Correlation Analysis

Monica Franzese and Antonella Iuliano, Institute for Applied Mathematics "Mauro Picone", Napoli, Italy

Introduction

In the analysis of health sciences disciplines, usually we are interested to understand the type of relationship that exists between two or more variables. For example, the association between blood pressure and age, height and weight, the concentration of an injected drug and heart rate or the intensity of a stimulus and reaction time. Generally, the strength of relationship between two continuous variables is examined by using a statistical technique called correlation analysis. The concept of correlation was introduced by Sir Francis Galton (1877) in the mid 19th century as the most important contribution to psychological and methodology theory. In 1896, Karl Pearson published his first rigorous treatment of correlation and regression in the Philosophical Transactions of the Royal Society of London (Pearson, 1930). Here, he developed the Pearson product-moment correlation coefficient (PPMCC), using an advanced statistical proof based on the Taylor expansion. Today, the term correlation is used in statistics to indicate an association, connection, or any form of relationship, link or correspondence between two or more random variables. In particular, the PPMCC is used to study the linear relationship between two sets of numerical data. The correlation is connected to the concept of covariance that is the natural measure of the joint variability of two random variables. Usually, when we use measures of correlation or covariance from a set of data, we are interested in the degree of the correlation between them. In fact, we cannot prove that one variable causes a change in another if there are no connections between the two variables analyzed. An example is to test if the efficacy of a specific treatment is connected with the dose for the drug in a patient. In this case, a change in the drug variable changes the treatment variable, then the two variables are correlated. Therefore, correlation analysis provides information about the strength and the direction (positive or negative) of a relationship between two continuous variables. No distinction between the explaining variable and the variable to be explained is necessary. On the other hand, regression analysis is used to model or estimate the linear relationship between a response variable and one or more predictor variables (Gaddis and Gaddis, 1990). The simplest regression models involve a single response variable (dependent variable) and a single predictor variable (independent variables). For example, the blood pressure measured at the wrist depends on the dose of some antihypertensive drug administered to the patient.

In this work, our discussion is limited to the exploration of the linear (Person correlation) and non-linear relationship (Spearman and Kendall correlations) between two quantitative variables. In particular, we first introduce the concepts of covariance and correlation coefficient, then we discuss how these measures are used in statistics for estimating the goodness-of-fit of linear and non-linear trends and for testing the relationship between two variables. Finally, we explain how the regression analysis is connected to the correlation one. Simulated and real data sets are also presented for the identification and characterization of relationship between two or more quantitative variables.

Measures of Correlation Analysis

In this section, we discuss three correlation measures between two quantitative variables X and Y to determine the degree of association or correlation between them. For example, if we are interested to investigate the association between body mass index and systolic blood pressure, the first step for studying the relationship between these two quantitative variables is to plot a scatter diagram of the data. The points are plotted by assigning values of the independent variable X (body mass index) to the horizontal axis and values of the dependent variable Y (systolic blood pressure) to the vertical axis. The pattern made by the points plotted on the scatter plot usually suggests the nature and strength of the relationship between two variables. For instance, in **Fig. 1**, the first plot (A) shows that there is no relationship between the two variables X and Y, the second one (B) displays that exist a positive linear relationship between the two variables X and Y, the third one (C) exhibits a negative linear trend between the two variables X and Y. In the last two cases (B and C) the strength of linear relationship is the same but the direction is different, i.e., in (B) the values of Y increase as the values of X increase while in (C) the values of Y decrease as the values of X increase. In addition, the higher the correlation in either direction (positive or negative), the more linear the association between two variables.

Covariance

The covariance quantifies the strength of association between two or more sets of random variables.

Let X and Y be two random variable with the same population size N, we define the covariance as the following expectation value

$$\mathrm{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)] \tag{1}$$

doi:10.1016/B978-0-12-809633-8.20358-0

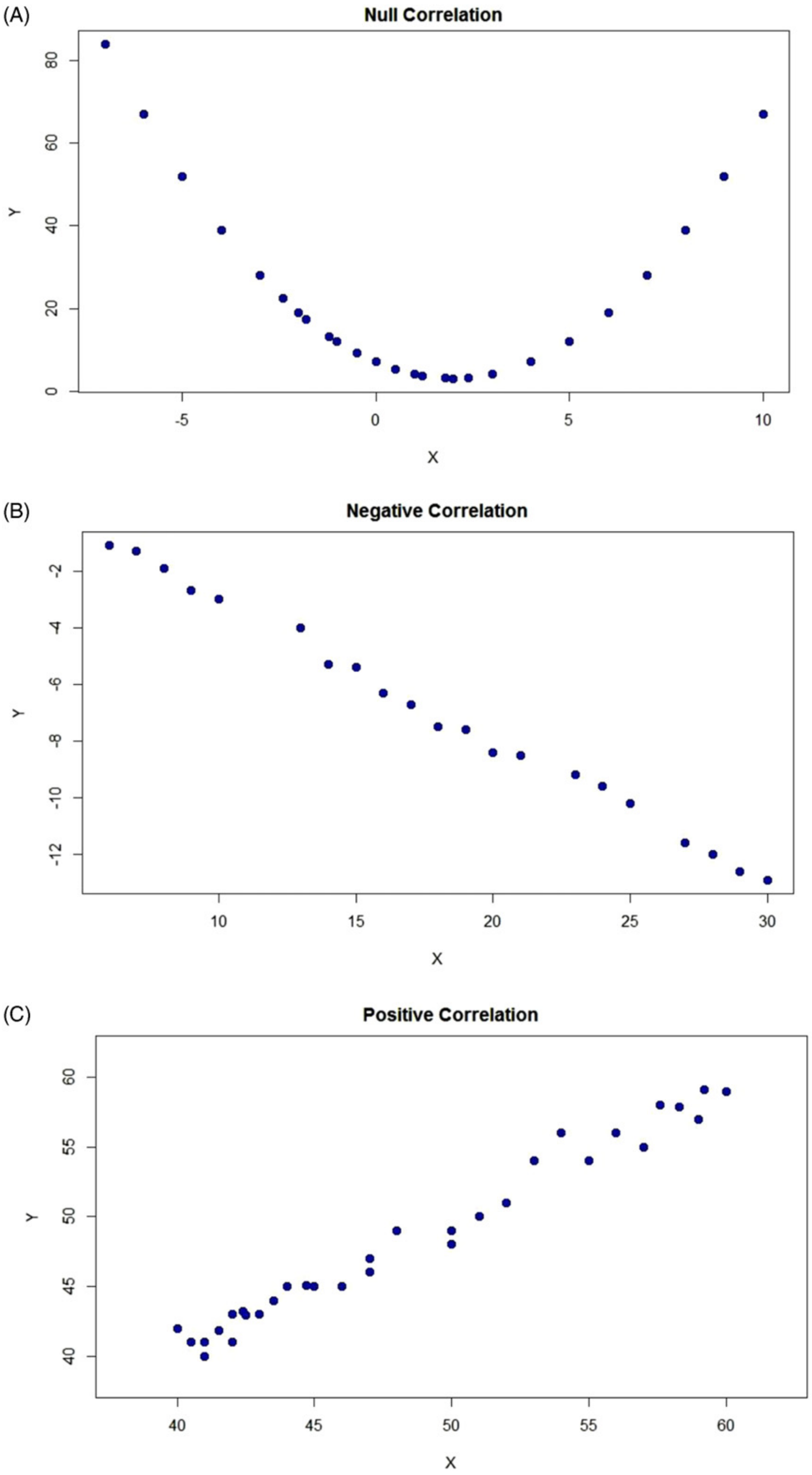

Fig. 1 The plot (A) shows that there is no relationship between the two variables *X* and *Y*; the plot (B) displays that exist a positive linear relationship between the two variables *X* and *Y*; the plot (C) exhibits a negative linear trend between the two variables *X* and *Y*.

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$ are the population means of X and Y, respectively. We often denote covariance by σ_{XY}. The variance is a special case of the covariance when the two variables X and Y are identical. In fact, let $\mu = E[X]$ the population mean of X. Then, the covariance is given by

$$\sigma_{XX} = \mathrm{Cov}(X, X) = E\left[(X - \mu_x)^2\right] = \mathrm{Var}(X) = \sigma_X^2 \tag{2}$$

By using the linearity property of expectations, formula (1) can be further simplified as follow

$$\begin{aligned}\sigma_{XY} &= \mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - XE[X] - E[X]Y + E[X]E[Y] \\ &= E[XY] - E[X]E[X] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y]\end{aligned}$$

In other words, the covariance is the population mean of the pairwise cross-product XY minus the cross-product of the means. The formula is given by

$$\sigma_{XY} = \mathrm{Cov}(X,Y) = \mu_{XY} - \mu_X\mu_Y \tag{3}$$

where μ_{XY} is the joint population mean of X and Y. Given a set of paired variables (x,y) with simple size n, the sample covariance is given by

$$\sigma_{xy} = \mathrm{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{4}$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

are the sample means of variable x and y, respectively. Positive values of covariance, $\sigma_{xy} > 0$, means that the variables are positively related, i.e., the terms $(x_i - \overline{x})(y_i - \overline{y})$ in the sum are more likely to be positive than negative; a negative covariance, $\sigma_{xy} < 0$, indicates that the variables are inversely related, i.e., the terms $(x_i - \overline{x})(y_i - \overline{y})$ in the sum are more likely to be negative than positive. If $\sigma_{xy} = 0$, then the variables x and y are uncorrelated or independent between them. In other words, the sample covariance is positive if y increases with increasing x, negative if y decreases as x increases, and zero if there is no linear tendency for y to change with x. An alternative formula of sample covariance is the following formula (similar to that for a sample variance),

$$\sigma_{xy} = \mathrm{Cov}(x,y) = \frac{n(\overline{xy} - \overline{xy})}{n-1} \tag{5}$$

where

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

is the joint sample mean of x and y. In probability theory, the covariance is a measure of the joint variability of two random variables. In particular, if X and Y are discrete random variables, with joint support S, then the covariance of X and Y is:

$$\mathrm{Cov}(X,Y) = \sum\sum\nolimits_{(x,y)\in S}(x_i - \mu_x)(y_j - \mu_y)f(x,y) \tag{6}$$

While on the contrary, if X and Y are continuous random variables with supports S_1 and S_2, respectively, then the covariance of X and Y is:

$$\mathrm{Cov}(X,Y) = \int_{S_1}\int_{S_2}(x_i - \mu_x)(y_j - \mu_y)f(x,y)dxdy \tag{7}$$

In Eqs. (6) and (7), the function $f(x,y)$ is a joint probability distribution, i.e., a probability distribution that gives the probability that X and Y fall in any particular range or discrete set of values specified by the two variables. For instance, we consider two discrete random variable X and Y (bivariate distribution) as illustrated in **Table 1**. By using the formula (6) the absolute value of covariance, is equal to 0.005. This value indicates a weak degree of association between X and Y.

Correlation Coefficients

In this section, we discuss the most widely used measures of associations between variables: Pearson's product moment correlation coefficient, Spearman's rank correlation coefficient and Kendall's correlation coefficient (Chok, 2010). The correct use of correlation coefficient depends on the type of variables. Pearson's product moment correlation coefficient is used only for continuous variables while the Spearman's and Kendall's correlation coefficients are adopted for either ordinal or continuous variables

Table 1 Joint probability mass function (or bivariate distribution) of two discrete random variables *X* and *Y*

x|y	*1*	*2*	*3*	$f_X(x)$
1	0.25	0.00	0.25	0.50
2	0.15	0.10	0.25	0.50
$f_Y(y)$	0.40	0.10	0.50	1

Note: The function $f_X(x)$ and $f_Y(y)$ are called marginal distributions. The mean of *X* and *Y* are equal to 1.5 and 2, respectively.

(Mukaka, 2012). In addition, the first correlation coefficient is used to quantify linear relationships, the last two correlation coefficients are applied for measuring non-linear (or monotonic) associations.

Pearson's product moment correlation coefficient

Pearson's product moment correlation coefficient r, called also linear correlation coefficient, measures the linear relationship between two continuous variables (Pearson, 1930). Let x and y be the quantitative measures of two random variables on the same sample of n. The formula for computing the sample Pearson's correlation coefficient r is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{8}$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

are the sample means of variable x and y, respectively. In other words, assuming that the sample variances of x and y are positive, i.e., $s_x^2 > 0$ and $s_y^2 > 0$, the linear correlation coefficient r can be written as the ratio of the sample covariance of the two variables to the product of their respective standard deviations s_x and s_y,

$$r = \frac{\mathrm{Cov}(x, y)}{s_x s_y} \tag{9}$$

Hence, the correlation coefficient is a scaled version of covariance. The sample correlation measurement r ranges between -1 and $+1$. If the linear correlation between x and y is positive (i.e., higher levels of one variable are associated with higher levels of the other) results $r > 0$, while if the linear correlation between x and y is negative (i.e., higher levels of one variable are associated with lower levels of the other) results $r < 0$. The value $r = 0$ indicates absence of any association (positive or negative) between x and y. The sign of the linear correlation coefficient indicates the direction of the association, while the magnitude of the correlation coefficient denotes the strength of the association. If the correlation coefficient is equal to $+1$ the variables have a perfect linear positive correlation. This means that if one variable increases, the second increases proportionally in the same direction. If the correlation coefficient is zero, no relationship exists between the variables. If correlation coefficient is equal to -1, the variables are perfectly negatively correlated (or inversely correlated) and move in opposition to each other. If one variable increases, the other one decreases proportionally. In addition, when two random variables X and Y are normally distributed, the population Pearson's product moment correlation coefficient is given by

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{10}$$

where σ_x and σ_y are the population standard deviations of X and Y, respectively. This coefficient is affected by extreme values and it is therefore not significant when either or both variables are not normally distributed.

Spearman's rank correlation coefficient

Spearman's correlation coefficient evaluates the monotonic relationship between two continuous or ordinal variables (Spearman, 1904). In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. Given two random variables x and y, Spearman's rank correlation coefficient computes the correlation between the rank of the two variables. The sample Spearman's rank correlation coefficient r_s is given by the following expression

$$r_s = \frac{\sum_{i=1}^{n}(x_i' - \overline{x}')(y_i' - \overline{y}')}{\sqrt{\sum_{i=1}^{n}(x_i' - \overline{x}')^2}\sqrt{\sum_{i=1}^{n}(y_i' - \overline{y}')^2}} \tag{11}$$

where x' is the rank of x and y' is the rank of y. In other words, it is a rank-based version of the Pearson's correlation coefficient. It ranges from -1 to $+1$. A strong monotonically increasing (or decreasing) association between two variables usually leads to positive (or negative) values of all correlation coefficients simultaneously. Moreover, for weak monotone associations, different correlation coefficients could also be of a different sign. Similar to the Pearson correlation coefficient, Spearman's correlation coefficient is 0 for variables that are correlated in a non-monotonic way. Unlike the Pearson's correlation coefficient, r_s is equal to $+1$ for both linearly and not linearly correlated variables. In addition, there is no requirement of normality for the variables. The corresponding population Spearman's rank correlation coefficient, denoted as ρ_s, describes the strength of a monotonic relationship. This coefficient is computed when one or both variables are skewed or ordinal and it is robust when extreme values are present. An alternative formula used to calculate the Spearman rank correlation is

$$r_s = 1 - \frac{6\sum_{i=1}^{2} d_i}{n(n^2 - 1)} \tag{12}$$

where d_i is the difference between the ranks of corresponding values x_i and y_i.

Kendall's correlation coefficient

Kendall's correlation coefficient τ is used to measure the monotonic association between two ordinal (not necessarily continuous) variables (Kendall, 1970). The formula to compute τ is

$$\tau = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{sign}\left(x_i - x_j\right)\operatorname{sign}\left(y_i - y_j\right)}{n(n-1)} \tag{13}$$

where

$$\operatorname{sign}(x_i - x_j) = \begin{cases} 1 \; if \; sign\left(x_i - x_j\right) > 0 \\ 0 \; if \; sign\left(x_i - x_j\right) = 0 \\ -1 \; if \; sign\left(x_i - x_j\right) < 0 \end{cases} \text{ and } \operatorname{sign}(y_i - y_j) = \begin{cases} 1 \; if \; sign\left(y_i - y_j\right) > 0 \\ 0 \; if \; sign\left(y_i - y_j\right) = 0 \\ -1 \; if \; sign\left(y_i - y_j\right) < 0 \end{cases}$$

This coefficient measures the discrepancy between the number of concordant and discordant pairs. Any pairs of ranks (x_i, y_j) and (x_i, y_j) are said to be concordant when $x_i < x_j$ and $y_i < y_j$, or $x_i > x_j$ and $y_i > y_j$, or $(x_i - x_j)(y_i - y_j) > 0$. Similarly, any pairs of ranks (x_i, y_j) and (x_i, y_j) are said to be discordant when $x_i < x_j$ and $y_i < y_j$, or $x_i > x_j$ and $y_i > y_j$, or $(x_i - x_j)(y_i - y_j) < 0$. As the two previous correlation coefficients, also τ ranges from -1 to $+1$. It is equal to $+1$ for concordant pairs and -1 for discordant pairs. Both Kendall and Spearman coefficients are formulated as special cases of the Person correlation coefficient.

Confidence Intervals and Testing Hypothesis

Suppose that x and y are two normally distributed variables (mean 0 and standard deviation 1), then the joint distribution is still a normal distribution with probability density

$$f(x) = \frac{1}{2\pi(1 - r^2)}\exp^{-\frac{x^2+2xy+y^2}{2(1-r^2)}}$$

where r is the linear correlation coefficient. This function is called bivariate normal distribution. If the form of the joint distribution is not normal, or if the form is unknown, inferential procedures are invalid, although descriptive measures may be computed. Under the assumption of bivariate normality, given a sample correlation coefficient r, estimated from a sample size of n, we are interested to test if two variables X and Y are linearly correlated, i.e., if $\rho \neq 0$. To estimate ρ (usually unknown), we use the sample correlation statistic r. In particular, we consider the following test statistics

$$r_{obs} = r\sqrt{\frac{n-2}{1-r^2}} \sim T_{(n-2)} \tag{14}$$

which is a statistics distributed as a Student's t variable with $n - 2$ degrees of freedom. To test if $\rho \neq 0$, the statistician and biologist Sir Ronald Aylmer Fisher (1921) developed a transformation of r that tends to become normal quickly as the population size n increases. Fisher's z-transformation is a function of r whose sampling distribution of the transformed value is close to normal. It is also called the r to z transformation and it is defined as

$$z : = 0.5\ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r) \tag{15}$$

where ln is the natural logarithm function and arctanh is the inverse hyperbolic tangent function. Then, z is approximately normally distributed with mean

$$\bar{z} = 0.5\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{16}$$

and standard error

$$\sigma_z = 1/\sqrt{n-3} \tag{17}$$

where n is the sample size. Fisher's z-transformation and its inverses

$$r = \frac{\exp(2z) - 1}{\exp(2z) + 1} = \operatorname{arctanh}(z) \tag{18}$$

can be used to construct a confidence interval for r using standard normal theory and derivations. If the distribution of z is not strictly normal, it tends to be normal rapidly as the sample size increases for any values of ρ. A confidence interval gives an estimated range of r values which is likely to include an unknown population parameter ρ. Generally, it is calculated at a confidence level, usually 95% (i.e., the significance level α is equal to 0.05). If the confidence interval includes 0 we can say that the

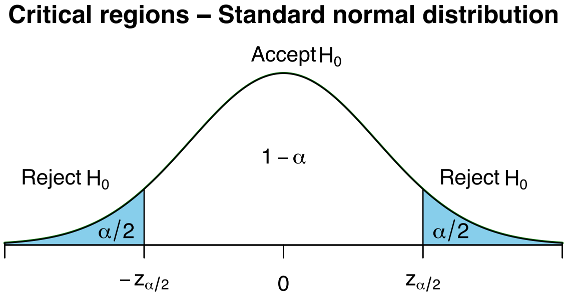

Fig. 2 Critical regions of the standard normal distribution.

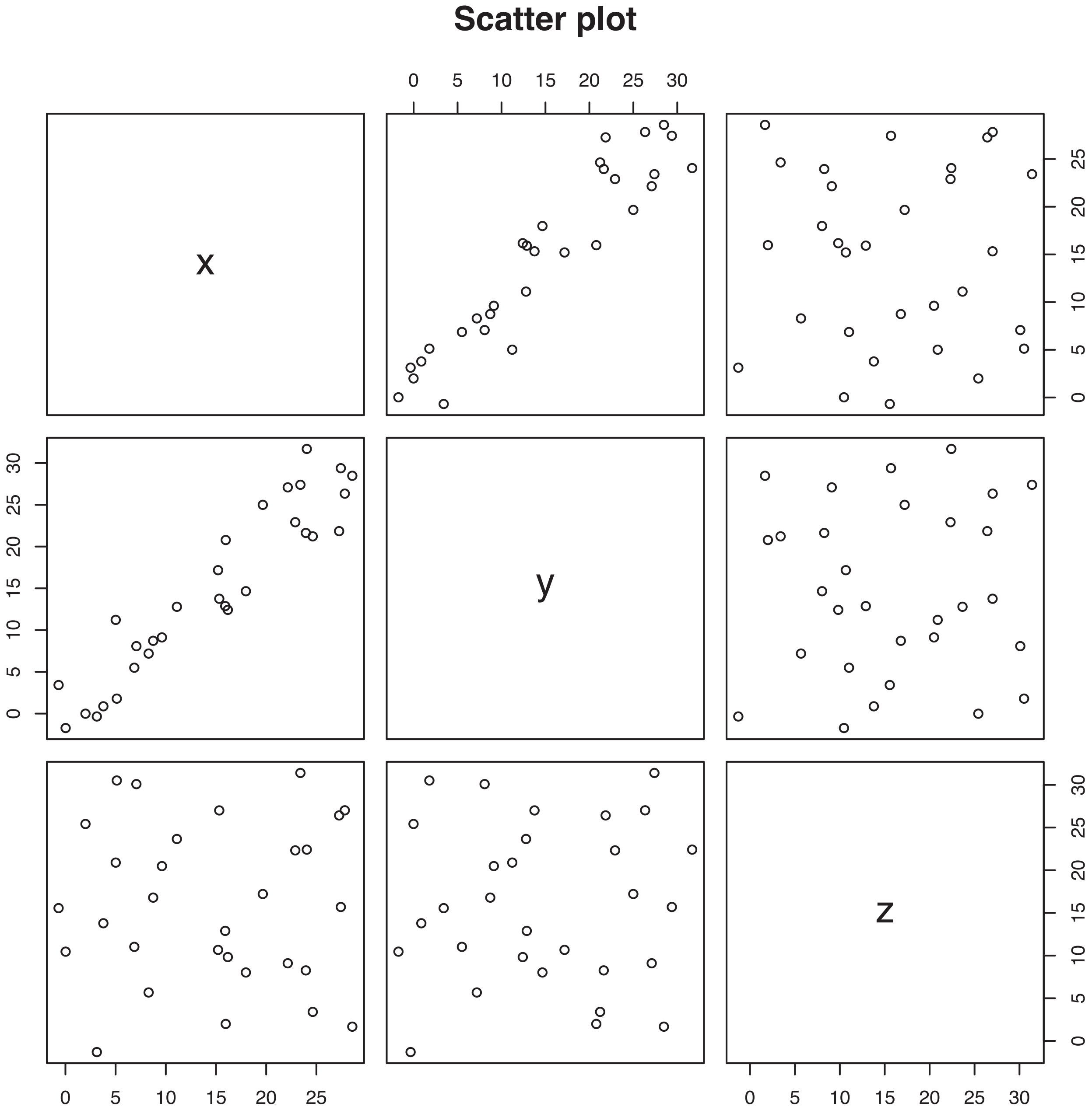

Fig. 3 Scatterplot shows a positive strong linear relationship between *x* and *y* variables (on the left of panel); while, it shows a weak association in the other cases; in particular, there is a negative correlation between variables *x* and *z* and a positive correlation between variables *y* and *z* (on the right of panel).

Table 2 Vital capacity data

Number	*Sex*	*Height*	*PEFR*	*vc*
1	1	180.6	522.1	4.74
2	1	168	440	3.63
3	1	163	428	3.40
4	1	171	536.6	3.75
5	1	177	513.3	3.81
6	1	169.4	510	2.80
7	1	161	383	2.90
8	1	170	455	3.88
9	1	158	440	2.40
11	1	161	461	2.60
11	1	163	370	2.72
12	1	155	503	2.20
13	1	171	430	3.38
14	1	171.5	442	2.99
15	1	167.6	595	3.06
17	1	166.6	455	3.06
18	1	167	500	3.72
19	1	163	548	2.82
20	1	172	463	2.83
21	1	155.4	475	3.06
22	1	165	485	3.07
23	1	174.2	540	4.27
24	1	167	415	3.80
25	1	162	475	2.88
26	1	172	490	4.47
27	1	161	470	3.40
28	1	155	450	2.65
29	1	162	450	3.12
30	1	174	540	4.02
31	1	161	475	2.80
32	1	166	430	3.69
33	1	166	510	3.66
34	1	161	470	2.56
35	1	168	430	2.78
36	1	167	470	3.48
37	1	166	440	3.03
38	1	164	485	2.90
39	1	162	550	2.96
40	1	176	535	3.77
41	1	166	485	3.50
42	1	160	360	2.30
43	1	161.2	480	3.39
44	1	167.8	480	3.70
45	2	181	580	–
46	2	170	560	–
47	2	171	460	–
48	2	184	611	–
49	2	184	600	–
50	2	188	590	–
51	2	186	650	–
52	2	187	600	–
53	2	181	630	–
54	2	181	670	–
55	2	177	515	–
56	2	167	470	–
57	2	182	550	–
58	2	172	620	–
59	2	190	640	–
60	2	178	680	–
61	2	184	600	–
62	2	170	510	–

Table 2 Continued

Number	*Sex*	*Height*	*PEFR*	*vc*
63	2	174	550	–
64	2	167	530	–
65	2	178	530	–
66	2	182	590	–
67	2	176	480	–
68	2	175	620	–
69	2	181	640	–
70	2	168	510	–
71	2	178	635	–
72	2	174	615.8	–
73	2	180.7	547	–
74	2	168	560	–
75	2	183.7	584.5	–
76	2	188	665	–
77	2	189	540	–
78	2	177	610	–
79	2	182	529	–
80	2	174	550	–
81	2	180	545	–
82	2	178	540	–
83	2	177	792	–
84	2	170	553	–
85	2	177	530	–
86	2	177	532	–
87	2	172	480	–
88	2	176	530	–
89	2	177	550	–
90	2	164	540	–
91	2	181	570	–
92	2	178	430	–
93	2	167	598	–
94	2	171.2	473	–
95	2	177.4	480	–
96	2	171.3	550	–
97	2	183.6	540	–
98	2	183.1	628.3	–
99	2	172	550	–
100	2	181	600	–
101	2	170.4	547	–
102	2	171.2	575	–

Note: The dataset contains the following variables: sex (1 = female, 2 = male), height, peak flow (PEFR), and vital capacity (vc) for 44 female and 58 male medical students.

population ρ is not significantly different from zero, at a given level of confidence α. More precisely, using Fisher's z transformation, we calculate the confidence interval at a confidence level α as following:

1. Given the observed correlation r, use Fisher's z transformation to compute the transformed sample correlation z, formula (15), the relative mean $\bar{z}$, formula (16), and standard deviation σ_z, formula (17).
2. Calculate the two-sided confidence limits (lower and upper) for z

$$\left(z - z_{\alpha/2}\sqrt{\frac{1}{n-3}}, z + z_{\alpha/2}\sqrt{\frac{1}{n-3}} \right) \tag{19}$$

   where the critical value $z_{a/2}$ depends on the significance level α. The value $\alpha/2$, is the area under the two tails of a standard normal distribution (see **Fig. 2**).
3. Un-transform the end points of the CI above using the arc tangent transformation, formula (19), in order to derive the confidence limits for the population correlation ρ as (r_{lower}, r_{upper}).

Confidence intervals are more informative than the simple results of hypothesis tests since they provide a range of possible values for the unknown parameter.

In order to perform hypothesis testing for the population correlation coefficient ρ, we first formulate the null and alternative hypothesis as follow

$$H_0 : \rho = 0 \text{ (}X\text{ and }Y\text{ are uncorrelated) versus } H_1 : \ \rho \neq 0 \text{ (}X\text{ and }Y\text{ are correlated).}$$

Then, we use the computed test statistic r_{obs}, formula (14), to find the relative p-value and to make a decision. If the p-value is smaller than the significance level α, then we reject the null hypothesis H_0 and accept the alternative H_1. In this case, we conclude that exists a linear relationship in the population between the two variables at the level α. If the p-value is larger than the significance level α, we accept the null hypothesis H_0. In this case, we deduce that there is no linear relationship in the population between the two variables at the level α. Typically, the significant level α is equal to 0.05 or 0.01. This approach is called p-value approach. An alternative method to make a decision is the critical value approach. We find the critical value $t_{a,n}$ using the Student's t distribution or t-table (where α is the significance level and n is the degree of freedom) and compare it to the observed r. If the test statistic is more extreme than the critical value, then the null hypothesis is rejected in favor of the alternative hypothesis. If the test statistic is not as extreme as the critical value, then the null hypothesis is not rejected. In other words, we reject the null hypothesis H_0 at the level α if $r_{obs} \leq t_{\alpha/2,n-2}$ and $r_{obs} \geq t_{\alpha/2,n-2}$. Vice versa, we accept the null hypothesis H_0 at the level α if $-t_{\alpha/2,n-2} \leq r_{obs} \leq t_{\alpha/2,n-2}$. For instance, see **Fig. 2**.

Note that the Fisher z-transformation and the hypothesis test explains above are mainly associated with the Person correlation coefficient for bivariate normal observations, but similar consideration can also be applied to Spearman (Bonett and Wright, 2000; Zar, 2014) and Kendall (Kendall, 1970) correlation coefficients in more general cases.

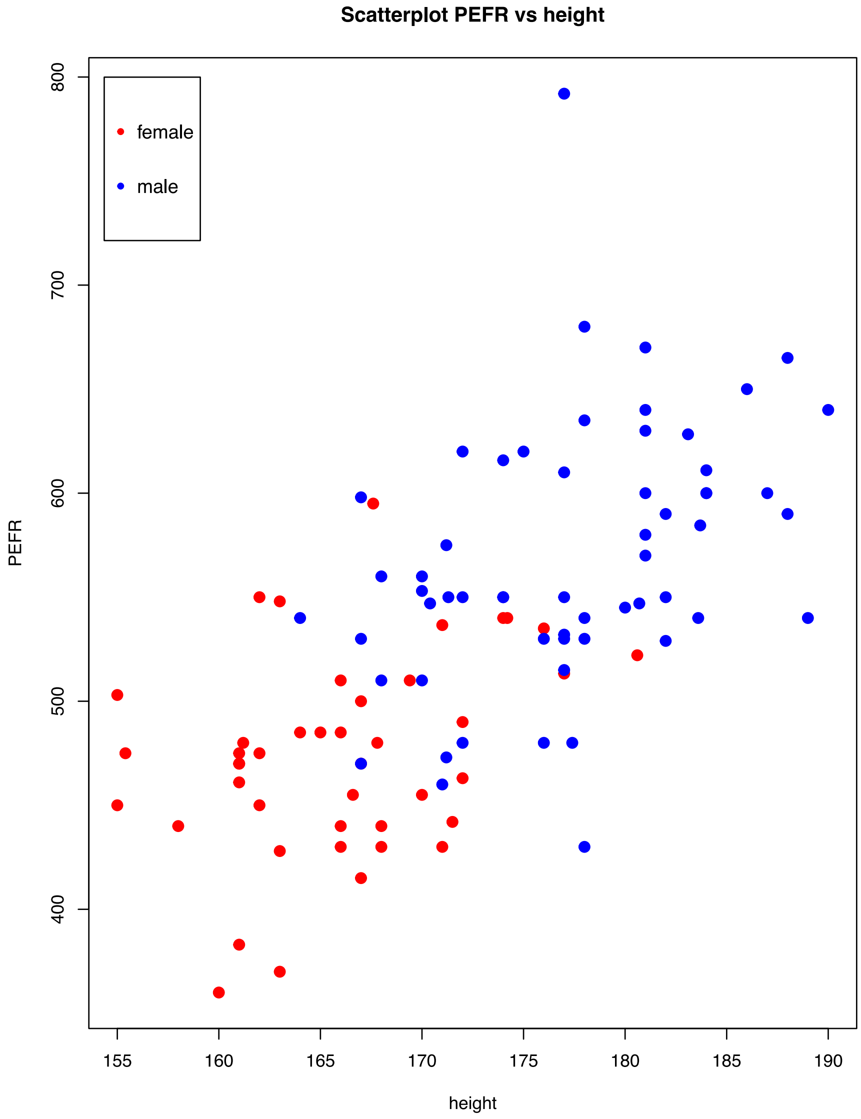

Fig. 4 Scatterplot between the peak expiratory flow rate (PEFR) and the height of medical students (male and female).

Table 3 In the table are reported the covariance and the three different correlation coefficients (Pearson, Spearman and Kendall) for the medical students dataset (female and male). The regression coefficients are also shown

		Height	*PEFR*
Covariance	height	69.02	396.20
	PEFR	396.20	5467.74
Pearson	height	1.00	0.64
Correlation	PEFR	0.64	1.00
Spearman	height	1.00	0.67
Correlation	PEFR	0.67	1.00
Kendall	height	1.00	0.48
Correlation	PEFR	0.48	1.00
Regression coefficients		$a = -461.84$, $b = 5.74$	

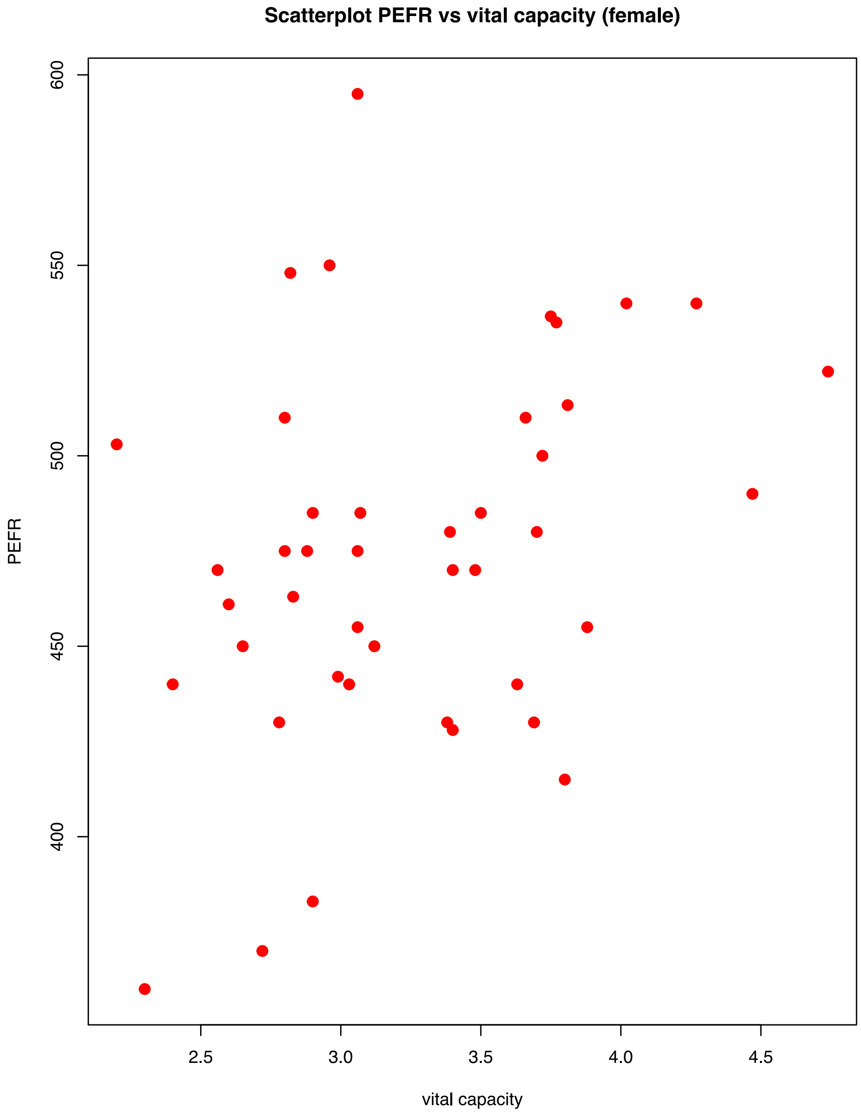

Fig. 5 Scatterplot between the vital capacity and the peak expiratory flow rate (PEFR) of female medical students.

Correlation and Regression

Regression analysis is used to model the relationship between a response variable (dependent variable) and one or more predictor variables (independent variables). Denoting by y the dependent variable and by x the independent variable, the simple linear correlation can be represented using the least squares regression line

$$y = a + bx + e \tag{20}$$

where a is the point where the line crosses the vertical axis y, and b shows the amount by which y changes for each unit change in x. We refer to a as the y-intercept and b as the slope of the line (or regression coefficient). The value e is the residual error. Letting $\hat{y} = \hat{a} + x\hat{b}$ be the value of y predicted by the model, then the residual error is the deviation between observed and the predicted values of the outcome y, i.e., $e = y - \hat{y}$. The aim of linear regression analysis is to estimate the model parameters a and b, in order to give the best fit for the joint distribution of x and y. The mathematical method of least-squares linear regression provides the best-fit solutions. Without making any assumptions about the true joint distribution of x and y, the least-squares linear regression minimizes the average value of the squared deviations of the observed y from the values predicted by the regression line $\hat{y}$. That is, the least-squares solution yields the values of a and b that minimize the mean squared residual, i.e., $e^2 = (y - \hat{y})^2$. The residual is the vertical distances of the data points. The least-squares regression line equation is obtained from sample data by simple arithmetic calculations. In particular, the estimations of a and b that

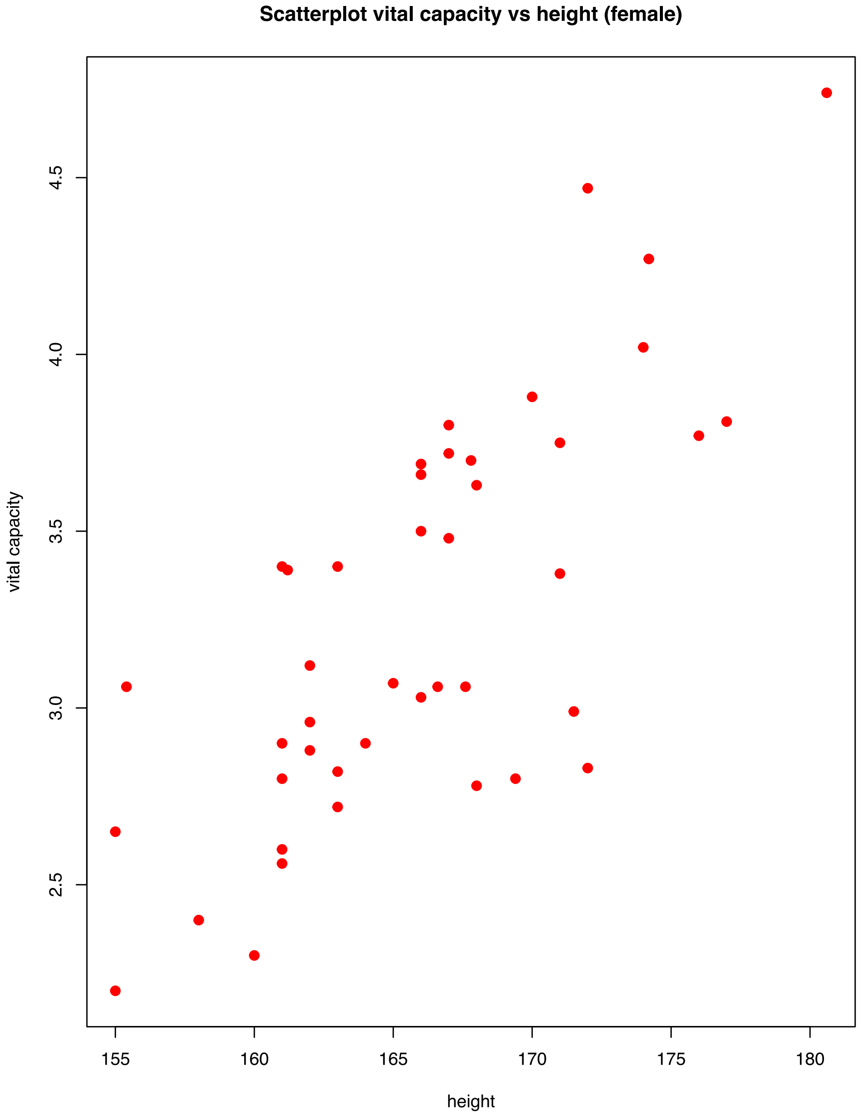

Fig. 6 Scatterplot between the height and the estimated vital capacity of female medical students.

minimizes the mean squared residual $e^2 = (y - \hat{y})^2$ are given by

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ \hat{a} = \overline{y} - b\overline{x} \tag{21}$$

where x_i and y_i are the corresponding values of each data point (X, Y), $\overline{x}$ and $\overline{y}$ are the sample means of the X and Y, respectively, and n the sample size. In other words, the estimate of intercept $\hat{a}$ and slope $\hat{b}$ are

$$\hat{b} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)},\ \hat{a} = \overline{y} - b\overline{x} \tag{22}$$

Thus, the least-squares estimators for the intercept a and slope b of a linear regression are simple functions of the covariances, variances and observed means and define the straight line that minimizes the amount of variation in y explained by a linear regression on x. In addition, the residual errors around the least squares regression are uncorrelated with the predictor variable x. Unlike the correlation analysis, that involves the relationship between two random variables, in the linear regression only the dependent variable is required to be random, while the independent variable is fixed (nonrandom or mathematical). In addition, as for the linear regression, in the correlation analysis we fit a straight line to the data either by minimizing

$$\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \text{ or } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In other words, we apply a linear regression of X on Y as well as a linear regression of Y on X. These two fitted lines in general are different. Also in this case we use the scatter diagram to plot the regression line. A negative relationship is represented by a falling regression line (regression coefficient $b < 0$), a positive one by a rising regression line ($b > 0$).

Data Analysis and Results

In this section, simulated and clinical data are presented for the study of correlation and regression analysis.

Simulated Data

We generate three random variables x, y, z from three bivariate normal distribution of 30 observations with means equal to zero and standard deviations 2, 3 and 4, respectively. For each pairs of variables, we first plot the scatter diagram in order to visualize the type of correlation that exists among them, and then we compute the covariance and the three correlation coefficients. Fig. 3 shows that the variables x and y have a positive covariance and a positive correlation, while the variables y and z have a negative covariance and a negative correlation. In particular, the Person correlation coefficient $r_{xy} = 0.94$ indicates that there is a strong positive association between the variables x and y, whereas Person correlation coefficient $r_{yz} = 0.03$ shows a weak positive relationship between the variables y and z. These results are confirmed by Spearman correlation coefficients $r_{s_{xy}} = 0.93$and $r_{s_{yz}} = 0.05$ and Kendall correlation coefficients $\tau_{xy} = 0.78$ and $\tau_{yz} = 0.04$. On the contrary, the variables x and z present a negative covariance and a negative correlation. In fact, Person correlation coefficient is $r_{xz} = -0.01$ which indicates a

Table 4 In the table are reported the covariance and the three different correlation coefficients (Pearson, Spearman and Kendall) for the female medical students dataset. The regression coefficients are also shown

		Height	*PEFR*	*Vital capacity*
Covariance	Height	34.30	100.16	2.50
	PEFR	100.16	2407.32	9.62
	Vital capacity	2.50	9.62	0.33
Pearson Correlation	Height	1.00	0.35	0.74
	PEFR	0.35	1.00	0.34
	Vital capacity	0.74	0.34	1.00
Spearman Correlation	Height	1.00	0.34	0.70
	PEFR	0.34	1.00	0.33
	Vital capacity	0.70	0.33	1.00
Kendall Correlation	Height	1.00	0.23	0.54
	PEFR	0.23	1.00	0.24
	Vital capacity	0.54	0.24	1.00
Regression coefficients	PEFR vs. vital capacity		$a = 380.10$, $b = 28.87$	
	Vital capacity vs. height		$a = -8.81$, $b = 0.07$	

weak negative correlation. This result is validated by Spearman correlation coefficient $r_{s_{xz}} = -0.02$ and Kendall correlation coefficient $r_{xz} = -0.02$. Finally, we test if these three correlation coefficients are statistically significant. Using parametric assumptions (Pearson, dividing the coefficient by its standard error, giving a value that follow a t-distribution), for the pair (x,y) the confidence interval at level 95% ($\alpha = 0.05$) is (0.88,0.97). In other words, we reject the null hypothesis H_0. In fact, the p-value$=1.118e-14$ is statistically significant, since it is less than 0.05. Similar results are obtained when data violate parametric assumptions (Spearman and Kendall). Also in these cases the p-value is statistically significant (p-value$=5.562e-08$ and p-value$=1.953e-12$). On the contrary, for the pairs (x,z) and (y,z), we accept the null hypothesis H_0 for the three type of correlation coefficients (Person, Spearman and Kendall) take into account. In fact, the three tests are not statistically significant (p-value>0.05) (Table 2).

Clinical Data

The clinical data is downloaded (See Section Relevant Websites). The data is composed by 44 female and 58 male medical students. It contains information about the peak expiratory flow rate (PEFR, measured in litre/min), which is a form of pulmonary function test used to measure how fast a person can exhale, the height of students and the vital capacity (vc), which is the maximum amount of air a person can expel from the lungs after a maximum inhalation (measured in litre). In particular, the vital capacity (vc) is reported only for female students. The scatter plot in Fig. 4 shows a positive correlation between the peak expiratory

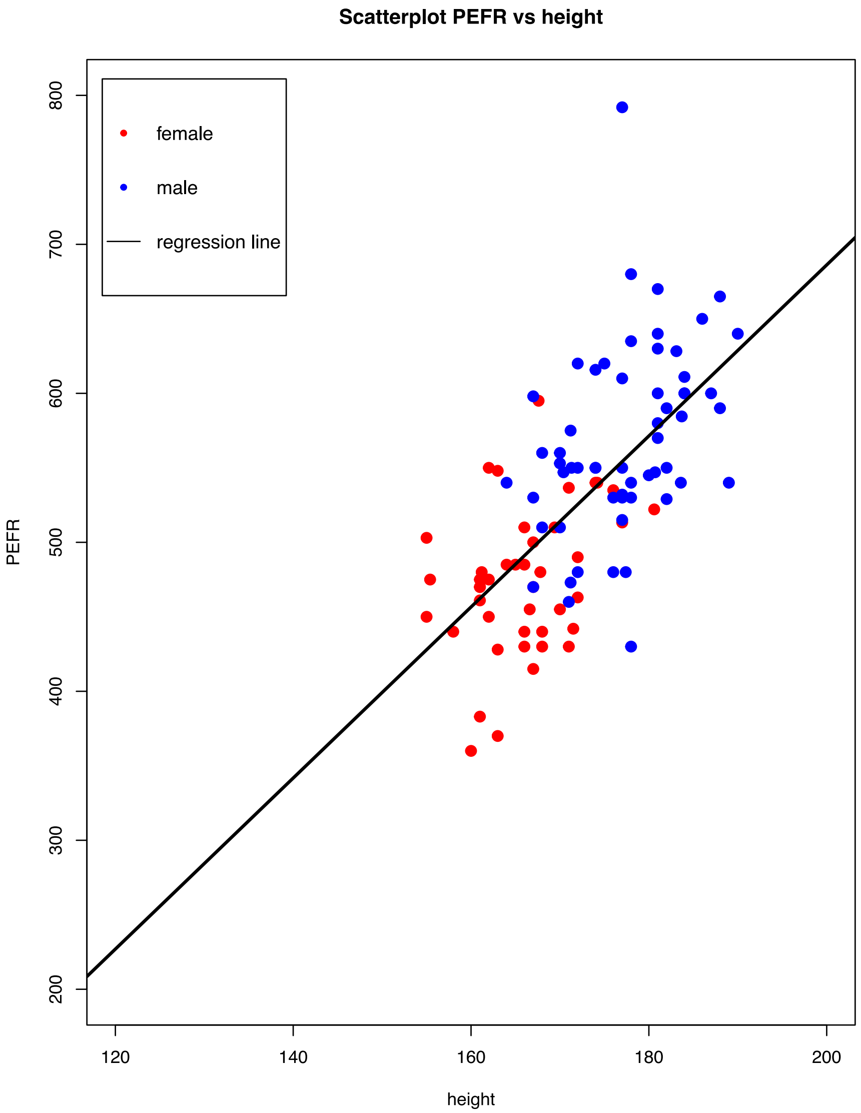

Fig. 7 Least squares line between the variables PEFR and height of the medical students (female and male).

flow rate (PEFR) and the height of all students (male and female). This result is confirmed by the computation of covariance and correlation coefficients (Person, Spearman and Kendall) as shown in Table 3. The Person and Spearman correlation coefficients indicate a moderate positive relationship ($r=0.64$ and $r_s=0.67$), while the Kendall correlation coefficient indicates a weak positive relationship ($\tau=0.48$). In particular, for the female group of medical students, Figs. 5 and 6 show a positive association for the variables (vc, PEFR) and (height, vc), respectively. Also in this case the positive relationship is validated by the computation of covariance and correlation coefficients (Person, Spearman and Kendall) as shown in Table 4. In particular, the three correlation coefficients indicate a weak positive relationship ($r=0.34$, $r_s=0.33$ and $\tau=0.24$) for the variables (vc, PEFR), while they indicate a strong positive relationship ($r=0.74$, $r_s=0.70$ and $\tau=0.54$) for the variables (height, vc). Finally, the regression analysis is also performed. In Fig. 7, we plot the least-squares line between the dependent variable PEFR and the independent variable height when female and male are considered together. The least-squares equation is $\hat{y}= -461.84+5.74\hat{x}$. On the other hand, in Figs. 8 and 9, the least-squares line between PEFR and vital capacity (vc) and vital capacity and height are plotted, respectively. The least-squares equations are $\hat{y}=380.10+28.87\hat{x}$ and $\hat{y}= -8.81+0.07\hat{x}$.

Software

We use the R statistical software (see Relevant Website section) to elaborate the strength of correlation and to analyze the regression relationship between the variables under investigated in both cases studies. In particular, we apply the common used statistical packages in R.

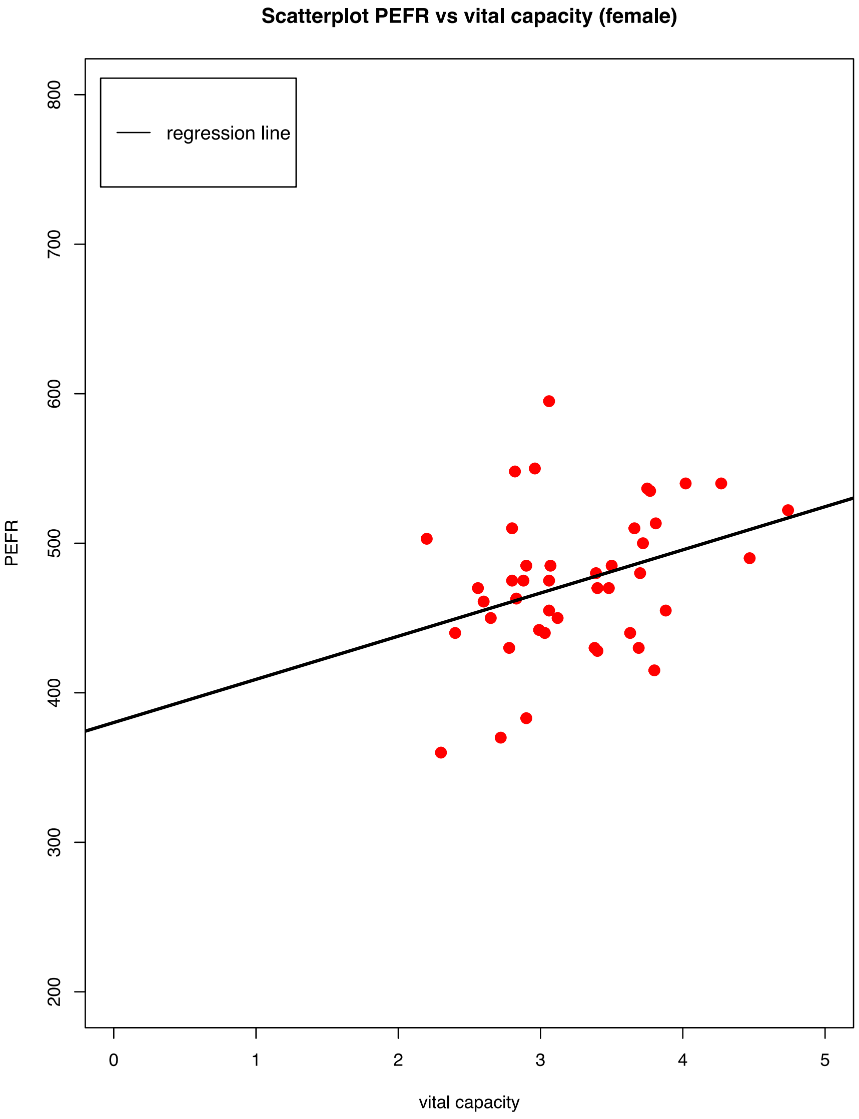

Fig. 8 Least squares line between the variables PEFR and vital capacity of the female medical students.

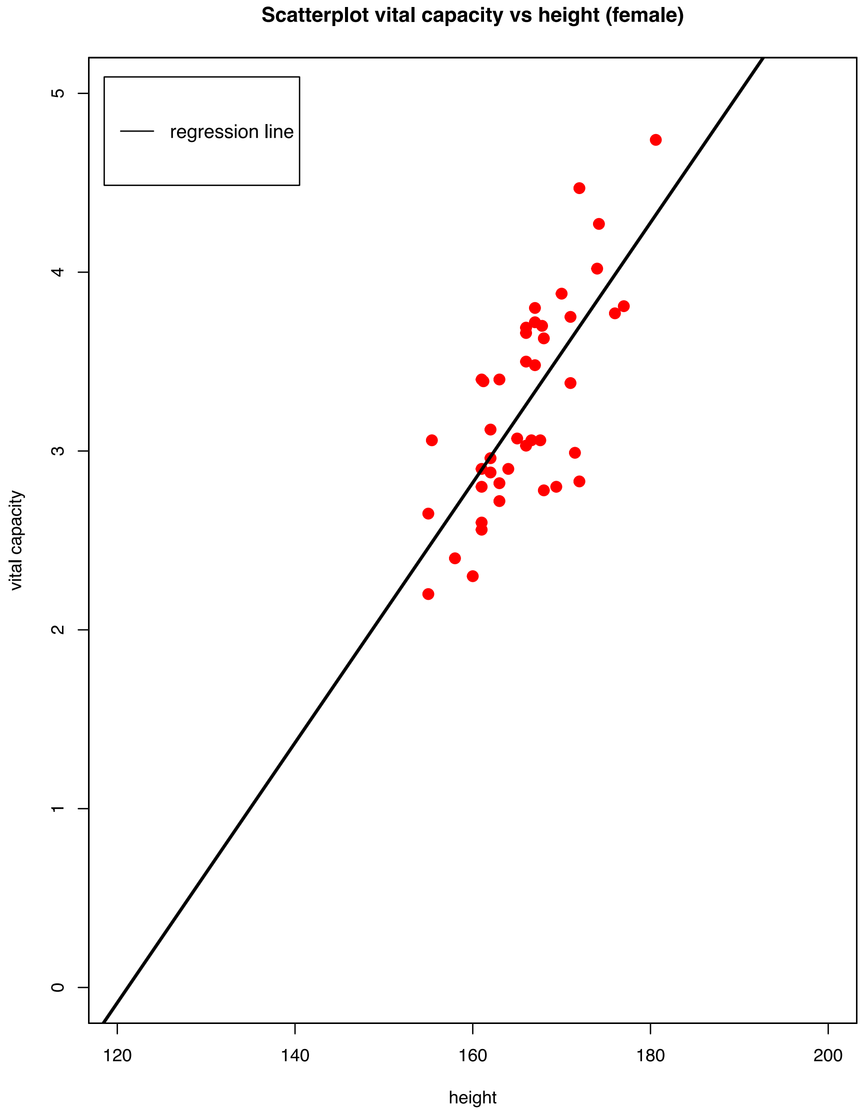

Fig. 9 Least squares line between the variables vital capacity and height of the female medical students.

Conclusions

Correlation analysis is an important statistical method for the analysis of medical data. It is used to investigate the relationship between two quantitative continuous variables. In particular, a correlation coefficient measures the strength of the relationship (magnitude) between two variables and the direction of the relationship (sign). We analyzed three different type of correlation coefficient. Pearson correlation coefficient quantifies the strength of a linear relationship between two variables. Spearman and Kendall correlation coefficients are two rank-based (or non-parametric) version of the Pearson coefficient. When two variables are normally distributed we use Pearson coefficient, otherwise, we apply Spearman or Kendall coefficient. Moreover, Spearman coefficient is more robust to outliers than is Pearson coefficient. Finally, since the correlation analysis does not establish if one variable is dependent and the other is independent, we introduce the regression analysis which is another statistical method used to describe a linear relationship between a depend variable (response or outcome) and one or more independent variables (predictors or explanatory variables). Therefore, the correlation analysis can be defined as a double linear regression of X on Y and of Y on X.

See also: Deep Learning. Introduction to Biostatistics. Natural Language Processing Approaches in Bioinformatics

References

Bonett, D.G., Wright, T.A., 2000. Sample size requirements for Pearson, Kendall, and Spearman correlations. Psychometrika 65, 23–28.
Chok, N.S., 2010. Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. Dissertation, University of Pittsburgh
Gaddis, M.L., Gaddis, G.M., 1990. Introduction to biostatistics: Part 6, correlation and regression. Annals of Emergency Medicine 19 (12), 1462–1468.
Kendall, M.G., 1970. Rank Correlation Methods, fourth ed. London: Griffin.
Mukaka, M.M., 2012. A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal 24 (3), 69–71.
Pearson, K., 1930. The Life, Letters and Labors of Francis Galton. Cambridge University Press.
Spearman, C., 1904. The proof and measurement of association between two things. American Journal of Psychology 15, 72–101.
Zar, J.H., 2014. Spearman Rank Correlation: Overview. Wiley StatsRef: Statistics Reference Online.

Further Reading

Bland, M., 2015. An Introduction to Medical Statistics. Oxford: Oxford University Press.
Cox, D.R., Shell, E.J., 1981. Applied Statistics. Principles and Examples. London: Chapman and Hall.
Daniel, W.W., Cross, C.L., 2013. Biostatistics: A Foundation for Analysis in the Health Sciences, tenth ed. John Wiley & Sons.
Dunn, O.J., Clark, V.A., 2009. Basic Statistics: A Primer for the Biomedical Sciences. John Wiley & Sons.
Gibbons, J.D., Kendall, M.G., 1990. Rank Correlation Methods. Edward Arnold.

Relevant Websites

https://www.users.york.ac.uk/∼mb55/datasets/datasets.htm
Selected Data-sets from Publications by Martin Bland.
https://www.r-project.org
The R Project for Statistical Computing.