PROJECT 3, LECTURE 3: MORE DATA EXPLORATION IN R, AND SOME PROPERTIES OF COVARIANCES AND CORRELATIONS
GRA4110

Part I: Further exploratory analysis of the brain-data

This is a continuation of Project 2. You are expected to have completed this project before starting the following exercises.

Please note: The first part of this project is *technically easy*, in the sense that you are told what to do in R. But please do not underestimate this part: spend enough time to think about what you see, what it means, and make notes about your thoughts and ideas as you go along.

We will here revisit the dataset from Project 2. The dataset comes from Willerman et al. (1991), where the authors investigated the relationship between the size and weight of the brain and a persons mental capacity.

Recall the setup:

The study was conducted at a large southwestern university, with a sample of 40 right–handed Anglo introductory psychology students who had indicated no history of alcoholism, unconsciousness, brain damage, epilepsy, or heart disease. These subjects were drawn from a larger pool of introductory psychology students with total Scholastic Aptitude Test Scores higher than 1350 or lower than 940 who had agreed to satisfy a course requirement by allowing the administration of four subtests (Vocabulary, Similarities, Block Design, and Picture Completion) of the Wechsler (1981) Adult Intelligence Scale-Revised. They used Magnetic Resonance Imaging (MRI) to determine the brain size of the subjects. The researchers also took into account the gender and two body measurements (hight and weight) when they analysed the connection between brain size and intelligence. With prior approval of the University's research review board, students selected for MRI were required to obtain prorated full-scale IQs of greater than 130 or less than 103, and were equally divided by sex and IQ classification.

The datafile `BrainSize.tsv` contains 40 samples (rows), and 7 different types measurements/variables (columns):

1. `Gender`: Male or Female
2. `FSIQ`: Full Scale IQ scores based on the four Wechsler (1981) subtests
3. `VIQ`: Verbal IQ scores based on the four Wechsler (1981) subtests
4. `PIQ`: Performance IQ scores based on the four Wechsler (1981) subtests
5. `Weight`: body weight in pounds
6. `Height`: height in inches
7. `MRI_Count`: total pixel Count from the 18 MRI scans

In Project 2, we explored some of the variables of the brain size dataset using simple plots and summary statistics. In the exercises below, we will continue this exploration, and investigate the relationship between pairs of variables.

Exercise 1:

a) When you change the current directory, R outputs the command

```
setwd(" ... ")
```

where "..." is the *folder* you change the directory to. You can copy-paste this output to an R-script, and when you execute this line of code, R will change the current directory to the specified place.

Modify and execute the attached R-code, so that it changes the directory to where the dataset is located, load it and then remove the observations with the missing values that was discovered in Project 2.

b) Let us start by re-capping parts of the previous exploratory analysis.

The study contains three different measurements for intelligence, FSIQ, VIQ and PIQ; each measuring a different type of intelligence. We will be interested in modelling intelligence, and

Date: September 8, 2022.

it is natural to wonder how correlated these measurements really are, and if they measure some fundamental different aspects of intelligence; perhaps we can limit the analysis to only one of these? Etc. We start by comparing the distributions.

The following line(s) of code can be found below the comment `# Exercise b)` in the R script `BrainSize2.R` (and is included here for completeness)

```
with(brain_data, plot(density(FSIQ, bw = 6),
col  = "red",
main = "",
xlab = "FSIQ (red), VIQ (blue) and PIQ (green)"))
```

In order to add the other estimated densities to the plot, (i) include

```
with(brain_data, lines(density(VIQ, bw = 6), col = "blue"))
```

below the code used to plot the density of full scale IQ score (FSIQ) in the script, and (ii) modify this line of code (change VIQ with PIQ and change the colour) to add the density for the performance IQ score (PIQ) in green to the same plot. Remember that a estimated density is essentially a smoothed histogram, and that we may use this as an alternative to the traditional histogram. In particular, they are convenient when we want to plot different "histograms" in the same plot.

These plots indicate that there are more than one group present in the dataset (since the estimated densities are so-called bimodal, meaning two tops). And from the description of the datasets, we learn that this is actually the case. They have intentionally included students with FSIQ less than 103 and above 130 in the study. Even if all the densities are bimodal, however, it does not necessarily imply that the measurements divide the participants into the same two groups (e.g. a high FSIQ does not necessarily imply a equally high score across different IQ tests).

c) To investigate this further, we look at the correlations between the different IQ scores.

(i) Run the command `cor(VAR1, VAR2)` where `VAR1` and `VAR2` are all pairwise combinations of `brain_data$FSIQ`, `brain_data$VIQ`, `brain_data$PIQ`.

(ii) include and run the following command to the R script:

```
cor(brain_data[c("FSIQ", "VIQ", "PIQ")])
```

Note the accessing technique of the variables in the above code. The command `brain_data[c("FSIQ", "VIQ", "PIQ")]` returns a new data matrix with just these sub-columns, i.e., the `FSIQ`, `VIQ` and `PIQ`. `cor` then returns all pairwise correlations between these columns in a matrix, the so-called correlation matrix.

Note: An alternative way to generate a matrix whose columns are the variables `FSIQ`, `VIQ` and `PIQ` is the command

```
cbind(brain_data$FSIQ, brain_data$VIQ, brain_data$PIQ)
```

Verify this. Note that `cbind` (a command we will study in more detail later) concatenates column-vectors into a new matrix. Explain why

```
c(brain_data$FSIQ, brain_data$VIQ, brain_data$PIQ)
```

does not produce the same result as `brain_data[c("FSIQ", "VIQ", "PIQ")]`, and will not be useful for computing the correlation matrix of these variables.

(iii) Why is the correlation matrix symmetric? *Hint:* Recall the mathematical formula for the empirical correlation.

(iiii) Add the line `pairs(brain_data[c("FSIQ", "VIQ", "PIQ")])` to the script to construct the corresponding scatter plot matrix; do this and run the code. They all appear to be highly correlated. Do you agree, and do you see anything else? Does using correlation appear to be unproblematic for this dataset? From the description, however, we learn that the different IQ scores are derived from a common set of test scores and measurements; which may indicate a strong connection/correlation. Therefore, we will continue to work with only the full scale IQ score (FSIQ) as our response variable.

d) Now, since we have decided to focus on the full scale IQ score (FSIQ) as our response, we are ready to explore the relationship between intelligence (FSIQ) and the size of the brain (MRI counts). Start by (i) including the following line (to compute the correlation between intelligence and size) to the R script

```
with(brain_data, cor(MRI_Count, FSIQ))
```

In this context, do you see this as a high or low correlation? Then, (ii) add the following code to compute the corresponding scatter plot (with a so-called smoothing line)

```
with(brain_data, plot(MRI_Count, FSIQ, cex = 2.5, pch = 21, bg = "grey"))
with(brain_data, lines(smooth.spline(MRI_Count, FSIQ, df = 5), col = "red"))
```

Feel free to improve the appearance of such plots by changing e.g. `cex = ...` the size of the points, `pch = ...` the type of points used or `col = ...` for colour and `bg = ...` for background colour (for those points that have a background colour).

Write down a description of the relationship between `MRI_count` and `FSIQ`. Does the relationship seem linear, i.e., explainable by a straight line (the simple linear regression model)? Keep in mind that the dataset contains data from two groups, those with FSIQ above 130 and those with a score below 103. If you look at the scatter plot, are you able to differentiate the two groups in the plot? If not, (iii) try to run the following R command in the console

```
with(brain_data, plot(MRI_Count, FSIQ, cex = 2.5, pch = 21, bg = (FSIQ >= 130)))
```

The `bg = (FSIQ >= 130)` colours the individuals whose `FSIQ` equals or exceeds 130. We will shortly spend considerable time on such data selection commands, and will currently be content with just stating the above functionality.

Does the relationship appear linear for each group? Later we will divide the two groups and study them separately.

e) Next, we will explore how the remaining predictors (i.e. height, weight and gender) are related to the response FSIQ. In the R script (`BrainSize2.R`) there are commands needed to compute the correlation between height (in centimeters) and the full scale IQ score (FSIQ), and also the corresponding scatter plot (with a smoothing line); (i) run these lines. Is the correlation strong and what do you see from the plot? It looks like the overall relationship is negative. However, if you imagine that you split the points into the two groups (those with high and low full scale IQ score), the relationship (if any) looks increasing within each group; do you agree?

Then, use the three lines of code in the R script which computes the the correlation and scatter plot for `HeightCM` and `FSIQ`, as a template to compute (ii) the correlation and the scatter plot (with the smoothing line) between Weight (in kilograms) and FSIQ. Add the modified code to the R script (below the code for `HeightCm` and `FSIQ`), run the commands, do you see anything of interest?

Finally, we will explore the relationship between gender and the full scale IQ score (FSIQ). We will use a box plot to inspect the relation, (iii) add the code

```
with(brain_data, boxplot(FSIQ ~ Gender, horizontal = TRUE))
```

to the R script, run the code, do you see anything that is worth reporting? Note that `horizontal = TRUE` merely changes the orientation of the box-plots (try running the above command without `horizontal = TRUE` to see what I mean).

f) As it turns out, it is apparently well known (see `https://en.wikipedia.org/wiki/Brain_size`) that certain external factors are correlated to brain size (here represented by MRI count); the relationship between size and intelligence is not clear. To investigate this, we will now explore (i) how the external measurements (`HeightCm`, `WeightKg` and `Gender`) are related to MRI counts (the size of the brain). The following two lines of code computes the full covariance matrix and the corresponding scatter plot matrix (add these to the R script)

```
cor(brain_data[c("MRI_Count", "HeightCm", "WeightKg")])
pairs(brain_data[c("MRI_Count", "HeightCm", "WeightKg")])
```

Note that these are all quite strongly correlated, an issue which we will explore with more care at a later stage of our analysis (not covered in this project).

Then, (ii) we will look at the relationship between brain size (i.e. the MRI count) and gender; add and run

```
with(brain_data, boxplot(MRI_Count ~ Gender, horizontal = TRUE))
```

to the R script. What do you see? And how to interpret this in relation to the findings above (think about differences in height and weight for men and women)? Also, try to Google "brain size and gender".

g) (Optional) Let's here note that the brain is, to say the least, not well-understood. We should therefore be careful and not over-interpret patterns we observe that relates to the brain. A dominant view of the brain is that it functions in much the same way as our physical computers, but there are many other views of the brain. For example, Rupert Sheldrake argues that the brain is more like a receiver and sender than where consciousness resides, using sources such as the paper Lewin (1980) (uploaded on Itslearning). In Lewin (1980), a striking case of a mathematics student with a very small brain is reported. Another example of alternative views is the philosophical direction of panpsychism, see `https://iep.utm.edu/panpsych/`.

Part II: Selecting subsets of a dataset

Exercise 2: We will here continue our work with the brain-size data, with a focus on how to access subsets of the data. We will here be "less statistically interested" than in Exercise 1, and will instead focus more on the purely technical aspects of accessing data.

a) Start a new R-script which goes to the directory with the brain-size dataset and loads it. Copy-paste in the code which removes the observations which has missing values and run it. Organize this R-script as you have been taught in earlier exercises, and as you will be expected to do in all your future work, including the home exam.

Note that we will here completely ignore the removed observations in the rest of this exercise, and use the *new* enumeration of the observations after having removed the two above mentioned cases. That is: we remove the original "observation 2", making the original third observation the new *second* observation.

Most functions in R will not be affected by this, but a main exception is the `View`-command that we will use later on to verify some of our steps. The `View`-command shows the `rownames`

of the dataset, which is the enumeration of the observations before the removal of the two observations with missing values, and therefore presently skips 2 and 21 (verify this). In order to avoid any confusion at this point, we will create new row-names (which is just a label for each row, and will in most cases be irrelevant). Therefore, add the following lines of code to your script:

```
n <- length(brain_data$Height)
rownames(brain_data) <- (1:n)
```

The first line extracts the number of elments from one of the variables in the dataset, which is the sample-size, i.e., the number of rows. The second line assignes the integers from 1 to n to the rownames. After running this code, verify that the row-names displayed on the left of the spreadsheet shown by `View(brain_data)` no longer skips over observation 2 and 21.

b) Use the material from Project 2 to access the height of the first 20 individuals in the study.

Hint: To access the height of the first two individuals in the study, you can write `brain_data$HeightCm[c(1,2)]`, or alternatively `brain_data$HeightCm[(1:2)]`. A third option is

```
firstPersons <- (1:2)
brain_data$HeightCm[firstPersons]
```

Hint: Avoid writing in many numbers manually in your script!

c) Consider the following R-code.

```
brain_data$Gender == "Male"
```

The result is a vector of `TRUE` and `FALSE` values. The elements which are `TRUE` are where the gender variable is registered as male. The elements which are `FALSE` are where the gender variable is registered as female. For example, the first value is `FALSE`, meaning that the first observation is a woman. Use `View(brain_data)` to verify this for the first observations.

d) A vector with values of the type `TRUE` and `FALSE` is called a Boolean vector. An important command in R is `which`. When given a Boolean vector `X`, it returns a new vector with the indices of the elements `X` which are `TRUE`.

For example,

```
which(brain_data$Gender == "Male")
```

returns

```
[1]  2  3  8  9 11 12 17 19 20 22 24 26 30 31 32 35 37 38
```

Check (using the `View`-command) that this indeed is a complete list of the indices whose observations are made on men.

Save the result of `which(brain_data$Gender == "Male")` in a vector. Use this vector to access the height of the men in the dataset.

e) Make a histogram of the height of the men and compute their mean height. Similarly, modify the above points to make a histogram of the height of the woman, and also compute their mean height. Then make a single plot where the histograms of the weights of each group are placed next to each other. Also make a histogram with the weights of all people in the dataset. Comment.

Hint: Recall the `mfrow`-command from Project 2.

f) If we multiply a Boolean vector by 1, R converts the vector to a sequence of zeros and ones. `TRUE` is converted to 1, and `FALSE` is converted to 0 (try this!)

This conversion is also done by R automatically if we attempt to use arithmetic operations on a Boolean vector.

What does the following code find? Explain.

```
n <- length(brain_data$Gender)
nMale <- sum(brain_data$Gender == "Male")
nFemale <- sum(brain_data$Gender == "Female")
n
nMale + nFemale
```

Part III: Some elementary properties of covariances and variances

Part of the motivation for studying the following mathematical tasks is to understand the formulas underlying linear regression, formulas we will study later in the course. Recall that if you do not have time to complete the following assignment during class, it is to be considered homework.

Exercise 3: Let $X_1, X_2, \ldots, X_n$ be numbers.

a) Let us define

$$\tilde{X}_i = X_i - \bar{X}_n \qquad \text{for } i = 1, 2, \ldots, n,$$

where we recall that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Show that the average of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ is zero.

b) Show that the empirical variance of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ is equal to the empirical variance of $X_1, X_2, \ldots, X_n$.

Hint: Recall that the empirical variance of some observations, say, $Y_1, Y_2, \ldots, Y_n$ is given by $\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, where $\bar{Y}_n$ is the average of $Y_1, Y_2, \ldots, Y_n$.

c) Let us now re-define $\tilde{X}_i$, and rather set

$$\tilde{X}_i = \frac{X_i - \bar{X}_n}{s_X} \qquad \text{for } i = 1, 2, \ldots, n,$$

where s_X is the empirical standard deviation of $X_1, X_2, \ldots, X_n$, i.e.,

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

Show that the average of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ is equal to zero, and that the standard deviation of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ is equal to one.

d) Let us again re-define $\tilde{X}_i$, and rather set

$$\tilde{X}_i = aX_i + b \qquad \text{for } i = 1, 2, \ldots, n,$$

for numbers a, b. Show that

$$\frac{1}{n}\sum_{i=1}^n \tilde{X}_i = a\bar{X}_n + b$$

and

$$s^2_{\tilde{X}} = a^2 s^2_X,$$

where $s^2_{\tilde{X}}$ is the empirical variance of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$, and s^2_X is the empirical variance of $X_1, X_2, \ldots, X_n$. Explain why the results of (a), (b) and (c) are special cases of this result.

Exercise 4: Recall that the empirical covariance between $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ is given by

$$s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

Also recall that the empirical correlation between $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ is given by

$$r_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{X_i - \bar{X}_n}{s_X}\frac{Y_i - \bar{Y}_n}{s_Y}$$

where s_X, s_Y are the standard deviations of $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ respectively.

a) Define the so-called standardized variables

$$\tilde{X}_i = \frac{X_i - \bar{X}_n}{s_X}, \quad \tilde{Y}_i = \frac{Y_i - \bar{Y}_n}{s_Y}, \quad i = 1, 2, \ldots, n.$$

Show that

$$s_{\tilde{X},\tilde{Y}} = r_{X,Y},$$

and

$$r_{\tilde{X},\tilde{Y}} = r_{X,Y},$$

where $s_{\tilde{X},\tilde{Y}}$ and $r_{\tilde{X},\tilde{Y}}$ are respectively the empirical covariance and correlation between $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n$.

Hint: You may denote the means of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n$ by $\bar{\tilde{X}}$ and $\bar{\tilde{Y}}$. Answer the following questions before you start this task: 1) What is the mean of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$. And 2) What is empirical standard deviation of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

b) *(Optional)* Let a, b, c, d be numbers, where $b \neq 0$ and $d \neq 0$. Let

$$\tilde{X}_i = a + bX_i, \qquad \tilde{Y}_i = c + dY_i, \quad i = 1, 2, \ldots, n.$$

Show that

$$s_{\tilde{X},\tilde{Y}} = bds_{X,Y}.$$

Hint: Results from Exercise 3 will be useful here.

c) *(Optional)* With the same definitions as b), show that

$$r^2_{\tilde{X},\tilde{Y}} = r^2_{X,Y}.$$

References

Lewin, R. (1980). Is your brain really necessary? *Science*, 210(4475):1232–1234.

Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. D. (1991). In vivo brain size and intelligence. *Intelligence*, 15(2):223–228.