Problem Set 2

Applied Statistics and Econometrics II
Spring 2018, NYU
Ercan Karadas

(Due: February 15, in class)

[1] Store the values $-20, -15, -5, 8, 12, 9, 2, 23, 19$ in the R variable `x`.

a) Use the R command `sum` to verify that the sum of the values is 33.

b) Compute an average by using the R command `mean`?

c) Compute the average with using the R command `sum`?

d) Use R to sum the positive values in `x`.

e) Use the `which` command to get the average of the values ignoring the largest value.

f) Speculate about the values corresponding to the command `x[abs(x)>=8 & x<8]`. Verify your speculation running this R command.

[2] Let `x = c(1, 8, 2, 6, 3, 8, 5, 5, 5, 5)`

a) Describe two different R commands for summing the values in `x` ignoring the value 2 stored in `x[3]` and the value 3 stored in `x[5]`.

b) Use two different R commands to sum all of the values not equal to 5.

c) Use a single R command to change all values equal to 8 to 7.

[3] a) Create a 10×5 matrix M whose elements are random draws from a normal distribution with mean 5 and variance 2.

b) Create another matrix N of the same size which contains all zeros, except 5 NA and locations of these NAs are randomly determined (i.e. in the sense that each time you run your code their locations are expected to change).

c) Using M and N generate a random matrix from $N(5, 2)$ which contains 5 NA values that are arbitrarily located in the matrix.

d) Describe how the R function `is.na` can be used to eliminate the rows with missing values.

[4] R has a built-in data set called `chickwts`, which is stored in a data frame with two columns. The first column contains the weight of chicks, and the second column indicates the type of feed they received, one of which is labeled `horsebean`. Use R to compute the average weight among chicks that were fed horsebean.

[5] a) Create a vector named `my_vec` containing the integers 1 through 100 and then divide each element of `my_vec` by 3 and store the result as `my_vec2`. (Your answer should contain two lines of R commands).

b) Compute the average of the vector `my_vec` you created before without using built-in R function `mean`.

c) Create a vector named `my_vec3` containing the elements of `my_vec` that are between 20 and 35. (Your answer should contain a single line of R commands)

[6] Briefly explain what would be the output of following R command: `mean(rnorm(1000))`

[7] Make a script file which constructs two random normal vectors of length 10. Call these vectors `x1` and `x2`. Make a data frame called `T` with two columns (called `a` and `b`) containing respectively `x1` and `x1` + `x2`.

[8] When the function `head` called on the data frame `Orange` it produces the following output:

```
> head(Orange, n=2)
  Tree  age circumference
1    1  118            30
2    1  484            58
```

Write the command that adds up Y (defined as `circumference`) and X (defined as square root of `age`). (Your answer should contain at most three lines of R commands)

[9] `Orange` is a data frame with two numeric variables `circumference` and `age`. Create a dummy variable that is 0 when `age` is less than or equal to 900 and 1, otherwise.

[10] Put all even integers from 30 to 89 in a vector named P and then in a matrix with 6 rows and 5 columns named Q.

[11] Declare a function in R, named `my_function` which takes a vector, say `x`, as input and returns the sum of the elements in the vector and the mean of the values in the vector. Also, make sure that this function returns the message ”Data is not numeric!” when you feed in a non-numeric vector.

[12] Using a `for` loop in R, write a script that produces the sum of the first $n = 40$ integers.

[13] R has a built-in data set called `ChickWeight`. Verify that the R command

```
mean(ChickWeight[,1])
```

returns 121.8 but that the command

```
mean(ChickWeight[,3])
```

returns NA and a warning message even though the values in column 3 appear to be numeric. The reason for the warning message is that column 3 is stored as a factor variable. Arithmetic operations can only by performed on numeric or logical variables. Verify that

```
mean(as.numeric(ChickWeight[ ,3]))
```

returns 26.26.

[14] The final exam scores for 15 students are

$$73, 74, 92, 98, 100, 72, 74, 85, 76, 94, 89, 73, 76, 99$$

Compute the mean, 20% trimmed mean, and median using R.

[15] Let $\{x_1, x_2, \ldots, x_n\}$ be a sample and suppose that we declare x_i an *outlier* if it is more than two standard deviation of the mean, i.e.

$$\frac{|x_i - \bar{X}|}{s} > 2\,.$$

For the values

$$20, 121, 132, 123, 145, 151, 119, 133, 134, 130, 200$$

write an R command to determine whether any outliers exist.

[16] Importing data from a plain text file can be illustrated with an example of a data set available on the website ”Data and Story Library” (DASL). The Massachusetts lunatics data is available at `http://lib.stat.cmu.edu/DASL/Datafiles/lunaticsdat.html`.

a) Saved the data as `lunatics.txt` in a folder named ”R Practice” on your desktop.

b) Use the `read.table` function to read the file into a data frame.

c) The `str` (structure) function provides a quick check that 14 observations of six variables.

d) Convert `lunatics.txt` into `lunatics.csv` and save in the same folder.

e) Import `lunatics.csv` using the package `readr` and compute the total population of 14 counties.

[17] Many of the interesting data sets that one may wish to analyze are available on a web page. R provides an easy way to access data from a file on the internet using the URL of the web page. The function `read.table` can be used to input data directly from the internet.

a) The data file `PiDigits.dat`, located at `http://www.itl.nist.gov/div898/strd/univ/data/PiDigits.dat`, contains the first 5000 digits of the mathematical constant π. Use `read.table` to save the data in the folder ”R Practice”.

b) Redo the same thing but now skip the first 50 digits and save as `PiDigitsSkipped.csv`.

c) Using `head` command print some of the data on the console.

d) Are the digits of π uniformly distributed? Using `table` command compute the relative frequencies of each digit.

e) Use `barplot` to visualize the tabulated data you obtained above.

[18] Lets say you have a data frame named `mydata`, with variables `x1` and `x2`, and you want to create a new variable `sumx` that adds these two variables and a new variable called `meanx` that averages the two variables. Furthermore, you want to add these new variables to the original data frame. Here is how you do it:

```
# Prepare data
mydata <- data.frame(x1 = c(2, 2, 6, 4),
                     x2 = c(3, 4, 2, 8))

# Method 1: use $ to append these two variables to the data frame
   directly

# Method 2: use attach() function

# Method 3: use transform() function
```

[19] Suppose you have the following data set

```
manager <- c(1, 2, 3, 4, 5)
date <- c(”10/24/08”, ”10/28/08”, ”10/1/08”, ”10/12/08”, ”5/1/09”)
country <- c(”US”, ”US”, ”UK”, ”UK”, ”UK”)
gender <- c(”M”, ”F”, ”F”, ”M”, ”F”)

age <- c(32, 45, 25, 39, 99)
      q1 <- c(5, 3, 3, 3, 2)
      q2 <- c(4, 5, 5, 3, 2)
      q3 <- c(5, 2, 5, 4, 1)
      q4 <- c(5, 5, 5, NA, 2)
      q5 <- c(5, 5, 2, NA, 1)

leadership <- data.frame(manager, date, country, gender, age,
                         q1, q2, q3, q4, q5, stringsAsFactors=
   FALSE)
```

a) Recode the value 99 for `age` to indicate that the value is missing

b) Lets say you want to recode the ages of the managers in the leadership dataset from the continuous variable age to the categorical variable `agecat` (Young, Middle Aged, Elder).

c) Note that `agecat` is a character variable. Turn it into an ordered factor.

d) Lets say you want to change the variable `manager` to `managerID` and `age` to `ages`.

e) Suppose you have data `y <- c(1, 2, 3, NA)`.
Then `is.na(y)` returns `c(FALSE, FALSE, FALSE, TRUE)`. What would be the output of the following command: `is.na(leadership[4:5, 6:10])`

f) Save the `date` variable as a date variable

[20] Explain why `sum(c(1, -2, NA, 3))` would return `NA`, but not

`sum(c(1, -2, NA, 3), na.rm=TRUE)`?

[21] You can remove any observation with missing data by using the `na.omit()` function. It deletes any rows with missing data. Remove the row with NA values from `leadership` data frame.

[22] (Counting Missing Values) Create a vector of lenght 100 with every third element missing (namely ‘NA‘) by only using vector operations. Then count the number of missing values by utilizing only the following two function ‘is.na‘ and ‘sum‘.

[23] (Leadership Example continued) Suppose you have the following data

```
leadershipOriginal <- leadership
```

a) Create a new dataset ‘newdata‘ containing rows sorted from youngest manager to oldest manager.

b) Sort the rows into female followed by male, and youngest to oldest within each gender.

c) Sort the rows by gender, and then from oldest to youngest manager within each gender.