ST437/537 – HW #01- Solution

ZHOU LAN

Due date: January 17, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework before coming to class. Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.

Problems

Soleve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

1. (10 points) Let $\boldsymbol{x} = (5, 1, 3)^T$ and $\boldsymbol{y} = (-1, 3, 1)^T$ be two 3×1 column vectors.

a. Find the length of $\boldsymbol{x}$ and $\boldsymbol{y}$ (Do this by hand)

$||X|| = \sqrt{5^2 + 1^2 + 3^2} = 5.91, ||Y|| = \sqrt{(-1)^2 + 3^2 + 1^2} = 3.31$

```
x = c(5,1,3)
y = c(-1,3,1)

# length of x
sqrt( t(x) %*% x )
```

```
##          [,1]
## [1,] 5.91608
```

```
# length of y
sqrt( t(y) %*% y )
```

```
##          [,1]
## [1,] 3.316625
```

b. Find $\boldsymbol{x}^T\boldsymbol{y}$ (Do this by hand)

$\boldsymbol{x}^T\boldsymbol{y} = 5 \times (-1) + 1 \times 3 + 3 \times 1 = 1$

```
t(x)%*%y
```

```
##      [,1]
## [1,]    1
```

c. Are $\boldsymbol{x}$ and $\boldsymbol{y}$ orthogonal? Explain.

No, because $\boldsymbol{x}^T\boldsymbol{y}$ is not equal to 0.

d. Repeat (a) and (b) using R. Provide code and output.

See the parts above.

2. (10 points) Read the section about linear combinations in the "Multivariate summary statistics" lecture notes, and then answer the following questions.

You are given a random vector $\boldsymbol{X} = [X_1, X_2, X_3, X_4]^T$ with the mean vector $\boldsymbol{\mu} = [4, 3, 2, 1]^T$, and the variance-covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix}.$$

Also define the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -1 \end{bmatrix}.$$

a. Find $E(\boldsymbol{X})$ and $E(\boldsymbol{AX})$.

We have $E(\boldsymbol{X}) = [4, 3, 2, 1]^T$ and

$$E(\boldsymbol{AX}) = \boldsymbol{A}E(\boldsymbol{X}) = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -1 \end{bmatrix} [4, 3, 2, 1]^T = [10, 11, 0, 3]^T.$$

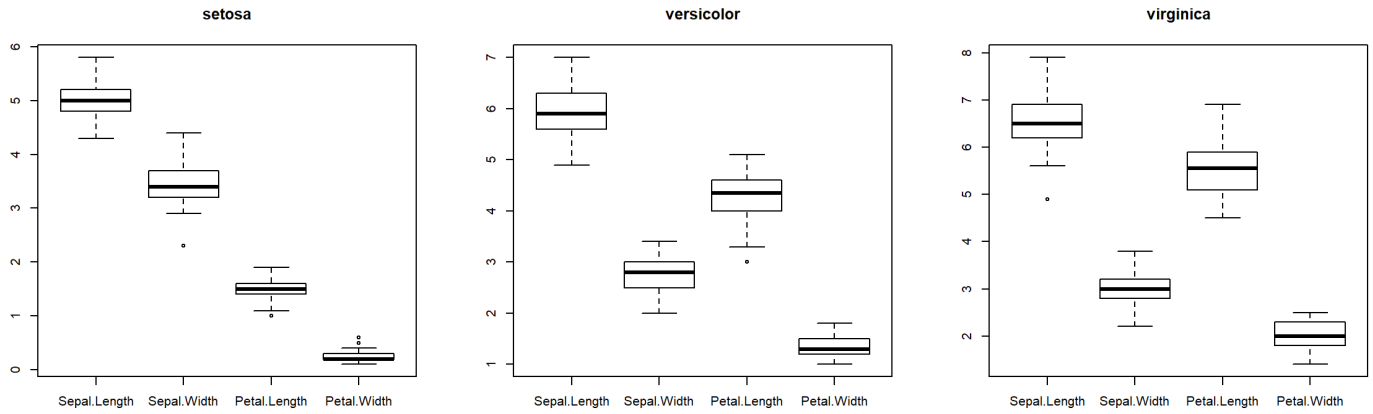
b. Find $cov(\boldsymbol{AX})$.

$$cov(\boldsymbol{AX}) = \boldsymbol{A}cov(\boldsymbol{X})\boldsymbol{A}^T = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -1 \end{bmatrix}^T = \begin{bmatrix} 7 & 8 & 0 & 6 \\ 8 & 13 & -3 & 6 \\ 0 & -3 & 33 & 36 \\ 6 & 6 & 36 & 48 \end{bmatrix}$$

3. (20 points) Consider the `iris` data available in `R`.

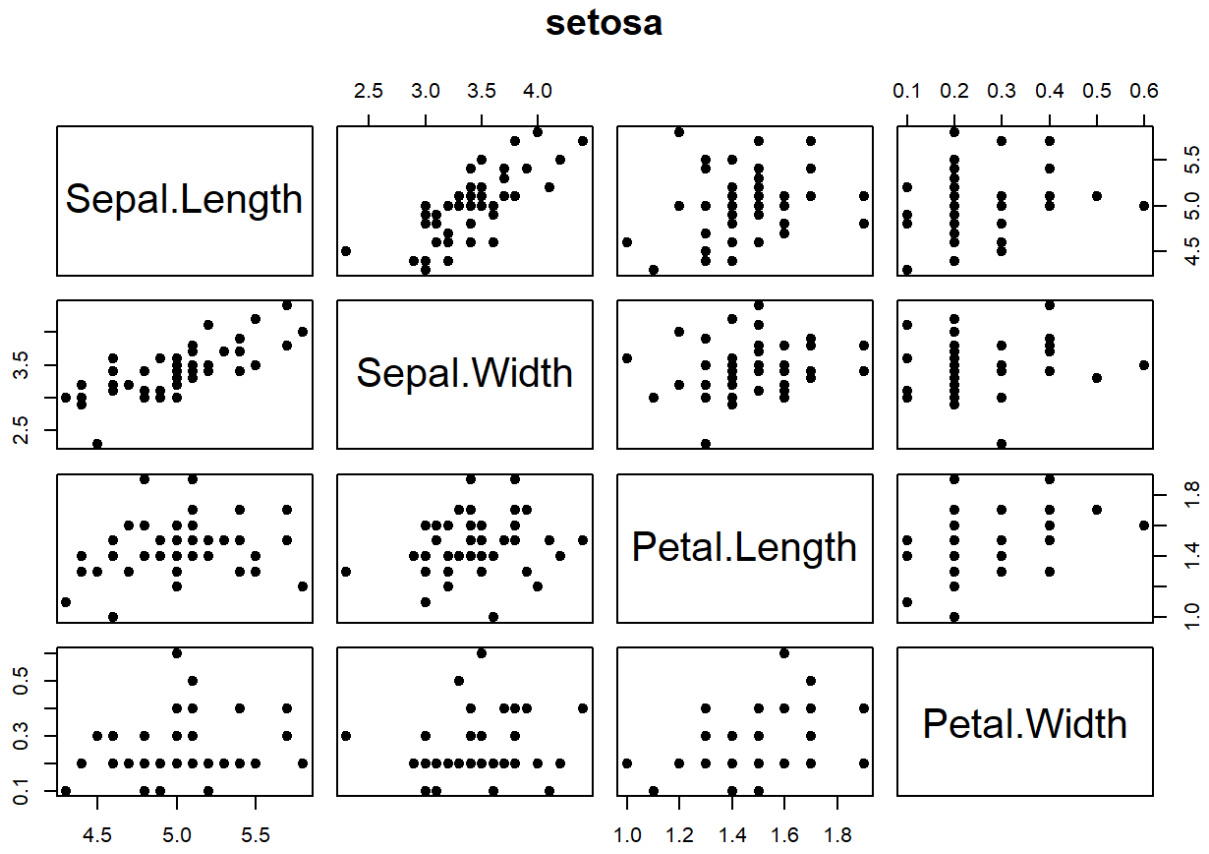
a. Construct side-by-side boxplots of the four quantitative variables (SL, SW, PL, PW) for each species. Do not forget to properly label the axes and give a proper title to the plots when needed. [Hint: your will have 3 plots, one for each species. Each plot will contain four boxplots. See the function `boxplot()` in `R`]

```
species=unique(iris[,5])
par(mfrow=c(1,3))
boxplot(iris[which(iris[,5]==species[1]),1:4],main="setosa")
boxplot(iris[which(iris[,5]==species[2]),1:4],main="versicolor")
boxplot(iris[which(iris[,5]==species[3]),1:4],main="virginica")
```

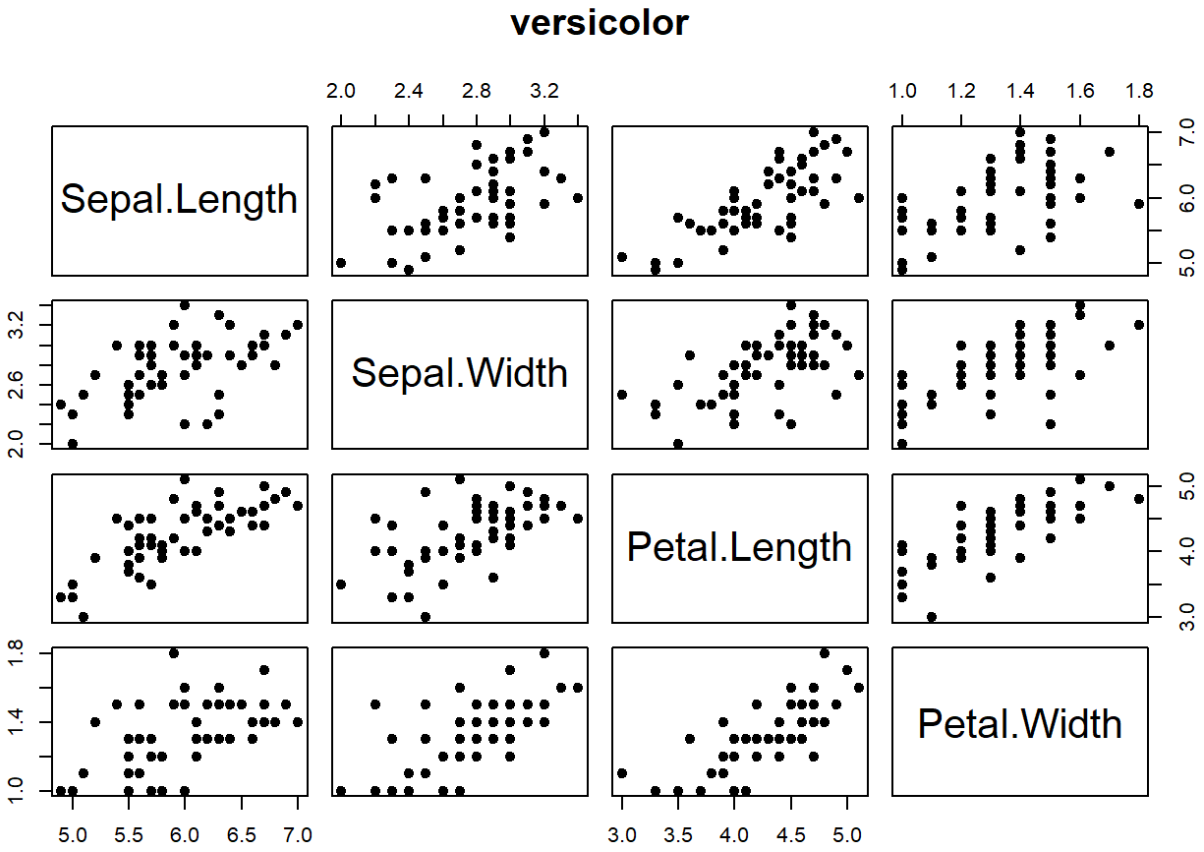


b. Construct a `pairs-plot` of the four variables *for each of the three species*.

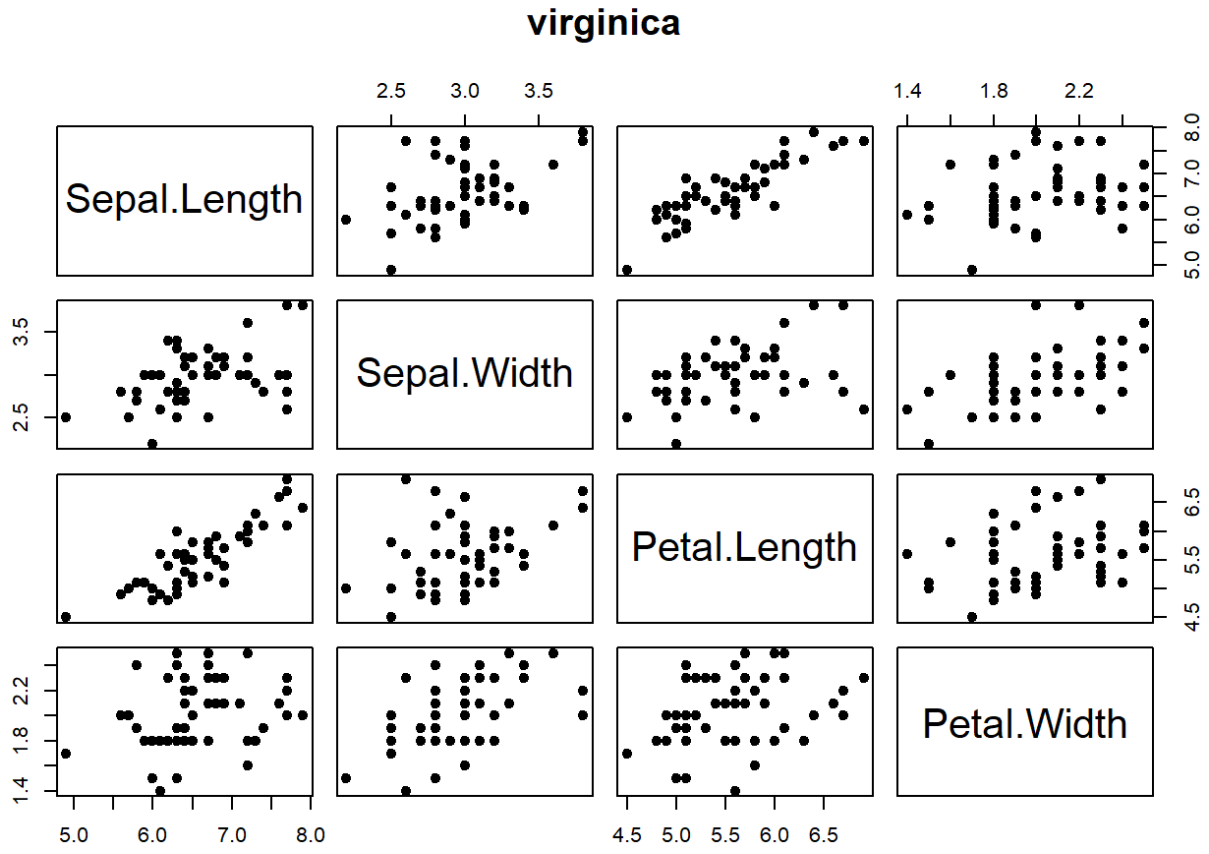
```
species=unique(iris[,5])
par(mfrow=c(1,3))
pairs(iris[which(iris[,5]==species[1]),1:4],main="setosa", pch=19)
```




```
pairs(iris[which(iris[,5]==species[2]),1:4],main="versicolor", pch=19)
```




```
pairs(iris[which(iris[,5]==species[3]),1:4],main="virginica", pch=19)
```



c. Define the vector $\boldsymbol{x} = [Sepal.Length, Sepal.Width, Petal.Length, Petal.Width]^T$. The dataset have 50 observations of this vector (one for each flower), and with 3 species. Thus, we have a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of size $n = 50$ for each of the three species. Compute the sample mean $\bar{\boldsymbol{x}}$ (a 4×1 vector), the sample covariance matrix $\boldsymbol{S}$ (a 4×4 matrix) and the sample correlation matrix $\boldsymbol{R}$ (a 4×4 matrix) *for each species*.

For setosa:

```
####The sample mean is
colMeans(iris[which(iris[,5]==species[1]),1:4])
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
```

```
####The sample covariance is
var(iris[which(iris[,5]==species[1]),1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.12424898 0.099216327  0.016355102 0.010330612
## Sepal.Width    0.09921633 0.143689796  0.011697959 0.009297959
## Petal.Length   0.01635510 0.011697959  0.030159184 0.006069388
## Petal.Width    0.01033061 0.009297959  0.006069388 0.011106122
```

```
####The sample correlation is
cor(iris[which(iris[,5]==species[1]),1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

For versicolor:

```
####The sample mean is
colMeans(iris[which(iris[,5]==species[2]),1:4])
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.936        2.770        4.260        1.326
```

```
####The sample covariance is
var(iris[which(iris[,5]==species[2]),1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.26643265  0.08518367   0.18289796  0.05577959
## Sepal.Width    0.08518367  0.09846939   0.08265306  0.04120408
## Petal.Length   0.18289796  0.08265306   0.22081633  0.07310204
## Petal.Width    0.05577959  0.04120408   0.07310204  0.03910612
```

```
####The sample correlation is
cor(iris[which(iris[,5]==species[2]),1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.5259107    0.7540490   0.5464611
## Sepal.Width     0.5259107   1.0000000    0.5605221   0.6639987
## Petal.Length    0.7540490   0.5605221    1.0000000   0.7866681
## Petal.Width     0.5464611   0.6639987    0.7866681   1.0000000
```

For virginica:

```
####The sample mean is
colMeans(iris[which(iris[,5]==species[3]),1:4])
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        6.588        2.974        5.552        2.026
```

```
####The sample covariance is
var(iris[which(iris[,5]==species[3]),1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.40434286  0.09376327   0.30328980  0.04909388
## Sepal.Width    0.09376327  0.10400408   0.07137959  0.04762857
## Petal.Length   0.30328980  0.07137959   0.30458776  0.04882449
## Petal.Width    0.04909388  0.04762857   0.04882449  0.07543265
```

```
####The sample correlation is
cor(iris[which(iris[,5]==species[3]),1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.4572278    0.8642247   0.2811077
## Sepal.Width     0.4572278   1.0000000    0.4010446   0.5377280
## Petal.Length    0.8642247   0.4010446    1.0000000   0.3221082
## Petal.Width     0.2811077   0.5377280    0.3221082   1.0000000
```

d. Looking at the pairs plot in (b) and the correlation matrices in (c), do you see any patterns or differences among the species? Explain.

There are some visible patterns in the correlation matrices. For example, in the setosa species, `sepal.length` and `sepal.width` have high correlation. But for the other two species, `sepal.length` and `petal.length` have high correlation. Similarly, `petal.length` and `petal.width` have high correlation in versicolor but small or moderate correlation for the remaining two species.

4. (20 points) Consider the `skulls` dataset in the `HSAUR3` package in `R`. You will first need to install the package in `R` to access the dataset. Use `?skulls` command to get more details on the data. A snapshot of the data is shown below.

```
library(HSAUR3)
```

```
## Loading required package: tools
```

```
head(skulls)
```

```
##     epoch  mb  bh  bl nh
## 1 c4000BC 131 138  89 49
## 2 c4000BC 125 131  92 48
## 3 c4000BC 131 132  99 50
## 4 c4000BC 119 132  96 44
## 5 c4000BC 136 143 100 54
## 6 c4000BC 138 137  89 56
```

a. Suppose we want to estimate the *population mean* of all the 4 variable for skulls with epoch `c4000BC`. Write down the population, parameter, the sample and the statistic you will use to answer the question above.

In this case,

- Population: all the skulls belonging to epoch `c4000BC`.

- Parameter: the vector of true means of the form variables, $\boldsymbol{\mu}_{4\times1}$.
- Sample: the 30 skulls we have observed the data on.
- Statistic: We use the sample mean vector, $\bar{\boldsymbol{X}}_{4\times1}$, to estimate $\boldsymbol{\mu}_{4\times1}$.

b. From the `skulls` data set, provide an estimate (numeric value) of the parameter mentioned above. Explain how you obtained it.

```
data=skulls[1:30,]
###estimated population mean (i.e., the sample mean)
colMeans(data[,2:5])
```

```
##        mb        bh        bl        nh
## 131.36667 133.60000  99.16667  50.53333
```

c. Now suppose we want to estimate the population variance-covariance matrix. When estimator will you use? Provide a numeric estimate.

We will use the sample covariance matrix as an **estimator**: $S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$

A numeric **estimate** is given by:

```
cov(data[,2:5])
```

```
##           mb         bh         bl         nh
## mb 26.309195  4.1517241  0.4540230  7.2459770
## bh  4.151724 19.9724138 -0.7931034  0.3931034
## bl  0.454023 -0.7931034 34.6264368 -1.9195402
## nh  7.245977  0.3931034 -1.9195402  7.6367816
```

d. Compute the variance covariance matrix of the estimator of the population mean in part (b).

We used the estimator $\bar{\boldsymbol{X}}_{4\times1}$ in part (b). Thus the covariance of the estimator is $cov(\bar{\boldsymbol{X}}_{4\times1}) = \boldsymbol{\Sigma}/n$, where $\boldsymbol{\Sigma}$ is the true population covariance matrix.

Since do not know the true value of $\boldsymbol{\Sigma}$, we need to estimate it by the sample covariance $\boldsymbol{S}$ as in part (c). Thus, an estimator of $cov(\bar{\boldsymbol{X}}_{4\times1})$ is $\boldsymbol{S}/n$. A numerical estimate id given below.

```
cov(data[,2:5])/30
```

```
##           mb          bh          bl          nh
## mb 0.8769732  0.13839080  0.01513410  0.24153257
## bh 0.1383908  0.66574713 -0.02643678  0.01310345
## bl 0.0151341 -0.02643678  1.15421456 -0.06398467
## nh 0.2415326  0.01310345 -0.06398467  0.25455939
```

e. Provide an estimate for the parameter vector $(\mu_{mb} - \mu_{nh}, \mu_{bh} - \mu_{nh})^T$ and compute the covariance matrix of the estimator.

Let $\theta = (\mu_{mb} - \mu_{nh}, \mu_{bh} - \mu_{nh})^T$ Let $\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \end{bmatrix}$. Then we can write $\theta = \mathbf{A}\mu$. Since we can estimate μ by $\bar{\boldsymbol{X}}$, we can estimate $\mathbf{A}\mu$ by $\mathbf{A}\bar{\boldsymbol{X}}$. A numerical estimate is computed below.

```
A=matrix(c(1,0,0,1,0,0,-1,-1),nrow=2)
# estimate
A %*% colMeans(data[,2:5])
```

```
##          [,1]
## [1,] 80.83333
## [2,] 83.06667
```

Also, $cov(\mathbf{A}\bar{\boldsymbol{X}}) = \mathbf{A}cov(\bar{\boldsymbol{X}})\mathbf{A}^T = \mathbf{A}(\boldsymbol{\Sigma}/n)\mathbf{A}^T$. We estimate this quantity by replacing $\boldsymbol{\Sigma}$ by $\mathbf{S}$, that is, $\mathbf{A}(\mathbf{S}/n)\mathbf{A}^T$. A numerical estimate is computed below.

```
# covariance of mean estimator
A %*% (cov(data[,2:5])/30) %*%t(A)
```

```
##           [,1]      [,2]
## [1,] 0.6484674 0.1383142
## [2,] 0.1383142 0.8940996
```

5. (10 points) Answer the following questions.

a. What is the basic difference between multivariate and longitudinal data?

Multivariate data is a data-set where multiple variables are measured for each subject (no specific ordering among the variables). Longitudinal data is a data-set where one or more variables are measured on each subject *at different time points*.

b. Why is investigating covariance (or correlation) between two variables important?

Examining covariance is important to understand the relashioship among the measured variables. This in turn will help us provide better inference about parameters.

c. Suppose that in your analysis, you found $cor(X, Y) = 0$. Can you say that "X and Y does not have *any* relationship at all?" Explain.

No. The correlation coefficient only measures linear relationship. There might be other (e.g., quadratic) relationship even in the correlation coefficient is zero.

d. In problem 4 above, you considered the `skulls` dataset. Is this a longitudinal dataset (since the skulls are coming from different era/time as given in the epoch)? Explain.

No. Each skull belongs to a different subject; a subject can only belong to one epoch. Thus, in this dataset, one subject has only one measurement.