NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

2019/20 Semester 1 ST2137 Computer Aided Data Analysis Tutorial 2

Two datasets (“tut2htwtfixed.txt” and “tut2test.csv”) have been uploaded to the course website in the IVLE. They are used for Questions 1 to 7.
There are five variables in the “tut2htwtfixed.txt” dataset. They are

- id: Identity of the subject (Columns 1-3)
- gender: Gender of the subject (Column 4)
- height: Height of the subject in cm (Columns 5-7)
- weight: Weight of the subject in kg (Columns 8-9)
- siblings: Number of siblings of the subject (Column 10)

The column numbers in parentheses are the columns occupied by these variables.

“tut2test.csv” is an Excel file with comma-separated values. There are two variables in it. They are

- id: Identity of the subject
- test: Test score of the subject

1. Create an R data frame “htwt2” by importing the “tut2htwtfixed.txt” file into the R.

2. Based on “htwt2”, create an R data frame “htwt2m” which contains the data for all the male subjects. How many males are there in the data frame “htwt2”?

3. Import “tut2test.csv” into the R. Merge the two datasets “htwt2” and “tut2test”. Let us call this new R data frame “htwttest2”. Identify individuals whose height is greater than 182 cm. What are the test scores of subjects whose height is greater than 182 cm?

4. Suppose that there was an error in the weight of the Subject 210 in the text file. Obtain a new R data frame “htwttest2remo” by removing the record related to the Subject 210 from the data frame “htwttest2”.

5. After checking with the Subject 210, we found out that his actual weight should be 68 kg instead of 88 kg. Modify the R data frame “htwttest2” by rectifying the mistake.

6. Who is the second tallest female in this group? What are her height, weight, and test score?

7. Create a new variable called “grade” using the following rules: (1) grade = “A” if test $\geq$ 80, (2) grade = “B” if $70 \leq$ test < 80, (3) grade = “C” if $60 \leq$ test < 70, (4) grade = “D” if $50 \leq$ test < 60 and (5) grade = “F” if test < 50. How many subjects who have “F” grade are there?

8. Suppose a matrix $X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \\ 1 & 11 \end{pmatrix}$ and $\underline{y} = \begin{pmatrix} 4 \\ 6 \\ 13 \\ 15 \\ 19 \end{pmatrix}$.

Define $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$. Using the matrix operations in R to find $\underline{\hat{\beta}}$.
[Some useful matrix operation commands in R:
matrix multiplication of A and B: A%*%B;
transpose of A: t(A);
inverse of A: solve(A)]

9. A sequence is generated using the following recursive relation

$$x_n = x_{n-1} - 2x_{n-2} \quad \text{for} \quad n \geq 3$$

with $x_1 = 0$ and $x_2 = 2$.
(i) Use the loop function in R to find the 18th term of the series.
(ii) Find the sum of the first 15 terms in this sequence.

10. Write a function that will calculate the mean, the second, the third and the fourth central moments a given data vector on variable *X*.
The r^{th} central moment for $r \geq 2$ is given by

$$M_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r, \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Hence obtain the mean, the second, the third and the fourth central moments for the height in the data frame “htwt2”.

Partial code is given as follows.

```
# The function "cenmom" finds the mean, the 2nd, 3rd & 4th
central moments
cenmom <- function(x){
n <- length(x)
s <- numeric(4)
```

code to compute the four center moments

```
return(s) # To return the 4 values in the object "s".
}
```

Answers/Hints to selected questions

2. 48 males
3. 3 individuals. Subjects 261, 271 and 285 with heights 183 cm, 188 cm and 184 cm, and test scores 55, 76 and 54 respectively.
6. Subject 273 whose height, weight and test score are 174cm, 64kg and 57 respectively.
7. 6 subjects have F grade
8. $\hat{\beta} = \begin{pmatrix} 3.565789 \\ 1.506579 \end{pmatrix}$
9. 542, −92
10. The first four central moments are 165.86667, 75.98222, 205.46193 and 13883.27461.