# Correlation and Regression

Basra has a long history of seasonal dust storms, several researchers decided to see what effect, if any, these storms had on the respiratory health of the people living in the area. They undertook (among other things) to see if there was a relationship between the amount of dust and sand particles in the air when the storms occur and the number of hospital emergency room visits for respiratory disorders at three community hospital in Basra.

## Introduction

The relationship between two variables is an area of inferential statistics. For example, educators are interested in determining whether the number of hours a student studies is related to the student's score on a particular exam. Medical researchers are interested in questions such as, if caffeine related to heart damage? Or is there a relationship between a person's age and his or her blood pressure? A zoologist may want to know whether the birth weight of a certain animal is related to its life span. These are only a few of the many questions that can be answered by using the techniques of correlation and regression analysis.

**Correlation** is a statistical method used to determine whether a relationship between variables exists.

**Regression** is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

The purpose of this lesson is to answer these questions statistically:

1. Are two or more variables related?
2. If so, what is the strength of the relationship?
3. What type of relationship exists?
4. What kind of predictions can be made from the relationship?

## Scatter Plots

In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables. For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of students, determine the hours each studies, and obtain their grades on the exam. A table can be made for the data, as shown here.

| Student | Hours of study *x* | Grade y (%) |
|---|---|---|
| A | 6 | 82 |
| B | 2 | 63 |
| C | 1 | 57 |
| D | 5 | 88 |
| E | 2 | 68 |
| F | 3 | 75 |

The two variables x and y are called independent and dependent variables, respectively. The independent and dependent variables can be plotted on a graph called a scatter plot. The independent variable x is plotted on the dependent variable y is plotted on the vertical axis.

The procedure for drawing a scatter plot is shown in Example 1 through 3.

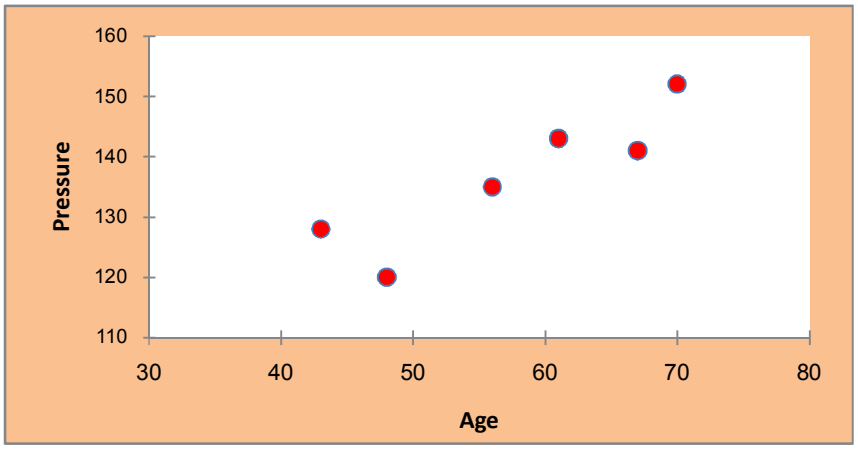
**Example 1**

Construct a scatter plot for the data obtained in a study of age and systolic blood pressure of six randomly selected subjects. The data are shown in the table.

| Subject | Age *x* | Pressure y |
|---|---|---|
| A | 43 | 128 |
| B | 48 | 120 |
| C | 56 | 135 |
| D | 61 | 143 |
| E | 67 | 141 |
| F | 70 | 152 |

**Solution:**

1. Draw and the x and y axes.
2. Plot each point on the graph, as shown in figure below.



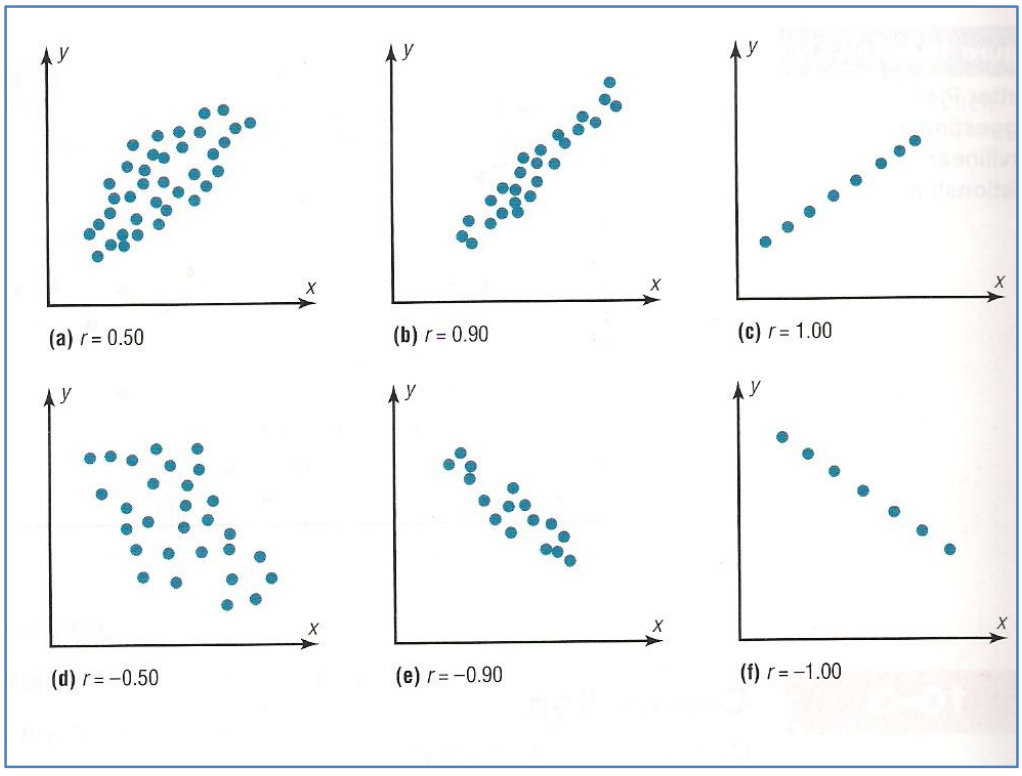
## Correlation

### *Correlation Coefficient*

Statisticians use a measure called the correlation coefficient to determine the strength of the relationship between two variables. There are several types of correlation coefficients. The one explained in this section is called the Pearson Product moment correlation coefficient (PPMC), named after statistician Karl Pearson, who pioneered the research in this area.

The range of the correlation coefficient is from -1 to +1. If there is a strong positive linear relationship between the variables, the value of r will be close to +1. If there is a strong negative linear relationship between the variables, the value of r will be close to -1. When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.

The graphs in Fig. below shows the relationship between the correlation coefficients corresponding scatter plots.



There are several ways to compute the value of the correlation coefficient. One method is to use the formula shown here:

**Formula for the correlation coefficient r**

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x^2)][n(\sum y^2) - (\sum y^2)]}}$$

where $n$ is the number of data points

---

**Example 2**

Compute the value of the correlation coefficient for the data obtained in the study of age and blood pressure given in Example 2.

Solution:

**Step 1:** Make a table as shown here

| *Subject* | *Age x* | *Pressure y* | *xy* | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| A | 43 | 128 | | | |
| B | 48 | 120 | | | |
| C | 56 | 135 | | | |
| D | 61 | 143 | | | |
| E | 67 | 141 | | | |
| F | 70 | 152 | | | |

**Step 2:** Find the values of xy, $x^2$, $y^2$ and place these values in the corresponding column of the table.

The completed table is shown

| *Subject* | *Age x* | *Pressure y* | *xy* | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| A | 43 | 128 | 5504 | 1849 | 16384 |
| B | 48 | 120 | 5760 | 2304 | 14400 |
| C | 56 | 135 | 7560 | 3136 | 18225 |
| D | 61 | 143 | 8723 | 3721 | 20449 |
| E | 67 | 141 | 9447 | 4489 | 19881 |
| F | 70 | 152 | 10640 | 4900 | 23104 |
| **Sum** | 345 | 819 | 47634 | 20399 | 112443 |

**Step 3:** substitute in the formula and solve for r.

$$r = \frac{6(47634) - 345 * 819}{\sqrt{[6(20399) - (345)^2][6(112443) - (819)^2]}} = 0.897$$

The correlation coefficient suggests a strong positive relationship between age and blood pressure.

**Example 3**

Compute the value of the correlation coefficient for the data obtained in the study of the number of absences and the final grade of the seven students in the statistics class given in table below

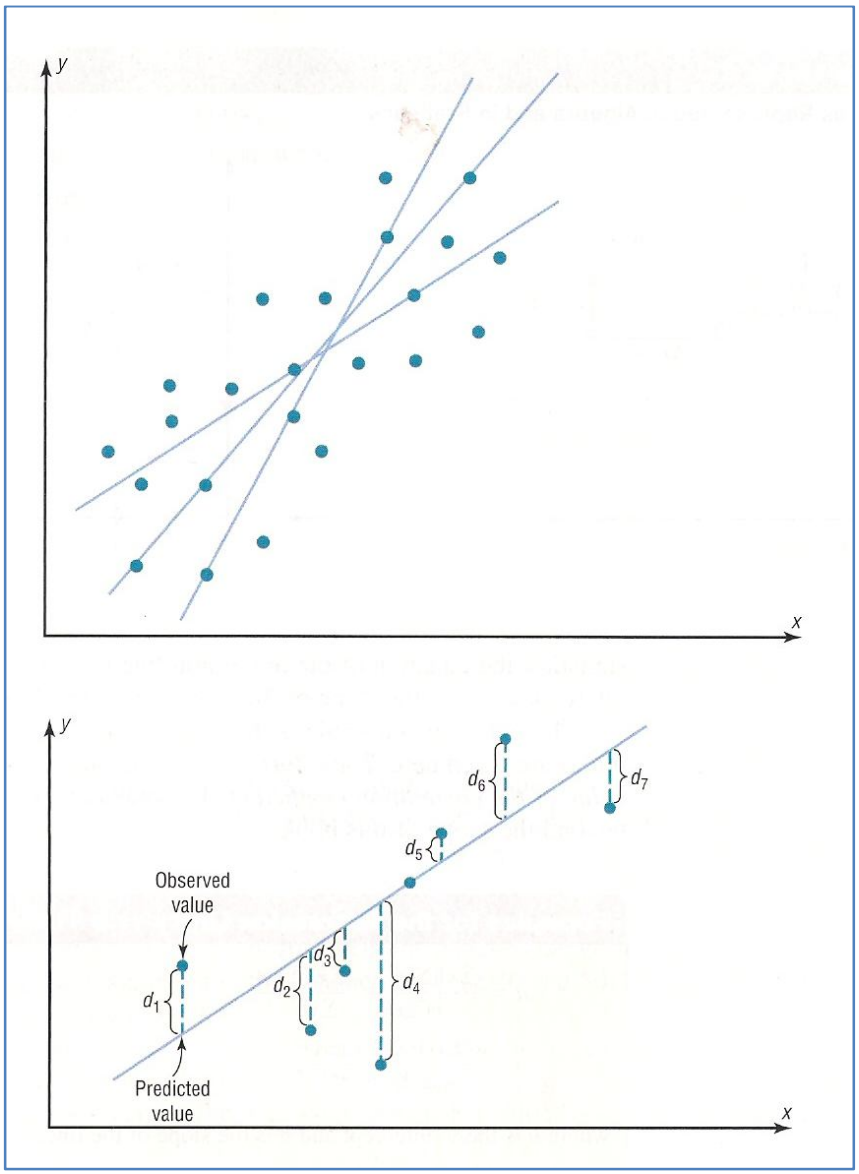
| *student* | *Number of absences x* | *Final grade y (%)* | *xy* | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| A | 6 | 82 | | | |
| B | 2 | 86 | | | |
| C | 15 | 43 | | | |
| D | 9 | 74 | | | |
| E | 12 | 58 | | | |
| F | 5 | 90 | | | |
| G | 8 | 78 | | | |

## Regression

In studying relationships between two variables, collect the data and then construct a scatter plot. The purpose of the scatter plot, as indicated previously, is to determine the nature of the relationship. The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or no discernible relationship. After the scatter plot is drawn, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship. If the value of the correlation coefficient is significant, the next step is to determine the equation of the regression line, which is the data's line of best fit. The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

### Line of Best Fit

Figure below shows a scatter plot for the data of two variables. It shows that several lines can be drawn on the graph near the points. Given a scatter plot, one must be able to draw the *line of best fit*. Best fit means that the sum of the squats of the vertical distances from each point to the line is at a minimum. The reason one needs a line of best fit is that the values of $y$ will predicted from the values of $x$; hence, the closer the points are to the line, the better the fit and the prediction will be.




**Determination of the regression line equation**

In algebra, the equation of a line is usually given as $y = mx + b$, where m is the slope of the line and b is the y intercept. In statistics, the equation of the regression line is written as $\acute{y} = a + bx$, where a is the $\acute{y}$ intercept and b is the slope of the line.

There are several methods for finding the equation of the regression line. Two formulas are given here. These formulas use the same values that are used in computing the value of the correlation coefficient.

Formula for the regression line $\hat{y} = a + bx$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

where $a$ is the $\hat{y}$ intercept and b is the slope of the line.

---

**Example 4**

Find the equation of the regression line for the data in Example below and graph the line on the scatter plot of the data.

| *Subject* | *Age x* | *Pressure y* | *xy* | $x^2$ |
|---|---|---|---|---|
| A | 43 | 128 | | |
| B | 48 | 120 | | |
| C | 56 | 135 | | |
| D | 61 | 143 | | |
| E | 67 | 141 | | |
| F | 70 | 152 | | |
| Sum | | | | |

**Solution:** make a table like below

| *Subject* | *Age x* | *Pressure y* | *xy* | $x^2$ |
|---|---|---|---|---|
| A | 43 | 128 | 5504 | 1849 |
| B | 48 | 120 | 5760 | 2304 |
| C | 56 | 135 | 7560 | 3136 |
| D | 61 | 143 | 8723 | 3721 |
| E | 67 | 141 | 9447 | 4489 |
| F | 70 | 152 | 10640 | 4900 |
| Sum | 345 | 819 | 47634 | 20399 |

Substituting in the formula, one gets

a = 81.048
b = 0.964
Hence, the equation of the regression line $\hat{y} = 81.048 + 0.964x$

The graph of the line is shown in Fig. below