A tutorial for Correlation Coefficients in R

Version 1.0.0

Konstantinos I. Bougioukas

04/11/2021

Contents

Objectives

- Applying hypothesis testing
- Investigate the possible association between two continuous variables
- Interpret the results

We will need the following packages for the notes:

```
library(GGally)
library(rstatix)
library(here)
library(tidyverse)
```

1 Introduction

Correlation is a statistical method used to assess a possible association between two continuous variables. It is measured by a statistic called the correlation coefficient, which represents the direction and strength of the association between the variables in question.

It is a dimensionless quantity that takes a value in the range -1 to +1. A positive correlation coefficient indicates that both variables increase (or decrease) in value together and a negative coefficient indicates that one variable decreases in value as the other variable increases and vice versa.

A correlation coefficient of **zero** indicates that no association exists between two continuous variables, and a correlation coefficient of **-1** or **+1** indicates a perfect negative or positive association, respectively. The strength of association can be anywhere between -1 and +1. The stronger the correlation, the closer the correlation coefficient comes to ± 1.

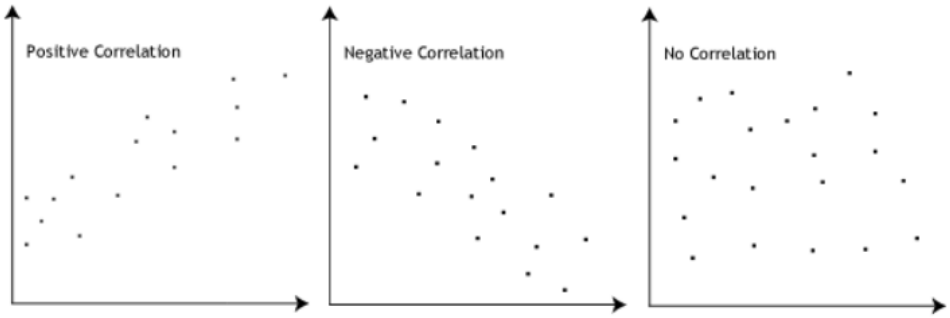


Figure 1: Correlation between two continuous variables

It is important to note that a significant association between two variables **does not** imply that they have a **causal** association.

On the other hand, a correlation coefficient that is not significant does not imply that there is no association between the variables because there may be a curvilinear or cyclical association.

Note An inherent problem of correlation coefficients is that the statistical significance of the test is often over-interpreted. For small samples it is possible to have a high correlation coefficient which is not significant. For large samples it is possible to have a small correlation coefficient without clinical importance which is statistically significant. Thus it is important to look at the value of r, the sample size as well as the p-value. In addition, outliers, the range of the data as well as the type of association between the two variables influence the correlation coefficient.

2 Types of correlation coefficients

There are several correlation coefficients measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear association between two variables. Other correlation coefficients – such as Spearman's rank correlation or Kendall's tau – have been developed to be more robust, that is, more sensitive to monotonic associations.

2.1 Pearson's correlation coefficient

Ho and H1 Hypotheses:

- H_0: there is not linear association between the two continuous variables ($\rho = 0$)
- H_1: there is a linear association between the two continuous variables ($\rho \neq 0$)

Appropriate correlation coefficient when:

- There is a linear association between the two continuous variables
- The two continuous variables (X, Y) are approximately normally distributed (bivariate normal distribution)
- There are no large outliers

The following guidelines have been proposed to interpet the strength of the association:

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Figure 2: Pearson's correllation interpretation

Note It is important to be clear that the Pearson correlation coefficient, r, does not represent the slope of the line of best fit. Therefore, if we get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It simply means that there is no variation between the data points and the line of best fit.

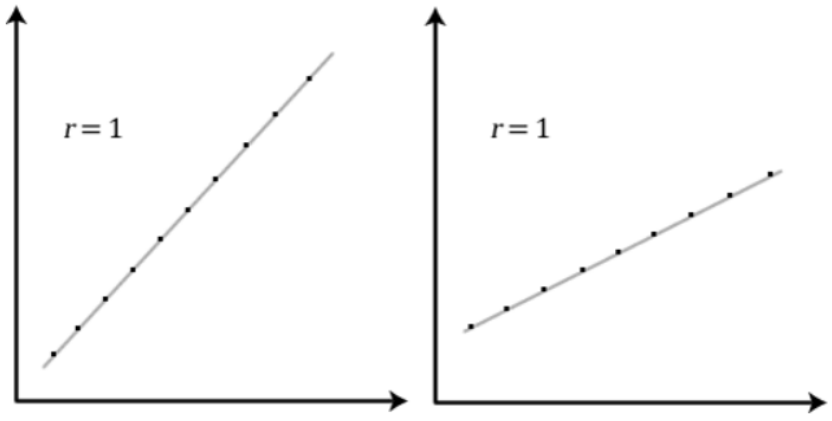


Figure 3: Pearson correlation coefficient does not represent the slope of the line of best fit

2.2 Spearman's correlation coefficient

- H_0: there is not (monotonic) association between the two ranked variables ($\rho = 0$)
- H_1: there is a (monotonic) association between the two ranked variables ($\rho \neq 0$)

Appropriate correlation coefficient when:

- The variables are continuous or ordinal
- There is a monotonic association between the two variables, meaning that as one variable increases, the other tends to either increase or decrease (not both)

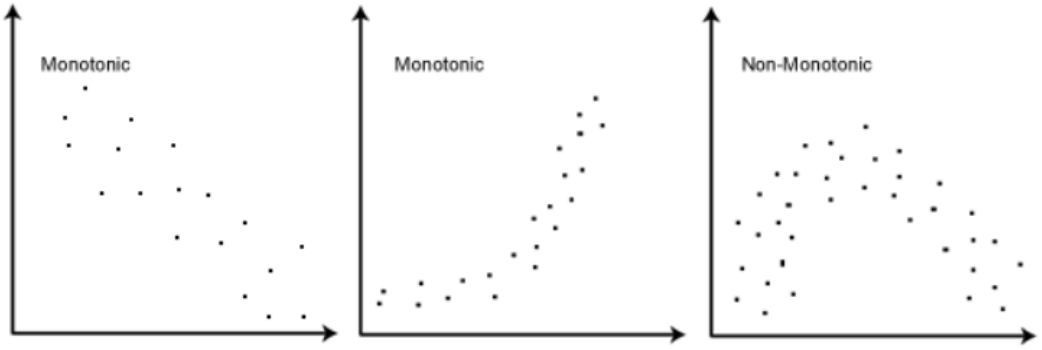


Figure 4: Monotonicity

Note Spearman's correlation coefficient is a non-parametric measure alternative to Pearson's r, it is based on the two ranked variables and is more robust to outliers.

3 Application in R

3.1 Dataset description

Data of 550 infants at 1 month age was collected (BirthWeight). The following variables were recorded:

- Body weight of the infant in kg (weight)
- Body height of the infant in cm (height)
- Head circumference in cm (headc)
- Gender of the infant (gender: Female, Male)
- Birth order in their family (parity: Singleton, One sibling, 2 or more siblings),
- Education of the mother (education: tertiary, year10, year12)

We import the data:

```
library(readxl)
BirthWeight <- read_excel(here("data", "BirthWeight.xlsx"), col_names=TRUE)
BirthWeight
```

Table 1: Birth Weight Data (first and last 5 rows)

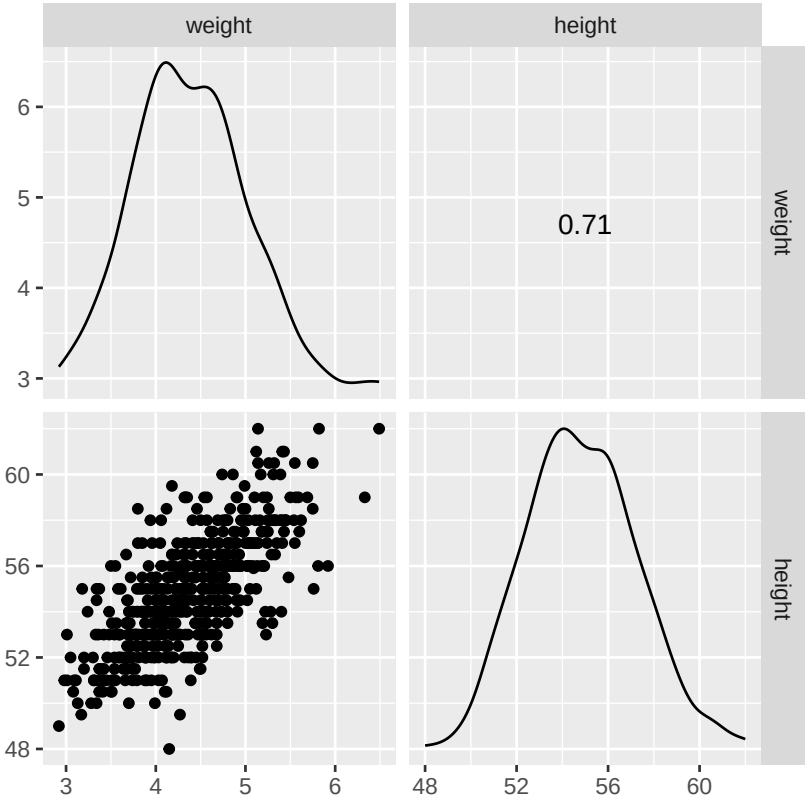
id	weight	height	headc	gender	education	parity
L001	3.95	55.5	37.5	Female	tertiary	2 or more siblings
L003	4.63	57	38.5	Female	tertiary	Singleton
L004	4.75	56	38.5	Male	year12	2 or more siblings
L005	3.92	56	39	Male	tertiary	One sibling
L006	4.56	55	39.5	Male	year10	2 or more siblings
NA	...	...	...	NA	NA	NA
W319	5.35	57	39.5	Male	tertiary	2 or more siblings
W320	5.39	60	40	Male	tertiary	Singleton
W321	3.88	52	36	Male	year10	One sibling
W322	5.23	57.5	40	Male	year10	2 or more siblings
W323	4.57	53.5	37.5	Female	tertiary	2 or more siblings

3.2 Correlation between two continuous variables

Let's say that we want to explore the association between weight and height for the sample of 550 infants of 1 month age. A first step that is usually useful in studying the association between two continuous variables is to prepare a scatter plot of the data. The pattern made by the points plotted on the scatter plot usually suggests the basic nature and strength of the association between two variables.

```
# correlation graph applying ggscatmat() function from GGally package

BirthWeight %>%
  select(weight, height) %>%
ggscatmat(corMethod = "pearson")    # alternative: "spearman", "kendall"
```



The above histograms show that the data are approximately normally distributed (we have a large sample so the graphs are reliable) for both weight and height.

Additionally, the points in the scatter plot seem to be scattered around an invisible line. The scatter plot also shows that, in general, infants with high height tend to have high weight (positive association). The Pearson's correlation coefficient $r = 0.71$, quantifies the strength of this association (alternatives are spearman and kendall).

We can also perform a correlation test:

```
BirthWeight %>%
  select(weight, height) %>%
cor_test(method="pearson")    # alternative: "spearman", "kendall"
```

var1	var2	cor	statistic	p	conf.low	conf.high	method
weight	height	0.71	23.813	1.4e-86	0.669	0.752	Pearson

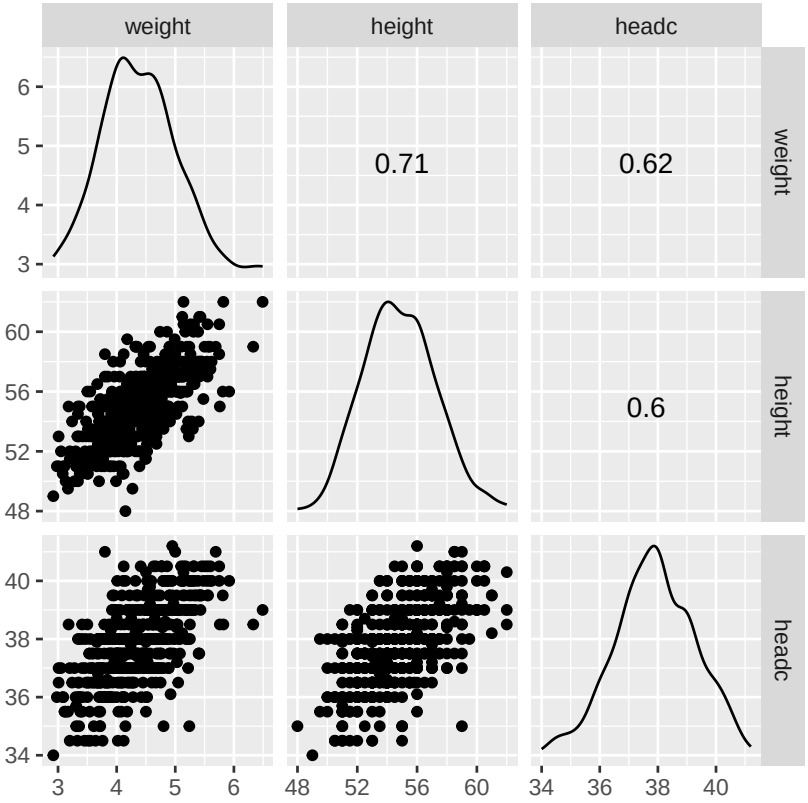
There is a significant high positive **linear** correlation (r=0.71, 95%CI: 0.67 to 0.75, p<0.001) between weight and height for infants of 1 month age.

3.3 Correlation for many pairs of continuous variables

- Plot many pairs of continuous variables

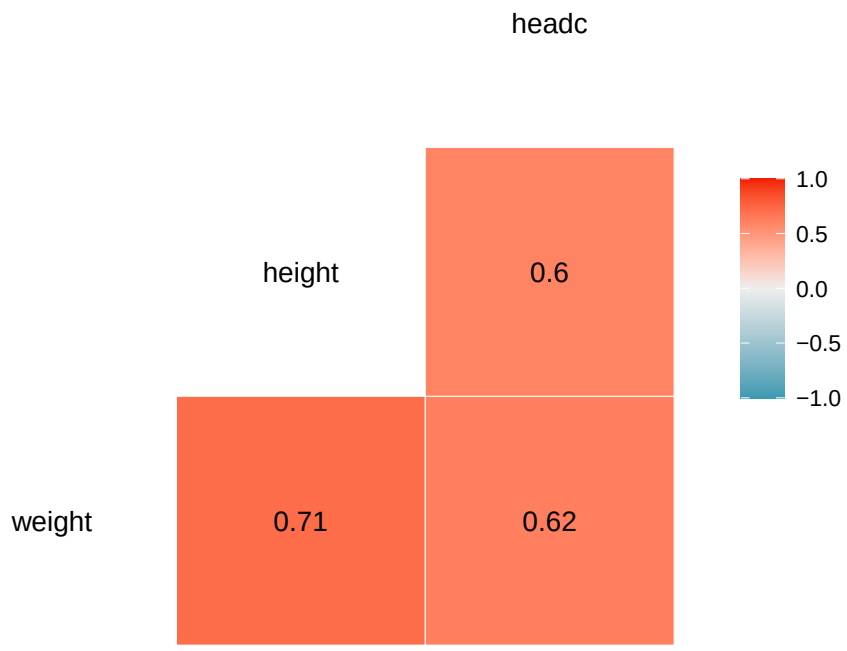
If we want to study the scatter plots and the association for many pairs of continuous variables, we can select them and run the `ggscatmax()` function:

```
BirthWeight %>%
  select(weight, height, headc) %>%
ggscatmat(corMethod = "pearson")  # alternative: "spearman", "kendall"
```



We can also create a correlation heat map chart:

```
BirthWeight %>%
  select(weight, height, headc) %>%
ggcorr(method = c("pairwise", "pearson"), label = TRUE, label_round = 2)
```



If there are many variables the use of a diverging color palette helps detect the strength and direction of the association between two variables.

Next, we perform the correlation tests:

```
BirthWeight %>%
cor_test(vars =c("weight", "headc"),
         vars2 = c("height", "headc"), method = "pearson") %>%
  filter(var1 != var2)

BirthWeight %>%
  select(weight, height, headc) %>%
  cor_mat(method = "pearson") %>%
  pull_lower_triangle()

BirthWeight %>%
```

```
  select(weight, height, headc) %>%
  cor_pmat(method = "pearson") %>%
pull_lower_triangle()
```

Table 2: Correlation Table

var1	var2	cor	statistic	p	conf.low	conf.high	method
weight	height	0.71	23.813	1.40e-86	0.669	0.752	Pearson
weight	headc	0.62	18.618	2.58e-60	0.568	0.671	Pearson
headc	height	0.60	17.478	1.08e-54	0.542	0.649	Pearson

Table 3: Lower correlation table with r

rowname	weight	height	headc
weight			
height	0.71		
headc	0.62	0.6	

Table 4: Lower correlation table with p

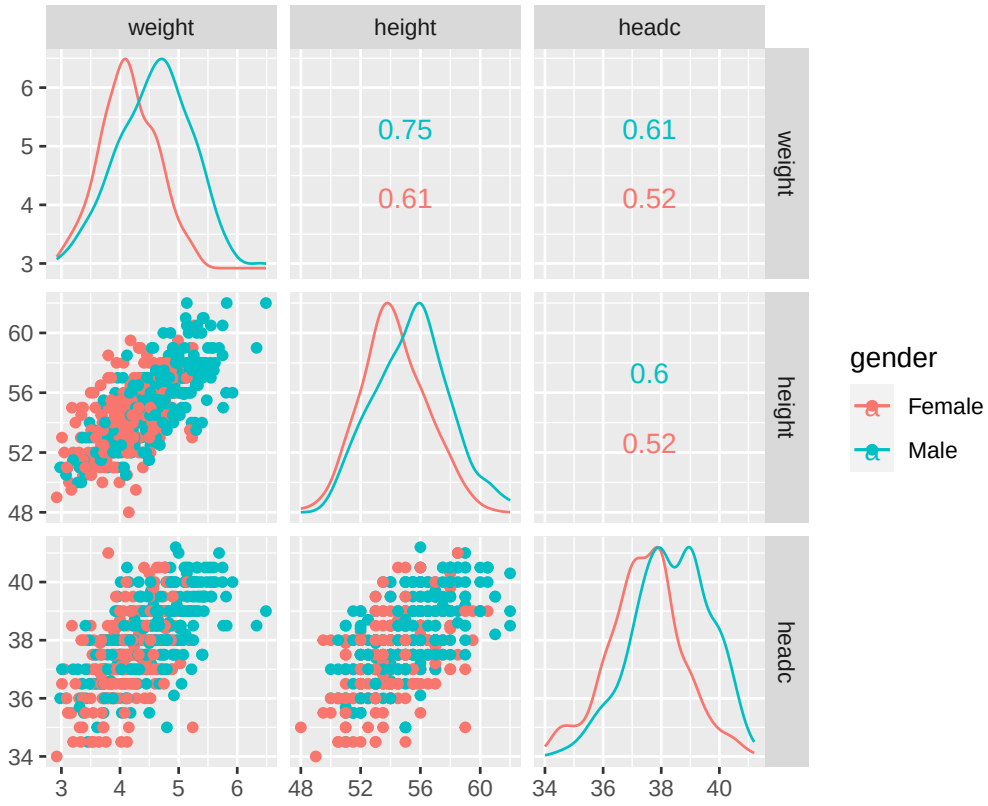
rowname	weight	height	headc
weight			
height	1.4e-86		
headc	2.58e-60	1.08e-54	

Note that we can also adjust the p-value to control for multiple comparisons using the `adjust_pvalue()` function if there are many comparisons.

- Plot many pairs of continuous variables according to a grouped variable

If we want to study the scatter plots and the association for many pairs of continuous variables according to a grouped variable (e.g., gender), we select the columns with the continuous variables and we add the `color` argument (e.g., `color = "gender"`):

```
BirthWeight %>%
ggscatmat(columns = 2:4, color = "gender", corMethod = "pearson")
```



Note that the `ggscatmat` function is set to convert the variable in the `color` argument to a factor automatically.