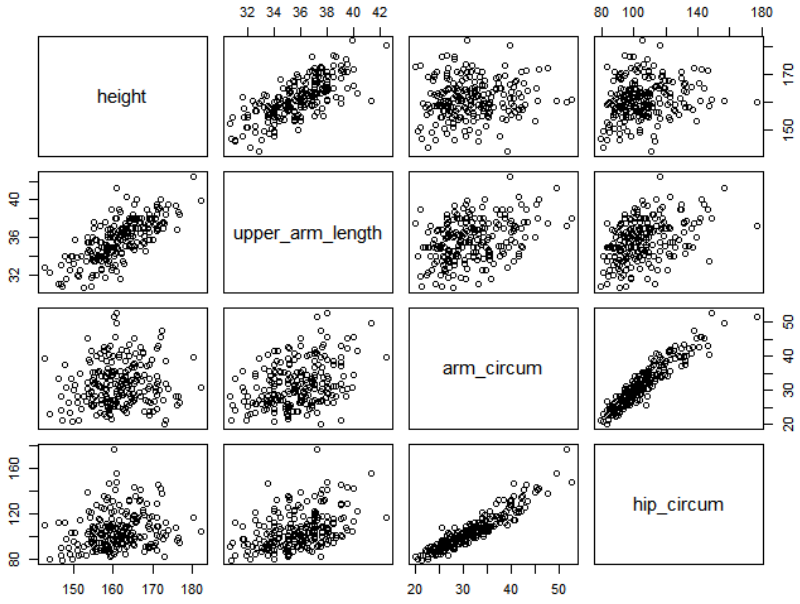
# Homework 10

**Problem.** We consider a data set extracted from the 2017-2018 US National Health and Nutrition Examination Survey consisting of 230 participants aged between 20-25 years. For each participant, data were collected about body measures to estimate the prevalence of overweight and obesity. Data in stored in bodydata.csv

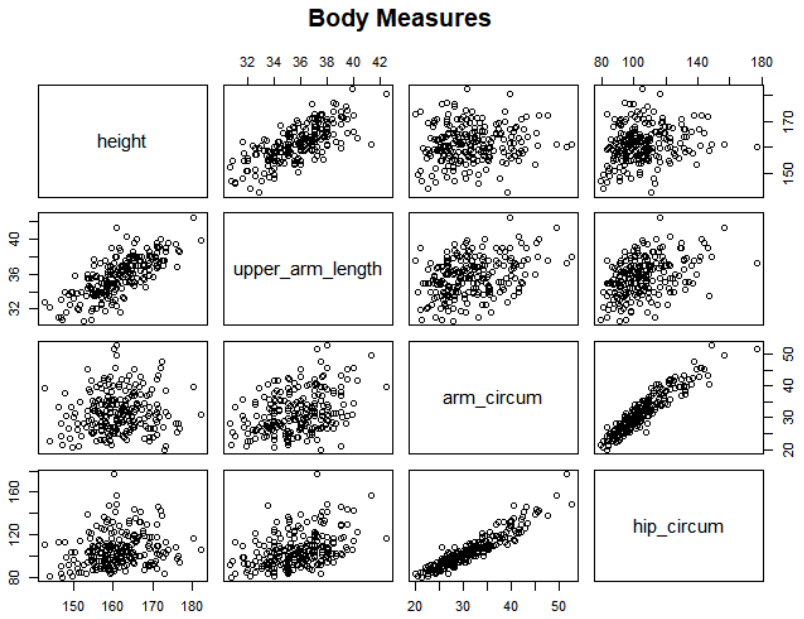
a) Draw a scatterplot of the data in pairs using the R command pair as in the notes.
b) Compute the correlation of the data matrix. Note: some data row have missing data. You will need to use this version of the command: cor(bodydata,use = "complete.obs")
c) Plot the correlation matrix using numbers as well as circles to display the size of correlation coefficients.
d) Apply hierarchical clustering as in the lectures on the correlation plot using 2 clusters.
e) Compute the p-values on the correlation matrix.
f) Analyze the results: which variables are strongly correlated (correlation coefficient > 0.7) to each other? Which variables are not statistically correlated (use alpha = 0.05)?

SOLUTION

```
> bodydata <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/bodydata.csv")
> head(bodydata)
  height upper_arm_length arm_circum hip_circum
1  158.4              36.0       26.5      101.1
2  164.7              38.1       30.5       97.4
3  156.9              34.0       28.5      101.7
4  158.1              35.0       22.2       88.7
5  158.2              35.0       32.0      100.3
6  162.0              34.4       32.7       99.3

> dim(bodydata)
[1] 230   4

> pairs(bodydata[c("height","hip_circum")])


> pairs(bodydata[c("height","upper_arm_length","arm_circum","hip_circum")])
```

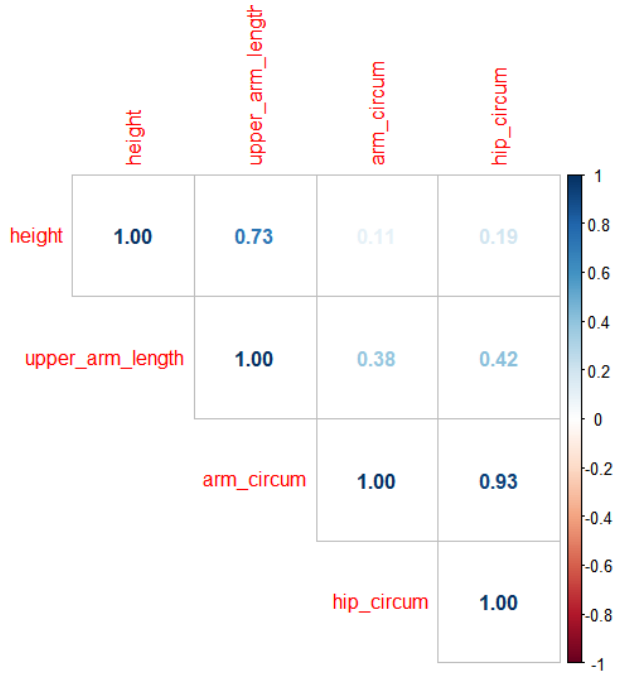


```
>plot(bodydata, main = "Body Measures")
```

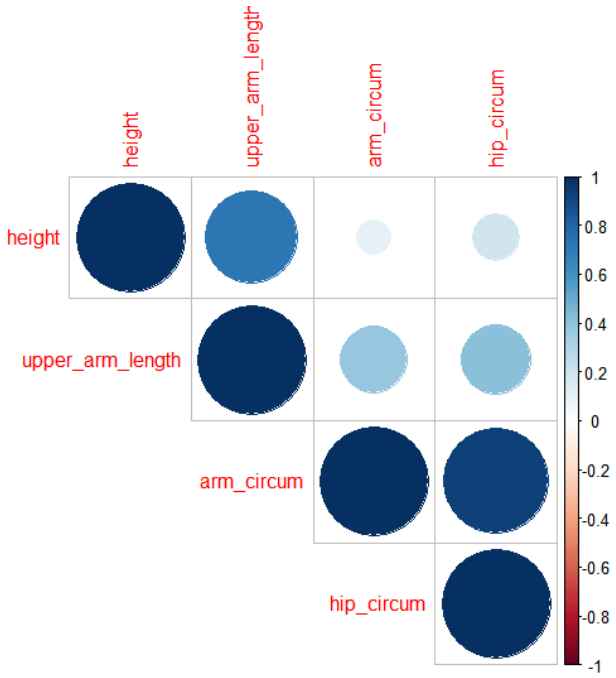



```
> cor(bodydata,use = "complete.obs")
                height upper_arm_length arm_circum hip_circum
height          1.0000000        0.7259228  0.1061935  0.1942494
upper_arm_length 0.7259228        1.0000000  0.3843140  0.4187889
arm_circum       0.1061935        0.3843140  1.0000000  0.9332575
hip_circum       0.1942494        0.4187889  0.9332575  1.0000000
```

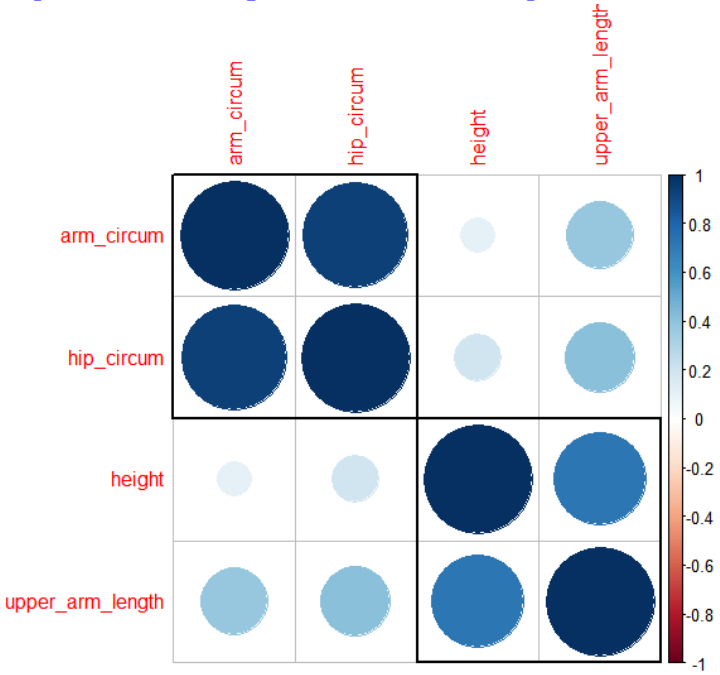
```
> corrplot(cor(bodydata,use = "complete.obs"),method = "number",type = "upper")
```



```
> corrplot(cor(bodydata,use = "complete.obs"),method = "circle",type = "upper")
```




```
> corrplot(cor(bodydata,use = "complete.obs"), order = "hclust", addrect = 2)
```




```
> X <- as.matrix(bodydata)
> res <-rcorr(X)
> round(res$P, 3)
                height upper_arm_length arm_circum hip_circum
height              NA                0      0.176      0.004
upper_arm_length 0.000               NA      0.000      0.000
arm_circum       0.176                0         NA      0.000
hip_circum       0.004                0      0.000         NA
```

**CONCLUSION:**

- **The variables arm-circum and hip_circum are strongly correlated; so are the variables height and upper_arm_length.**
- **The variables height and arm_circum are not statistically correlated.**