Simple Correlation

Measures of central tendency and measures of variations are not the only descriptive statistics that we are using to get a picture of what a set of values look like. You have already learned that knowing the values of the one most representative score (central tendency) and a measure of spread dispersion (variability) is critical for describing the characteristics of the distribution.

However, sometimes we are interested in the relationship between variables; that is how a value of one variable changes when a value of another variable changes. The way we express this interest is through computation of simple correlation. ***Correlation*** techniques are useful procedures for measuring the strength of a relationship between two variables (X and Y). For example, correlation procedures can be used to address questions as: To what extent is age related to blood pressure? A correlation between two variables are sometimes referred to as a ***bivariate*** correlation

Correlation between two variables can be plotted graphically or summarized through the calculation of an index that describe the extent and direction of the relationship.

A visual picture of correlation: The Scatter diagram

The relationship between two variables that have been measured on ratio or interval scale can be displayed graphically on a scatter diagram. This type of graph that plots the values of one variable, say X, on the X axis and plots the values of a second variable, say Y, on the Y axis is called **scatter diagram.**

Scatter diagrams are useful in showing both the magnitude and direction of relationships. Figures on the next page shows scatter plots for various relationships.

How to draw scatter diagram (usually the computer can do that):

1- Draw the X-axis and the Y- axis. Usually, the X variable goes on the horizontal axis and the Y variable goes on the vertical axis.

2- Mark both axes with the range of values that you think to be the case of the data.

3- Finally, for each pair of values (such as 3.5 and 49.4), as shown in Fig 1 for example 1, we entered a dot on the chart by making a place where 49.4 falls in the X-axis and 3.5 falls in the Y-axis. The dot represents a data point, which is the intersection of the two values, as you can see in Fig 1.

When all the data points are plotted, what does such illustration tell us about the relationship between the variables?

The general shape of the collection of data points indicates whether the correlation is direct (positive) or indirect (negative).

- A positive slope occurs when the data points group themselves in a cluster from the lower left-hand corner on the X-axis and Y-axis through the lower right-hand corner.
- A negative slope occurs when the data points group themselves in a cluster from the upper left-hand corner on the X-axis and Y-axis through the upper right-hand corner.

Here are some scatter plots showing very different correlations where you can see how the grouping of the data points reflects the sign and strength of the relationship.

- Fig 2 shows a perfect direct correlation where r =1and the data points are aligned along straight line with a positive slope.
- Fig 3 shows a perfect indirect correlation where r = -1and the data points are aligned along straight line with a negative slope.
- Fig 4 shows scatter plot for strong positive correlation (r = 0.76). The data align themselves a long a positive slope.
- Fig 5 shows scatter plot for strong negative correlation (r = -0.82). The data align themselves a long a negative slope. From the upper left-hand corner on the chart to the lower right-hand corner.

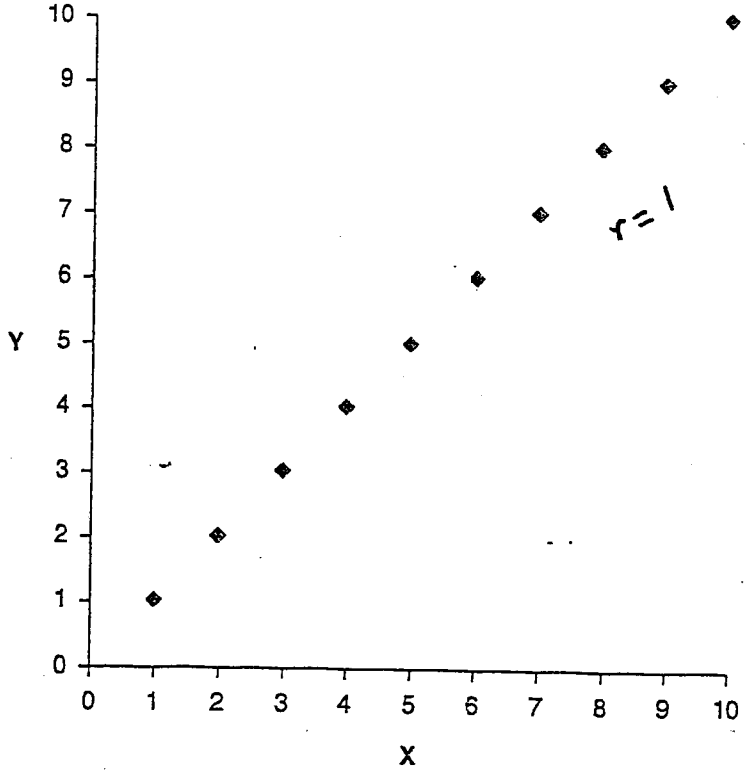
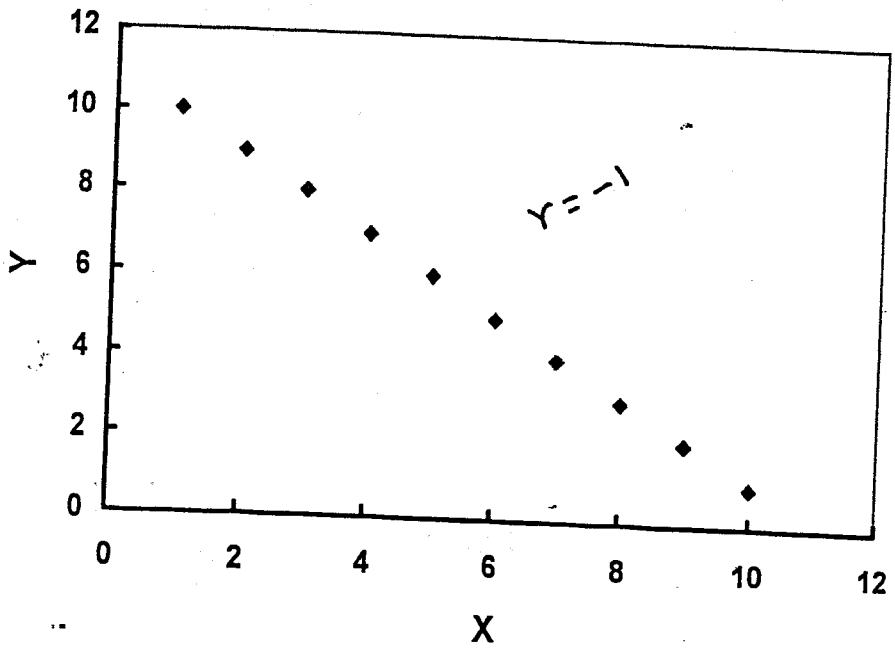


Figure 2 : A Perfect Direct, or Positive Correlation

Fig3 : **A perfect indrect, negative correlation**



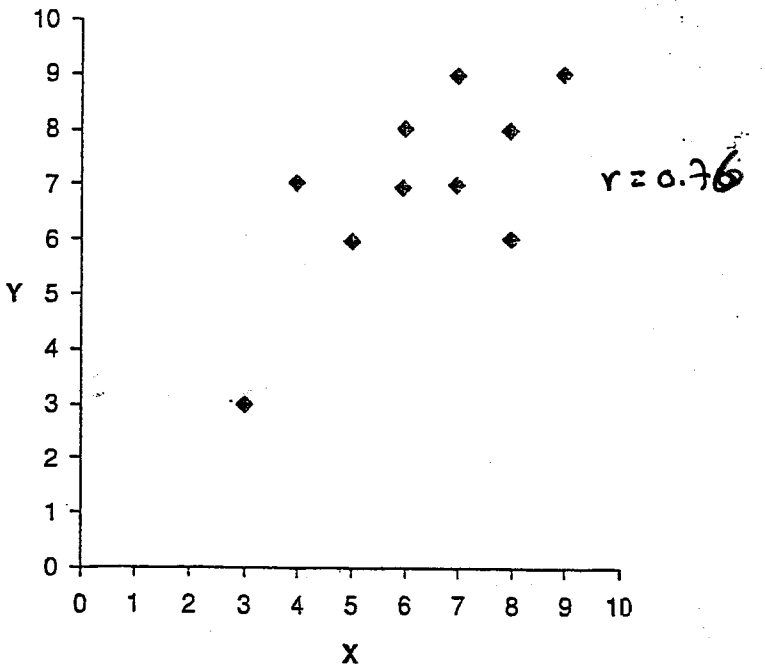


Figure 4 A Strong Positive, But Not Perfect, Direct Relationship

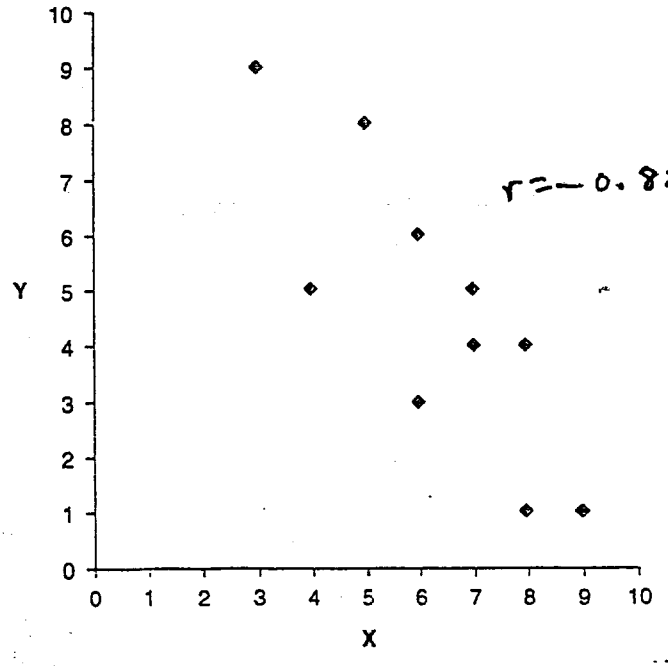


Figure 5 A Strong Indirect Relationship

(negative)

Example 1:

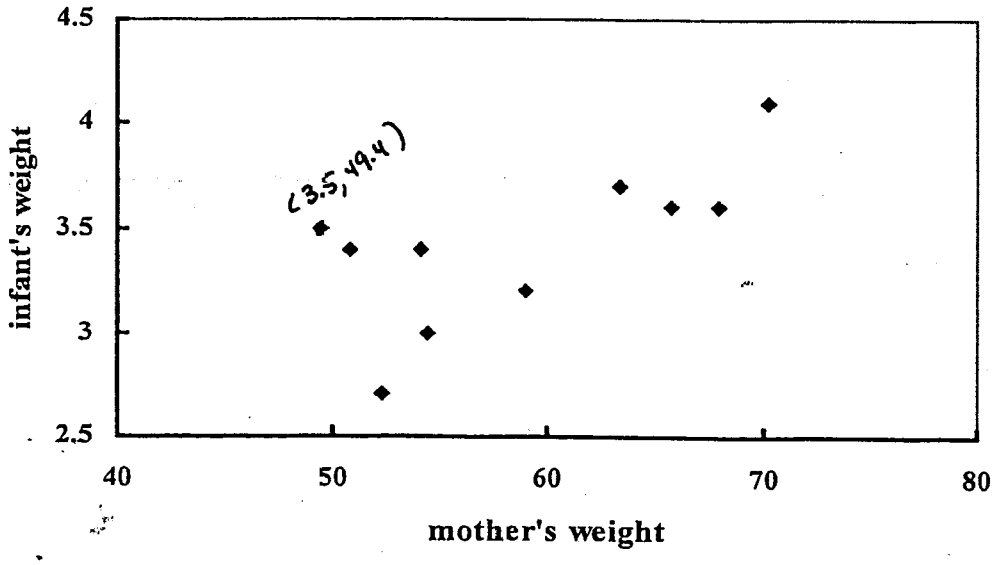
The below data were obtained on a study of the relationship between mothers' weight and birth weight of their infants. Let:

Y = infants birth weight (kg)

X= mother weight (kg)

Mothers weight (X)	Infant birth weight (Y)
49.4	3.5
63.5	3.7
68.0	3.6
52.2	2.7
54.4	3.0
70.3	4.1
50.8	3.4
65.8	3.6
54.1	3.4
59.0	3.2

The scatter pattern made by the plotted points usually suggests the basic form of the relationship between the two variables.



For example, the scatter diagram of the previous data revealed that when mother's weight increase infants weight also increase. High values of one variable, mother's

weight, are associated with high values of the other variable, infants' weight. Scatter diagram gives a visual impression of what the relationship between the two variables might be. We decide by visual inspection upon the kind of the line which best describes the overall pattern of the data.

Example 2:

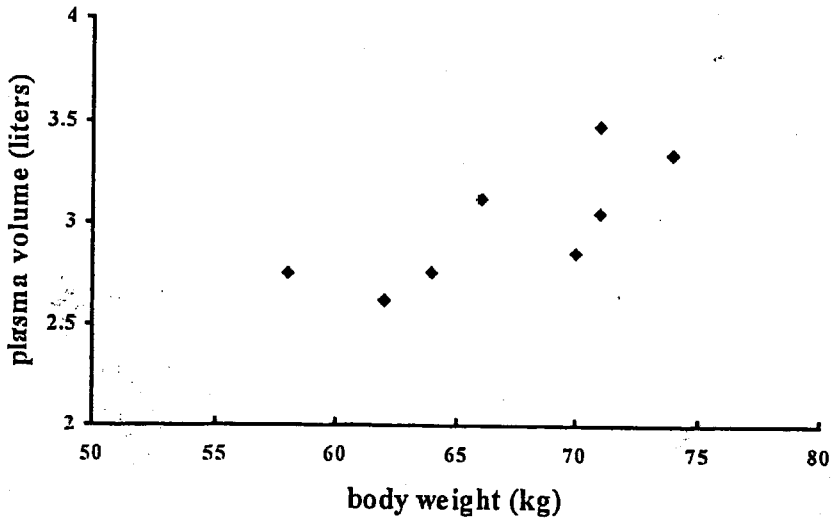
The table below shows plasma volume and body weight for eight healthy males.

Body weight (kg)	Plasma volume (l)
58	2.75
70	2.86
74	3.35
64	2.76
62	2.62
71	3.49
71	3.05
66	3.12
Total 536	24

$$\sum_{i=1}^{8} X_i = 536, \quad \sum_{i=1}^{8} Y_i = 24, \quad \sum_{i=1}^{8} X_i Y_i = 1616.94,$$

$$\sum_{i=1}^{8} X_i^2 = 36118, \sum_{i=1}^{8} Y_i^2 = 648.66$$

$$\overline{X} = 536/8 = 67, \quad \overline{Y} = 24/8 = 3$$



Researchers are much more likely to describe correlations by means of a statistic known as correlation coefficient. Correlation coefficient, like scatter plots, indicate both the magnitude and direction of a linear relationship between variables. Because they are expressed numerically, correlation coefficients are more precise about magnitude than scatter plots, to which we attach a more verbal labels such as "weak" or "moderate" or "strong".

Pearson correlation coefficient:

The most widely used correlation index is the Pearson correlation coefficient (known as Pearson's r). The Pearson correlation coefficient is employed with interval/ratio data to determine the association/correlation which assesses the relationship between two variables. We can some time apply Pearson correlation coefficient to rank-order data. The Pearson correlation coefficient determines the degree to which a linear relationship exists between two variables. The correlation coefficient can be calculated from the population or the sample. We denote the correlation coefficient calculated from the population by ρ (roh). We denote the correlation coefficient calculated from the sample by r.

The value of r is obtained using the formula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum y_i^2 - n\bar{y}^2$$

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)\left(\sum Y_i^2 - n\bar{Y}^2\right)}}$$

- The value of r can assume any value between –1 and +1, i.e. $-1 \le r \le 1$. Thus, the value of r can never be less than -1 or greater than +1.

- The absolute value of r (numerical value without any sign) indicates the strength of the relationship between the two variables. As the absolute value of r approaches 1, the degree of linear relationship between the two variables becomes stronger. The smaller the absolute value the weaker is the

relationship. For example, -0.9 indicates a very strong relationship, while 0.12 indicates a weak relationship.

- The correlation coefficient reflects the amount of variability that is shared between two variables and what they have in common. For example, you can expect an individual's height to be correlated with an individual's weight because they share many of the same characteristics, such as the individual's nutritional and medical history, general health and genetics.

Interpretation of the correlation coefficient

The sign of r indicates the nature or direction of the linear relationship which exists between the two variables.

1- If the correlation is greater than 0 (positive), subjects who have a high score on one variable will have a high score on the other variable and subjects who have a low score on one variable will have a low score on the other variable. The closer the positive value of r to +1, the stronger the direct relationship between the two variables, where as the closer the positive value of r to 0, the weaker the direct relationship between the variables (height and weight; IQ and GPA; cigarette consumption and heart disease risk, income and consumption). Two variables (X, Y) are positively correlated if as X increases, Y tends to increase, whereas as X decreases, Y tends to decrease.

2) If the correlation is less than 0, then the variables are said to be negatively or inversely correlated (an increase in one variable is associated with a decrease in the other variable and a decrease in one variable is associated with an increase in the other variable). When there is an inverse linear relationship, subjects who have a high score in one variable will have a low score on the other variable and vice versa). The close the negative r to -1, the stronger the inverse relationship between the two variables, whereas the closer the negative value of r to 0, the weaker the inverse relationship between the two variables (concentration of fluoride in drinking water and the prevalence of cavities in children's teeth, pulse rate and age, Two variables

(X, Y) are negatively correlated if as X increases, Y tends to decrease, whereas as X decreases, Y tends to increase.

3) If the correlation is close to 0, such as for birth weight and birthday, then the variables are said to be uncorrelated. Two variables (X, Y) are uncorrelated if there is no relationship between X and Y.

Using –Your _Thumb Rule:

1- Perhaps the easiest (but not the most informative) way of interpreting the value of a correlation coefficient is by eyeballing it and using the information in the table below:

Interpreting a correlation coefficient	
Size of a correlation	Coefficient general interpretation
0.8 – 1.0	Very strong relationship
0.6 – 0.8	Strong relationship
0.4 – 0.6	Moderate relationship
0.2 – 0.4	Weak relationship
0.0 – 0.2	Weak or no relationship

The Parson correlation is based on the following assumptions:

1- The sample of n subjects for which the value r is computed is randomly selected from the population of interest.
2- The level of measurement upon which each of the variables is based is interval or ratio data.
3- The two variables are normally distributed and any linear combination of the two variables is normally distributed.

Squaring the correlation coefficient (Coefficient of determination):

Coefficient of determination is the percentage of variance in one variable that is accounted for by the variance in the other variable. We previously stated that: variables that share something in common tend to be correlated with one another. If we correlate math and English grades for 100 intermediate school students, we will find the correlation to be moderately strong, because many of the reasons why

students do well (poor) in math tend to be the same reasons why they tend to well or poor in English. The number of hrs they study, how bright they are, how interested their parents in their school homework, the number of books they have at home and more are all related to math and English performance and account for differences between students.

The more these two variables share in common, the more they will be related. These two variables share variability- or the reason why students differ from one another. And on the whole, the brighter child who studies more will be better.

To determine exactly how much of the variance in one variable can be accounted for by the variance in another variable, the coefficient of determination is computed by squaring the correlation coefficient.

For example, if the correlation between GPA and number of hrs of study time is r = 0.70, then the coefficient of determination, represented by $r^2 = 0.70^2 = 0.49$. This means that 49% of the variation in GPA can be explained by the variation in study time. The stronger the correlation, the more of the variation can be explained.

Example 1:

A dentist conducted a study employing a sample of 10 children to determine whether or not there is a relationship between the number of ounces of sugar a ten-year old child eats per week (X) and the number of cavities in the child mouse (Y). The data is shown on the table below.

Ounces of sugar (X)	Number of cavities (Y)	XY	X^2	Y^2
20	7	140	400	49
0	0	0	0	0
1	2	2	1	4
12	5	60	144	25
3	3	9	9	9
5	4	20	25	16
10	5	50	100	25
2	1	2	4	1
8	5	40	64	25
5	3	15	25	9
$\bar{X} = 6.6$	$\bar{Y} = 3.5$	$\sum X_i Y_i = 338$	$\sum X_i^2 = 771$	$\sum Y_i^2 = 163$

$$r = \frac{338 - 10(6.6)(3.5)}{\sqrt{(771 - 10(6.6)^2}\sqrt{163 - 10(3.5)^2}} = 0.918$$

direct (positive) strong relation

It must be emphasized that when a researcher determines that two variables are correlated, this doe not imply that one variable caused the other. For example, if we determined that there is a negative correlation between self-esteem and depression, we cannot conclude that having low self-esteem causes people to become depressed. Nor can we conclude that being depressed reduces people's self-esteem. Either of these might be true, but it might also be that both are caused by some other factor (receiving a failing grade on an important test).

Steps in hypothesis testing.

It is a common practice to determine whether the value of the correlation coefficient r is large enough to allow us to conclude that the population correlation coefficient between the two variables other than 0. We wish to see if the sample value of r = 0.918 is of sufficient magnitude to indicate that in the population amount of sugar consumed by children and number of cavities are correlated.

1- Null and alternative hypothesis

$H_0: \rho = 0$ vs $H_1: \rho > 0$

2- Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

We substitute the values of r and n in the formula; we have r = 0.918 and n = 10

$$t = \frac{0.918\sqrt{10-2}}{\sqrt{1-0.918^2}} = \frac{2.596}{0.396} = 6.56$$

3- Critical value: we have $n - 2 = 10 - 2 = 8$ df; $\alpha = 0.05$, $1 - \alpha = 1 - 0.05 = 0.95$

$$t_{n-2,\,1-\alpha} = t_{10-2,\,0.95};\ t_{8,\,0.95} = 1.859$$

4- Decision: Since the calculated value (2.82) is greater than the critical value (1.859), we reject H_0.

5- conclusion: ounces of sugar consumed and number of cavities in children are correlated.

Example 2:

To study the relationship between age (years) and bone density (mg/cm^2) data was obtained for 7 subjects as shown in the below table:

Age (X)	Bone density (Y)	XY	X^2	Y^2
30	10.2	306	900	104.04
35	9.5	332.5	1225	90.25
47	9.1	427.7	2209	82.81
50	8.8	440	2500	77.44
60	7.9	474	3600	62.41
70	7.5	525	4900	56.25
80	7.1	568	6400	50.41
$\bar{X} = 53.14$	$\bar{Y} = 8.59$	$\sum XY =$ 3073.2	$\sum X^2 =$ 21734	$\sum Y^2 =$ 523.61

$$r = \frac{3073.2 - 7\,(53.14)\,(8.59)}{\sqrt{21734 - 7\,(53.14)^2}\,\sqrt{523.61 - 7\,(8.59)^2}}$$

Steps in hypothesis testing:

1- Null and alternative hypothesis

$H_0: \rho = 0$ vs $H_1: \rho < 0$

2- Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

substitute the values of r and n into the formula: In our case:

$$r = -0.987, \quad n = 7$$

$$t = \frac{-0.987\sqrt{7-2}}{\sqrt{1-(-0.987)^2}} = \frac{-2.21}{0.161} = -13.726$$

From the table of t distribution with $n - 2 = 7 - 2 = 5$ df

3- Critical value:

$$\alpha = 0.05, \; 1 - \alpha = 1 - 0.05 = 0.95$$

$$-t_{n-2,\,1-\alpha} = -t_{7-2,\,0.95}; \; -t_{5,\,0.95} = -2.015$$

4- Decision: Since the calculated value (-13.726) is less than the value obtained from the table (-2.015), we reject H_0.

5- Conclusion: we conclude that, in the population, age and bone density are linearly correlate.

2 Spearman Rank Correlation معامل سبيرمان للرتب

As in the case for the Pearson correlation coefficient; Spearman's rank-order correlation coefficient can be used to evaluate data for n subjects, each of whom has contributed a score on two variables(denoted as X and Y). Within each of the variables, the n scores are rank-ordered.

In computing Spearman's rank correlation coefficient, one of the following is true with regard to the rank- order data that are evaluated:

1- The data for both variables are in rank-order format

2- The original data orders are in a rank-order format for one variable and in an interval ratio format for the second variable. In such instance, data on the second variable are converted to rank-order format in order that both sets of data represent the same level of management.

3- The data for both variables have been transformed into a rank-order format from an interval ratio format, since we believed that one or more of the assumptions underlying Pearson correlation coefficient have been violated.

The same general guidelines that are described for interpreting the value of the Pearson correlation coefficient can be applied to Spearman's rank-order correlation coefficient.

To calculate Spearman rank correlation coefficient for a given set of data, we follow the steps:

1- Rank the values of X from low to high

2- Rank the values of Y from low to high

3- Compute d_i for each pair of observations by subtracting the rank of y from the rank of X.

$$\mathbf{d_i = Rank\ (X_i) - Rank\ (Y_i)}$$

4- Square each d_i and compute $\sum d_i^2$, the sum of the square values.

5- The Spearman rank correlation is given by:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where n is the number of pairs of X's and Y's. Note that the value of r_s is between 1 and -1.

Example 1:

The table below shows number of hrs studied by 10 students and the grades they obtained:

Number of hrs studied (X)	Grade in exam (Y)	Rank (X)	Rank (Y)	d_i	d_i^2
9	56	5	4	1	1
4	44	2	2	0	0
11	79	7	8	-1	1
13	72	8	7	1	1
10	70	6	6	0	0
5	54	3	3	0	0
18	94	10	10	0	0
15	85	9	9	0	0
2	33	1	1	0	0
8	65	4	5	-1	1
				total	4

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6(4)}{10(100-1)} = 0.97$$

Steps in hypothesis testing:

1-Null and alternative hypothesis

H_0: $\rho_s = 0$ $\quad$ H_1: $\rho_s > 0$

2- Test statistic:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

Substitute the values of n and $\mathbf{r_s}$ in the formula. We have $\mathbf{n = 10,\ r_s = 0.97}$

$$\mathbf{t = \frac{0.97\sqrt{10-2}}{\sqrt{1-0.97^2}} = 11.292}$$

3- Critical value: let $\alpha = 0.01$; **then** the critical value is 2.896. **reject** H_0 if $t \geq t_{1-\alpha}$

4- Decision: since t = 11.292 is greater than the critical value (2.896), we reject H_0.

5- Conclusion: There is a strong positive relationship between study time and scores in the exam.

Example 2:

To study the relationship between age (years) and bone density (mg/cm^2) data was obtained for 8 subjects as shown in the below table:

Age (X)	Bone density (Y)	Rank (X)	Rank (Y)	d_i	d_i^2
30	10.2	1	8	-7	49
35	9.1	2	6	-4	16
47	9.5	3	7	-4	16
50	8.8	4	5	-1	1
60	7.9	5	4	1	1
70	7.5	6	3	3	9
80	7.3	7	2	5	25
85	6.5	8	1	7	49
					$\sum d_i^2 = 166$

Substitute the values of n and i $\sum d_i^2$ into the formula for $\mathbf{r_s}$. From the given data:

n = 8 and $\sum d_i^2 = 166$; then:

$$r_s = 1 - \frac{6(166)}{8(64-1)} = \mathbf{-0.976}$$

Steps in hypothesis testing:

1- Null and alternative hypothesis

$H_0: \rho_s = 0$ vs $H_1: \rho_s < 0$

2- Test statistic:

$$t = \frac{-0.976\sqrt{8-2}}{\sqrt{1-(-0.976)^2}} = \mathbf{-12.665}$$

3- Critical value: let $\alpha = 0.05$; **then** critical value is -1.943; **reject** H_0 if : $t \leq -t_{1-\alpha}$

4- Decision: since t = -12.665 is less than -1.943, that is: We reject H_0.

5- Conclusion: There is a high negative relationship between bone density and age.

Correlation and causation:

Correlation is a measure of the degree to which two variables are related. A strong correlation between two variables, whether positive or negative, does not mean that one of the variables caused the other. For example, if the correlation between math score and high blood pressure is 0.93. This does not mean high blood pressure causes poor mathematics performance or that poor mathematics performance causes high blood pressure. The correlation of -0.93 shows that there is a very strong relationship between math test scores and systolic blood pressure, but that correlation tells us nothing about what causes that relationship. Correlation coefficient only indicates there is a relationship between two variables, but does not explain why the relationship occurs.

Things to remember:

- The correlation can range in value from -1 to +1.
- The absolute value of the coefficient reflects the strength of the correlation. So a correlation of -0.70 is stronger than a correlation of +0.5. A common mistake among students occurs when they assume that a positive correlation is stronger than a negative correlation.
- Pearson correlation is used with interval/ratio scale data.

Types of correlations and the corresponding relationship between variables				
What happens to variable X	What happens to variable Y	Type of correlation	Value	Example
X increases in value	Y increases in value	positive	0.0 – 1.0	The more time you spend studying, the higher will be your test score
X decreases in value	Y decreases in value	positive	0.0 – 1.0	The less money you put in the bank, the less interest you will earn
X increases in value	Y decreases in value	negative	– 1.0 – 0.0	The more you exercise, the less you will weigh
X decreases in value	Y increases in value	negative	– 1.0 – 0.0	The less time you take to complete a test, the more you will get more mistakes

Simple linear Regression and Correlation

Interpreting and plotting a line:

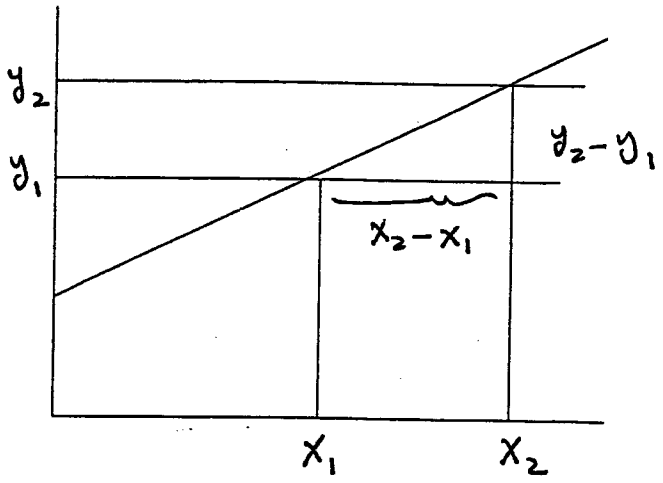
If we have two variables, say, X and Y. A straight line relationship between them can be expressed as:

$$Y = a + bX$$

Where:

a : is the intercept (where line intercepts or cross Y-axis

b : slope (amount of change in Y when X change by 1 unit).



Slope is a measure of how fast the line is rising or falling. The rate at which Y increase. Compared with X between any two points on that line.

$$\text{Slope} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

The slope may be positive or negative

We call Y as a linear function of X. The term linear equation arises from the fact that when the equation is plotted on ordinary graph paper, all pairs of X and Y which satisfy the equation of the form

$$Y = a + bX$$

yields points, which fall on straight line.

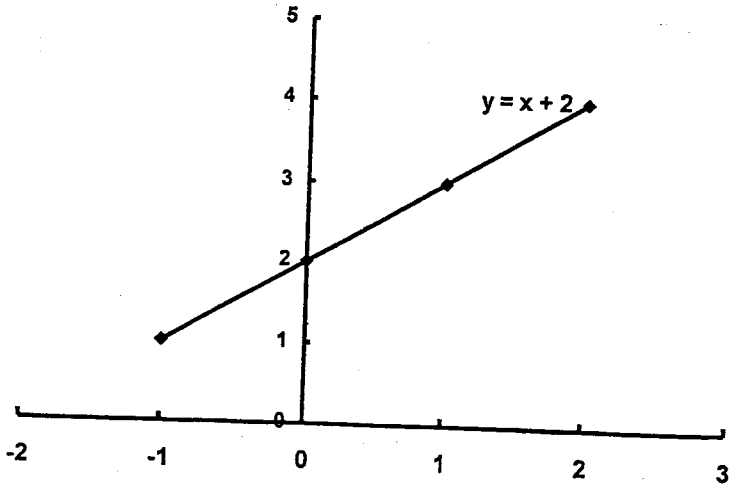
For the above linear equation, If we let a = 2 and b = 1, then:

$$Y = 2 + X$$

Any equation of the form $Y = a + bX$ has a graph. If we plot points on the (X, Y) plane, all the pairs of values satisfying the equation, they lie on the line. (0, 2), (-1, 1), (1,3) all satisfies the equation.

The above equation can be graphed using the following data:

X	-1	0	1	2
Y	1	2	3	4

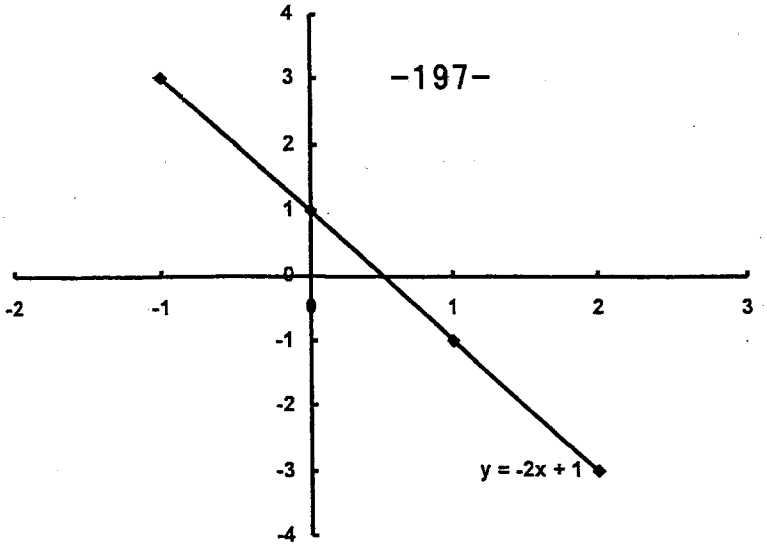


Slope is positive. A positive slope, such as b = 1 in our above equation, means that the line slants upwards to the right; that is, y increases as X increases. Y increases by 1 as increases by 1. The slope may be negative. If we let a = 1 and b = -2; we are going to have the following linear equation:

$$Y = 1 - 2X$$

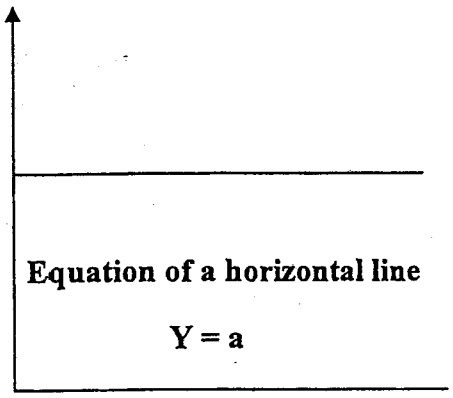
The above equation can be graphed using the below table.

X	-1	0	1	2
Y	3	1	-1	3



A negative slope, as Y = 1 – 2X, with b = -2, Y changes by -2 (decreases by 2) when X increases by 1.

A horizontal line has b = 0 as its slope because Y does not change as X increases. When b = 0, the equation Y = a + bX becomes Y = a. Thus, Y = a is the equation of a horizontal line.



Interpretation of the slope:

The cost of renting a car for one day is 80 SR plus a charge of 0.80 (80 h) for each mile the car is driven. The rental cost for a day (SR) is:

Cost = 80 + 0.8 x (number of miles driven).

80 = cost incurred no matter how many miles are driven. 0.8 implies that total cost increases by 0.8 for each additional mile driven.

Plotting a line:

1- We choose two values of x arbitrary.

2- For each value of X we compute a value of Y.

3- We plot each (X, Y) pair as a point on a graph and draw a line through the points.

For example, for our line :

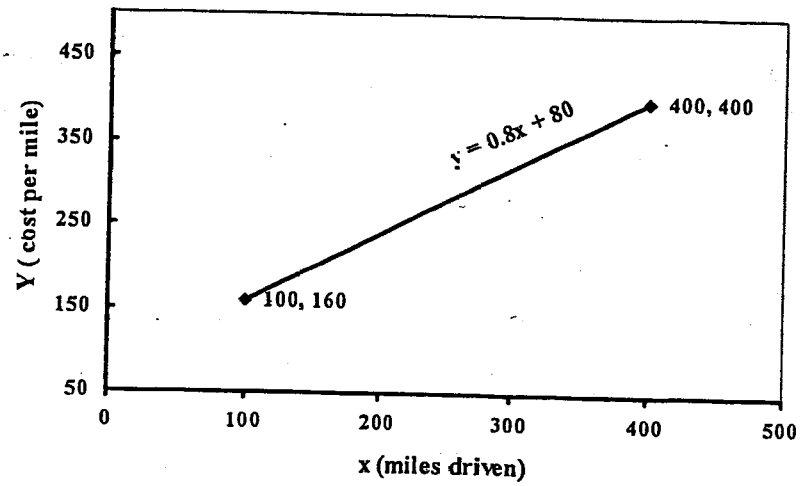
$$\textbf{Cost} = 80 + 0.8\ x$$

we choose $X_1 = 100$ and $X_2 = 400$. Next we compute the Y values:

For $X_1 = 100$, $Y_1 = 80 + 0.8\ (100) = 160$.

For $X_2 = 400$, $Y_2 = 80 + 0.8\ (400) = 400$

We have (X, Y) pairs as (100, 160), (400, 400), plot the points and draw the line through them, as shown in the figure below:



The main goals of many statistical techniques are to establish relationships, which make it possible to predict one or more variables in terms of others. For example:

1- Height and weight
2- Age and systolic blood pressure
3- Consumption and income
4- Amount of exercise and heart beat
5- Hrs of study and GPA

The nature and strength of the relationship between two variables may be examined by regression and correlation techniques.

Regression is used to discover the form of the relationship between two variables, say, X and Y, by defining an appropriate equation and to predict and estimate the value of

Y for a given value of X. Correlation is used to establish whether a relationship between two variables is strong or weak.

Simple linear regression:

In regression, a sample of n elements is taken from the population of interest and two variables X (height) and Y (weight) are measured on each of the chosen n elements obtaining:

$$(X_1, Y_1), (X_2, Y_2), \ldots\ldots, (X_n, Y_n)$$

Where, for example, X_1 (height of the first person) , Y_1 (weight of the first person), are paired since both are measured on the same element of the sample (first person). We usually selects the values of X, and is measured before the values of Y (For example, we measure age before we measure blood pressure).

The variable X is usually referred to as the independent variable (income, hrs of study, age, height) and Y as the dependent variable (expenditure, grade, blood pressure, weight) and we speak of the regression of Y on X or the regression of the dependent variable on te independent variable.

Scatter diagram:

A first step in studying the relationship between two variables is to prepare a scatter diagram of the data, which is obtained by plotting the pairs of measurements (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), with the values of the independent variable X on the horizontal axis and the values of the dependent variable Y on the vertical axis. Without examination of this scatter diagram it is impossible to analyze any relationship that might exists.

The scatter pattern made by the plotted points usually suggests the basic form of the relationship between the two variables.

Example:

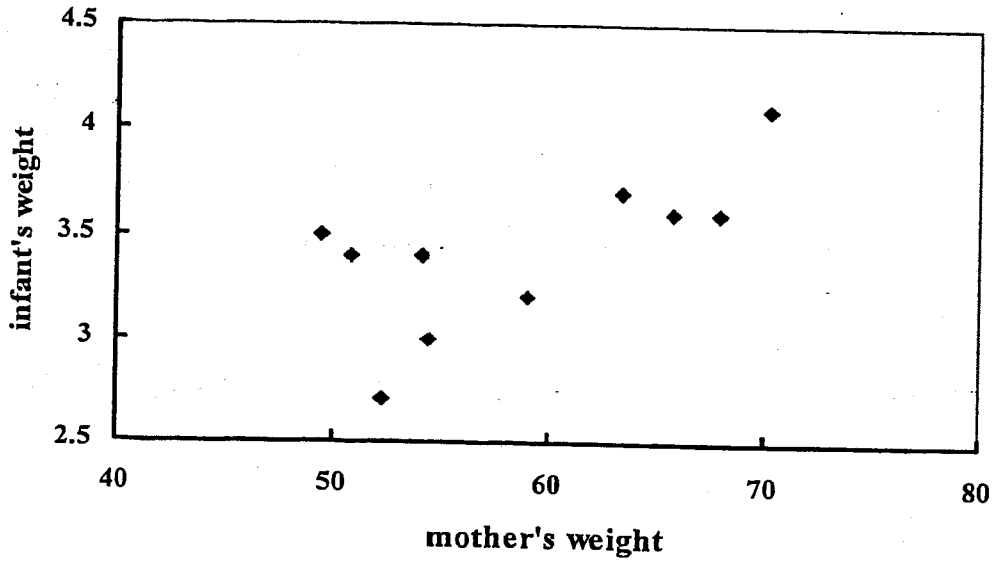
The below data were obtained on a study of the relationship between mothers' weight and birth weight of their infants. Let:

Y = infants birth weight (kg)

X= mother weight (kg)

Mothers weight (X)	Infant birth weight (Y)
49.4	3.5
63.5	3.7
68.0	3.6
52.2	2.7
54.4	3.0
70.3	4.1
50.8	3.4
65.8	3.6
54.1	3.4
59.0	3.2

The scatter pattern made by the plotted points usually suggests the basic form of the relationship between the two variables.



For example, the scatter diagram of the previous data revealed that when mother's weight increase infants weight also increase. High values of one variable, mother's weight, are associated with high values of the other variable, infants' weight. Scatter diagram gives a visual impression of what the relationship between the two variables might be. We decide by visual inspection upon the kind of the line which best describes the overall pattern of the data.

The best fitting line:

We know from the introduction of this chapter that, the equation of the straight line has the general form:

$$Y = a + b X$$

Where a is the y- intercept (where the line crosses y-axis), and b is the slope or amount Y changes per unit change in X.

For example, the equation that used to describe the exact relationship between the two common scales of measurement of temperature

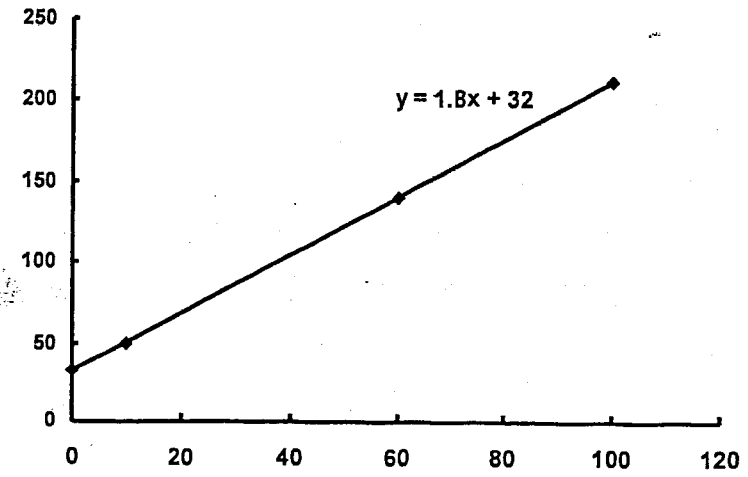
$$F = 32 + (1.8)\, C$$

Where C is the temperature in degrees Celsius or Centigrade (°C) and F is the same temperature of 32°F. For example °0C corresponds to a temperature of 32°F and 10°C corresponds to 50°F.

If this relationship were to be represented graphically for the data in the following table:

C	0	10	60	100
F	32	50	140	212

A straight line relationship as this is called linear. Thus every 1°C increase in temperature corresponds to an increase of 1.8 °F.



This equation implies an exact relationship, which in the case of two variables, say, X and Y, means that, given a value of X, a value of value can exactly be determined. When the two variables increase together, the relationship is aid to be direct and the value of b is **positive**. When the relationship is inverse, one variable increases and the other decreases. The value of b is **negative.**

Now, we go back to the example of the study of the relationship between mothers' weight and birth weight of their infants. From the scatter diagram, it is obvious that there is some form f relationship between the two variables, but this relationship cannot be described exactly by means of an equation as in the case of the temperature scales. No line could possibly pass through all the points in the scatter diagram.

However, some sort of equation could summarize this relationship and might be obtained by drawing a smooth line through the middle of the data points. We would like to draw a line tat best fit the data.

To avoid individual judgment in the fitting of the best line, it might be reasonable to define the best fitting for the data as the line where the distance between the line and the actual values (points) be as small as possible.

The line is called a regression line and its corresponding equation is called a regression equation. The regression equation can be expressed in the form:

$\hat{Y}$ predicted value → $\hat{Y} = a + bX$ ← regression eqn

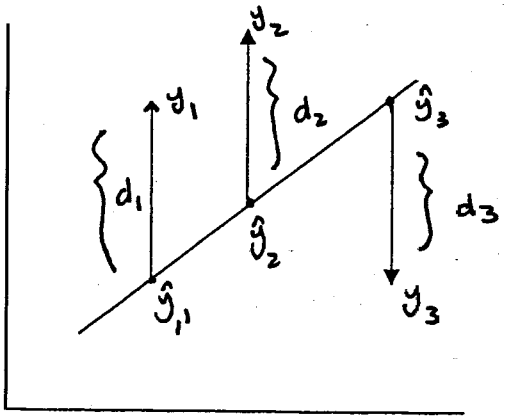
Where $\hat{Y}$ is the predicted or fitted value of a given observed Y. Where a is the y-intercept (where the line crosses y-axis), and b is the slope or amount Y changes per unit change in X.

The values of a and b are estimated from a given data and once they have been determined, we can substitute a value of X into the equation and calculate the corresponding value of Y.

How the values of a and b are obtained:

If we denote the distance between the actual value Y_i and the corresponding value $\hat{Y}_i$ by d_i, i.e $d_i = Y_i - \hat{Y}_i$

$d_1 = Y_1 - \hat{Y}_1$
$d_2 = Y_2 - \hat{Y}_2$
$d_3 = Y_3 - \hat{Y}_3$



We note, d_1 and d_2 are positive, while d_3 is negative. It might be reasonable to define the best fitting line as the line that produces the smallest d_i's, on the average. The problem with this definition is that some of the d_i's are positive and some are negative. It is possible that the average of the d_i's is 0, but non of the d_i's is 0. To avoid this problem, we square the d_i's and define the best fitting line as the line that produces the smallest sum of the squared deviations.

Criterion, which, is used today for determining "best fit", is known as the **method of least squares**.

The method requires the sum of the squares of the differences between the observed points $(X_1, Y_1), (X_2, Y_2), \ldots\ldots, (X_n, Y_n)$ and the fitted points on the line $(X_1, \hat{Y}_1), (X_2, \hat{Y}_2), \ldots\ldots\ldots (X_n, \hat{Y}_n)$ be as small as possible values of a and b are calculated from the following formulae:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{Y} - b\bar{X}$$

$$\text{Coefficient of Determination} = \frac{b\, S_{XY}}{S_{YY}}$$

Example:

From the data of mother's and infant's weight, the following information was obtained:

$$\sum_{i=1}^{10} X_i = 590, \quad \sum_{i=1}^{10} Y_i = 34, \quad \sum_{i=1}^{10} X_i Y_i = 2023.4, \quad \sum_{i=1}^{10} X_i^2 = 35280.84$$

$$\bar{X} = 590/10 = 59, \quad \bar{Y} = 34/10 = 3.4$$

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

substituting values from the information given, we obtain:

$$\frac{2023.4 - (10)(3.4)(59)}{35280.84 - (10)(59)^2} = \frac{17.4}{470.84} = 0.037$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 3.4 - 0.037\,(59) = 1.22$$

Thus, our sample regression line is:

$$\hat{Y} = \mathbf{a} + \mathbf{bX}$$

$$\hat{Y} = 1.22 + 0.037\,X \text{ (mother weight)}$$

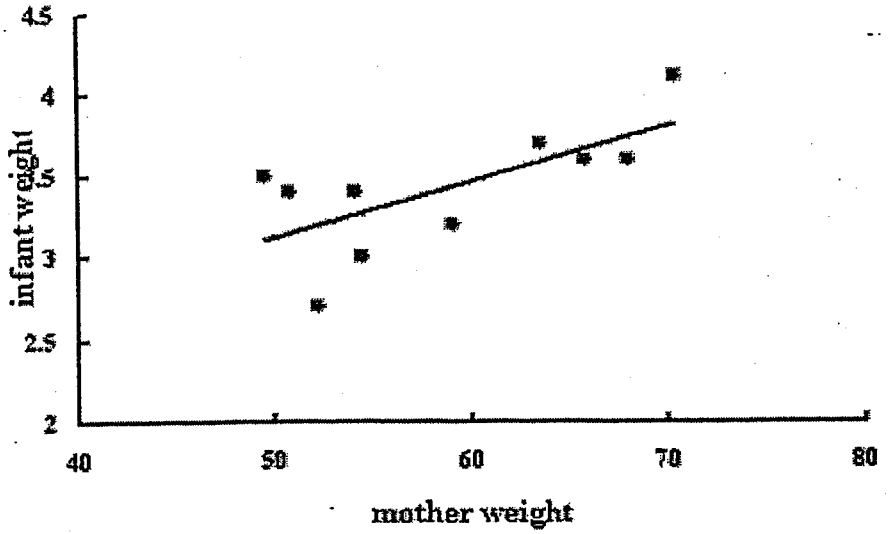
b = 0.037 is called ***regression coefficient***

Interpretation of regression coefficient:

For every unit increase in the independent variable x, we expect on average, the dependent variable (Y) to increase by the amount of b (regression coefficient).

For our example:

For every kg increase in the mother weight, we expect infant weight to increase by 0.037.



The regression line can be used for prediction:

What is the expected weight of an infant if the mother weight if mother weight is 60 kg. In this case, we substitute the value of mother weight (60kg) in the regression equation to obtain a predicted value for infant weight.

The **predicted** value is:

$$\hat{Y} = 1.22 + 0.037\ (60)$$
$$= 1.22 + 2.22 = 3.44$$

So, we expect infant weight for a mother with 60 kg of weight to be 3.44.

The below table shows the observed (Y) values and the predicted values ($\hat{Y}$)

Independent variable (X)	Observed value (Y)	Predicted value $\hat{Y}$	$d = Y - \hat{Y}$	$d^2 = (Y - \hat{Y})^2$
49.4	3.5	3.04	0.45	0.2025
63.5	3.7	3.57	0.13	0.0169
68.0	3.6	3.74	-0.34	0.1156
52.2	2.7	3.15	-0.45	0.2025
54.4	3.0	3.23	-0.23	0.0529
70.3	4.1	3.82	0.28	0.0784
50.8	3.4	3.10	0.30	0.0900
65.8	3.6	3.65	0.05	0.0025
54.1	3.4	3.22	0.18	0.0324
59.5	3.2	3.40	0.20	0.0400
				0.8337

The values in the third column are obtained from the equation

$$\hat{Y} = \mathbf{1.22 + 0.037\ X}$$

as follows:

$$\hat{Y}_1 = 1.22 + 0.037\ (49.2) = 1.22 + 1.82 = 3.04$$
$$\hat{Y}_2 = 1.22 + 0.037\ (63.5) = 1.22 + 2.35 = 3.57$$

$$\hat{Y}_{10} = 1.22 + 0.037\ (59) = 1.22 + 2.18 = 3.40$$

We need an indication of whether the points lie close or far from the line. The standard error (Se) provides a measure of the degree to which data are scattered about the regression line. If Se is large, then the observed values (Y) are widely scattered about the regression line, and we should place little confidence in prediction made from the equation. If Se is small, then we know that most of the points lie close to the line and we should have confidence in prediction made from the regression line.

The standard error is given by:

$$Se = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

An alternative simple formula is:

$$Se = \sqrt{\frac{\sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i}{n-2}}$$

For our example the standard error from the first equation is :

$$Se = \sqrt{0.8337 \,/\, (10-2)} = 0.323$$

While from the second equation is given by:

$$Se = \sqrt{\frac{34.2 - 1.22(34) - 0.037(2023.4}{10 - 2}} = 0.323$$

Example 2:

The table below shows plasma volume and body weight for eight healthy males.

Body weight (kg)	Plasma volume (l)
58	2.75
70	2.86
74	3.35
64	2.76
62	2.62
71	3.49
71	3.05
66	3.12
Total 536	24

① Find the regression equation

② find the determination Coefficient

③ the standard error

④ Find the plasma volume when the body weight equals 80

$$\sum_{i=1}^{8} X_i = 536, \quad \sum_{i=1}^{8} Y_i = 24, \quad \sum_{i=1}^{8} X_i Y_i = 1616.94,$$

$$\sum_{i=1}^{8} X_i^2 = 36118, \sum_{i=1}^{8} Y_i^2 = 648.66$$

$$\bar{X} = 536/8 = 67, \quad \bar{Y} = 24/8 = 3$$

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

substituting values from the information given, we obtain:

$$b = \frac{1616.94 - (8)(67)(3)}{36118 - (8)(67)^2} = 0.043$$

$$a = \bar{Y} - b\bar{X}$$

$$\mathbf{a = 3 - 0.043\ (67) = 0.12}$$

Thus, the relationship is described by:

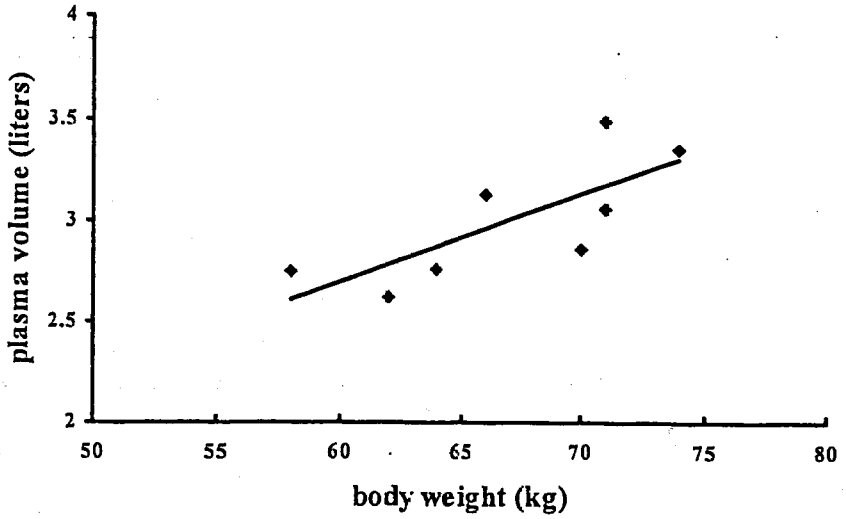
$$\hat{Y} = \mathbf{a + bX}$$

Plasma volume = 0.12 + 0.0437 body weight

The regression coefficient (b = 0.043) indicates: with each 1 kg increase in body weight, plasma volume will increase by 0.043 liters.

We can predict plasma volume from body weight: For example if body weight is 60kg, plasma volume will be:

Plasma volume = 0.12 + 0.043 (60) = 2.70 liter.



In Summary:

Main Goals in regression:

1- Establish a relationship between two or more variables

2- Enable us to predict the values of the dependent variable from the values of the independent variables.

EXERCISES

Q1- The following scores represent nurses' assessment (X) and physicians' assessment (Y) of the condition of 10 patients at time of admission to trauma center.

X	18	13	18	15	10	12	8	4	7	3
Y	23	20	18	16	14	11	10	7	6	4

1- Construct scatter diagram for this data

1) find person Coefficient
2) the relation is
3) the strenght is
4) the Coefficient of determination is
5) the Regression eqtn is
6) If X = 20 then the prediction value is

2- Plot the following regression equations on the scatter diagram and indicate which one you think best fits the data.

1- $\hat{Y} = 8 + 0.5X$

2- $\hat{Y} = -10 + 2X$

3- $\hat{Y} = 1.21 + 1.08X$

Q2- Height is frequently named as a good predictor for weight among people of the same age and gender. The following are the heights (m) and weights of 10(kg) males between the ages of 19 and 26 years.

Weight (X)	83.9	99	63.8	71.3	65.3	79.6	70.3	69.2	56.4	66.2
Height (Y)	1.85	1.8	1.73	1.68	1.83	1.84	1.74	1.64	1.69	2.05

Information given:

1- $\sum X_i = 725$ 2- $\sum Y_i = 17.85$ 3- $\sum X_i Y_i = 1296.37$ 4- $\sum X_i^2 =$ 53888.72

5- $\sum Y_i^2 = 31.99$

1- Draw scatter diagram

2- Obtain the regression equation

3- Predict weight if height is 1.85m

5- Calculate Pearson correlation coefficient r.

6- Test the hypothesis:

$H_0 : \rho = 0.$ VS $H_1: \rho > 0$

Q3- Suppose we are interested in the relationship between forced vital capacity (FVC), a pulmonary function index, and the age of a Saudi boy. For a sample of 10 Saudi boys aged 6, ...15 we measured the FVC for each boy. Results are reported in the below table.

Age (X)	6	7	8	9	10	11	12	13	14	15
FVC (Y)	1.21	1.38	1.54	1.78	2.00	1.94	2.26	2.34	2.80	3.05
Rank (X)	1	2	3	4	5	6	7	8	9	10
Rank (Y)										

1- $\sum X_i = 105$ **2-** $\sum Y_i = 20.3$ **3-** $\sum X_i Y_i = 229.09$ **4-** $\sum X_i^2 = 1185$

5- $\sum Y_i^2 = 44.40$

1- What is the dependent and independent variables?

2- Obtain the regression equation

3- Predict FVC if age is 10.5 years.

4- Calculate the Pearson correlation coefficient r.

5- Test the hypothesis:

$H_0 : \rho = 0.$ VS $H_1: \rho > 0$

7- Calculate Spearman rank correlation r_s

8- Test the hypothesis

$H_0 : \rho_s = 0$ $H_1: \rho_s \neq 0$

4- Test the hypothesis:
$H_0 : \rho = 0.$ VS $H_1: \rho > 0$

5- Calculate Spearman rank correlation

6- Test the hypothesis
$H_0: \rho_s = 0$ $H_1: \rho_s > 0$

Q7- If the relationship between X and Y is positive, as variable Y decreases variable X:
1- increases 2- decreases 3- remains the same 4- changes

Q8- Which of the following statements is false:
1- Pearson's r is used when one or both variable is at least of interval scaling
2- The range of the correlation coefficient is from -1 to +1
3- A correlation of r = -0.85 implies a stronger association than r = 0.80
4- A negative correlation between X and Y indicates as X increases Y increases

Q9- Of the following measurement levels, which is required for the valid calculation of a Pearson correlation coefficient?
1- nominal 2- ordinal 3- interval 4- ratio

Q10- You are told that there is a high, positive correlation between measures of 'fitness' and hrs of exercise. The correlation coefficient consistent with the above statement is:
1- 0.3 2- 0.2 3- 0.8 4- -0.3 5- non of these

Q11- The lowest strength of association is reflected by which of the following correlation coefficients?
a) 0.95 b) -0.60 c) -0.33 d) 0.29 e) non of a, b, c, d

Q12- Of the following measurement levels, which is required for the valid calculation of Spearman correlation coefficient?
1- nominal 2- ordinal 3- interval 4- ratio

Q13- An investigator aims to establish for a sample of subjects the relationship between blood cholesterol levels (mg/cc) and blood pressure (mmHg). Below two questions refer to this investigation.
1- The correlation coefficient appropriate for establishing the degree of correlation between the two variables:
a) is determined by the sample size b) depends on the direction of the relationship
c) is Spearman's r d) is Pearson's r

Q14- If the correlation coefficient obtained is 0.80. A correlation of this direction and magnitude indicates that:
a) High blood pressure causes high cholesterol
b) High blood cholesterol causes high blood pressure
c) there might be a third variable which causes both high blood pressure and high cholesterol.
d) non of the statements a, b, c is consistent with the value or r
e) any of the statements a, b, c might be correct, we cannot be sure from the available r value.

Q15- A study is conducted to examine the relationship between daily caffeine intake and systolic blood pressure.
1- The correlation coefficient appropriate for establishing the degree of correlation between the two variables:
a) is determined by the sample size
b) depends on the direction of the relationship
c) is Spearman's r
d) is Pearson's r

2- The calculated correlation coefficient measures which of the followings?

a) The extent to which caffeine intake and blood pressure are causally related
b) The degree of association between caffeine intake and blood pressure
c) The likelihood that caffeine intake and systolic blood pressure are mutually exclusive
d) The statistical significance of the association between daily caffeine intake and blood pressure.

Q16- We can calculate the correlation coefficient between two sets of data if given:

a) the top score from one set and the lowest score from the other set
b) at least two scores from the same set
c) The two sets of measurements for the same subjects
d) Not possible to calculate r for the two sets