CORRELATION ANALYSIS

Correlation is another way of assessing the relationship between variables. To be more precise, it measures the extent of correspondence between the ordering of two random variables. There is a large amount of resemblance between regression and correlation but for their methods of interpretation of the relationship. For example, a *scatter diagram* is of tremendous help when trying to describe the type of relationship existing between two variables.

1 Measuring correlation

We make use of the *linear product-moment correlation coefficient*, also known as *Pearson's correlation coefficient*, to express the strength of the relationship. This coefficient is generally used when variables are of *quantitative* nature, that is, ratio or interval scale variables.

Pearson's correlation coefficient is denoted by *r* and is defined by

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

The value of *r* always lies between –1 and 1 inclusive, that is, $-1 \le r \le 1$. If *Y* increases when *X* increases, we say that there is *positive* or *direct* correlation between them. However, if *Y* decreases when *X* increases (or *vice versa*), then we say that they are *negatively* or *inversely* correlated. The reader must have noticed that *direct* and *inverse* are terms that are used in the context of variation or *proportionality*.

2 Interpretation of the correlation coefficient

The extreme values of *r*, that is, when *r* = ±1, indicate that there is *perfect* (positive or negative) correlation between *X* and *Y*. However, if *r* is 0, we say that there is *no* or *zero* correlation.

Note

When *r* = 0, we may *not* assert that there is no correlation *at all* between *X* and *Y*. Pearson's correlation coefficient is meant to measure *linear* relationship only. It should *not* be used in the case of *non-linear* relationships since it will obviously lead to an erroneous interpretation.

The remaining values, falling in subintervals of [–1, 1], describe the relationship in terms of its *strength*. **Fig. 2.1** below may be used as a guideline as to what *adjective* should be used for the values of *r* obtained after calculation to describe the relationship.

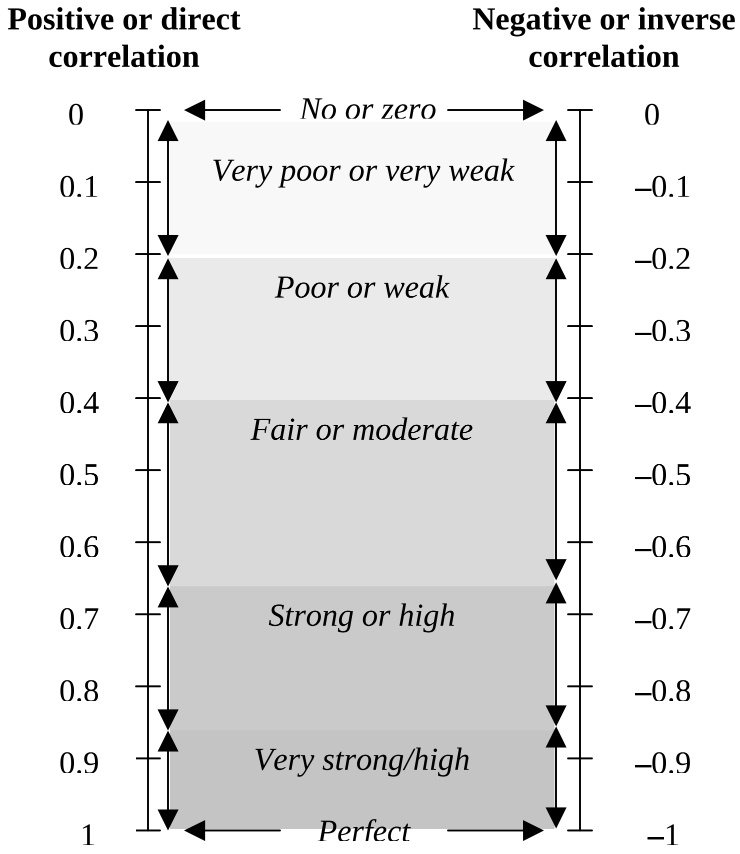


Fig. 2.1 Interpretation of correlation coefficient

Note that **Fig 2.1** is only to be used as a *guideline*. There are no set values that demarcate, for example, *moderate* from *strong* correlation.

We observe that the strength of the relationship between *X* and *Y* is the same whether $r = 0.85$ or -0.85. The only difference is that the there is *direct* correlation in the first case and *inverse* correlation in the second. We should bear in mind that *r* is the *linear* correlation coefficient and that, as mentioned earlier, its value can be wrongly interpreted whenever the relationship between *X* and *Y* is non-linear. That is the reason why we should have a look at a scatter diagram of points (x, y) and verify whether the relationship is, for example, of quadratic, logarithmic, exponential or trigonometric (briefly, *non-linear*) nature.

If r = 0, we should not jump to the conclusion that there is no correlation at all between X and Y. Consider the case where there is *perfect* (but unsuspected) *non-linear* correlation between the two variables, say, related by the equation $Y = X^2$ (see **Fig. 2.2** below). Taking an initial set of points (–3, 9), (–2, 4), (–1, 1), (0, 0), (1, 1), (2, 4) and (3, 9), then the reader may easily verify that both $\sum x$ and $\sum xy$ are equal to zero. Consequently, r = 0 (check the formula for r in Section 9.1). We deduce that the linear product-moment correlation coefficient cannot be used to interpret the strength of a non-linear relationship.

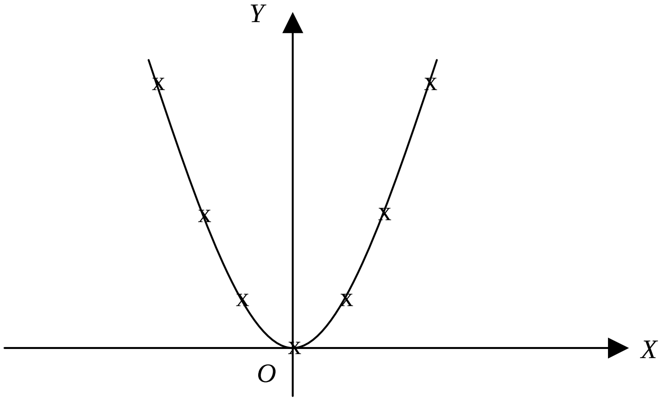
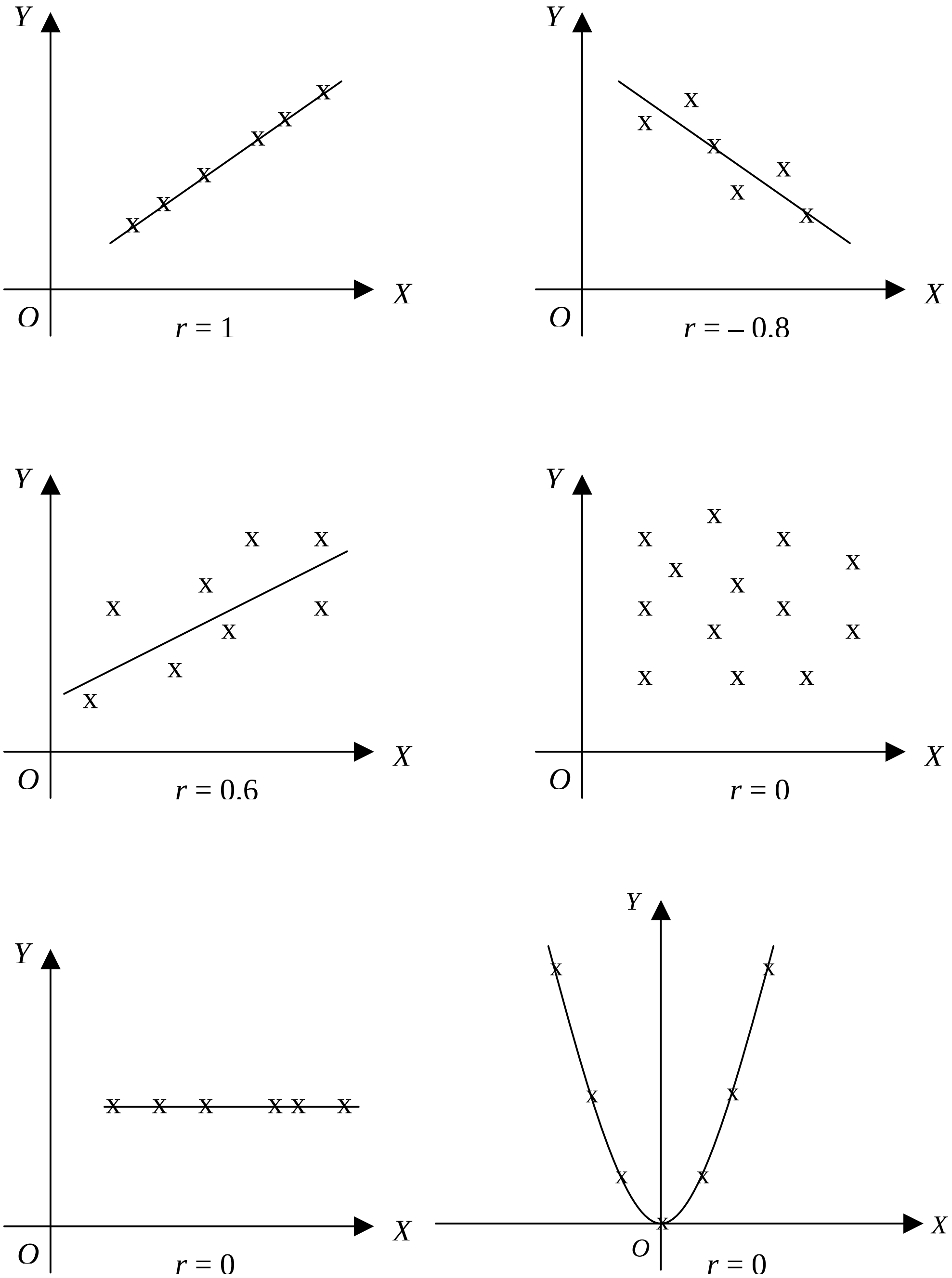


Fig. 2.2 Perfect non-linear relationship

With practice and experience, it is even possible to know approximately the value of r by inspection of a scatter diagram. The location (amount of scattering) of the points with respect to the least-squares regression line indicates the strength of the relationship between the variables. The *more scattered* the points are, the *weaker* is the relationship and the *closer* is the value of r to zero.

The sign of r is always the same as that of (the gradient) b in the regression equation $\hat{Y} = a + bX$. **Fig. 2.3** below shows how we can deduce the value of r to a certain degree of accuracy from a scatter diagram.

Note If the variables were *qualitative* in nature, that is, *nominal* or *ordinal*, then it would be advisable to use a *non-parametric* method of determining the correlation coefficient, namely, *Spearman*'s (not included in this course).



$r = 1$

$r = – 0.8$

$r = 0.6$

$r = 0$

$r = 0$

$r = 0$

Y is independent of X, that is, Y assumes the same value irrespective of X.

X and Y have a non-linear relationship.

Fig. 2.3 Using scattering diagrams to determine *r* approximately

Example

The yield of a particular crop on a farm is thought to depend principally on the amount of rainfall in the growing season. The values of the yield *Y*, in tonnes per acre, and the rainfall *X*, in centimetres, for seven successive years are given in the table below.

Rainfall (cm)	12.3	13.7	14.5	11.2	13.2	14.1	12.0
Yield (tonnes per acre)	6.25	8.02	8.42	5.27	7.21	8.71	5.68

Calculate the linear correlation coefficient and interpret your result.

Solution

We first summarise the data from the above table as follows:

$$\sum x = 91 \quad \sum x^2 = 1191.72 \quad \sum xy = 654.006 \quad \sum y = 49.56 \quad \sum y^2 = 362.1628$$

Pearson's correlation coefficient is calculated as

$$r = \frac{(7)(654.006) - (91)(49.56)}{\sqrt{\left[(7)(1191.72) - (91)^2\right]\left[(7)(362.1628) - (49.56)^2\right]}} = 0.9807$$

Hence, there is a *very strong direct* correlation between rainfall and yield. The relationship between these variables is most probably *linear*.

3 Causality

Causality, also known as *causation*, is defined as a *cause-effect* relationship between two variables. A significant correlation does not necessarily indicate causality but rather a *common linkage* in a sequence of events. One type of significant correlation situation is when both variables are influenced by a common cause and therefore are correlated with each other.

For example, individuals with a higher level of income have both higher levels of savings and spending. We might find that there is a positive correlation between level of savings and level of spending but this does not mean that one variable *causes* the other. We should mention the very interesting case where two related variables are separated by several steps in a *cause-effect chain* of events. **Fig. 3.1** illustrates this example.

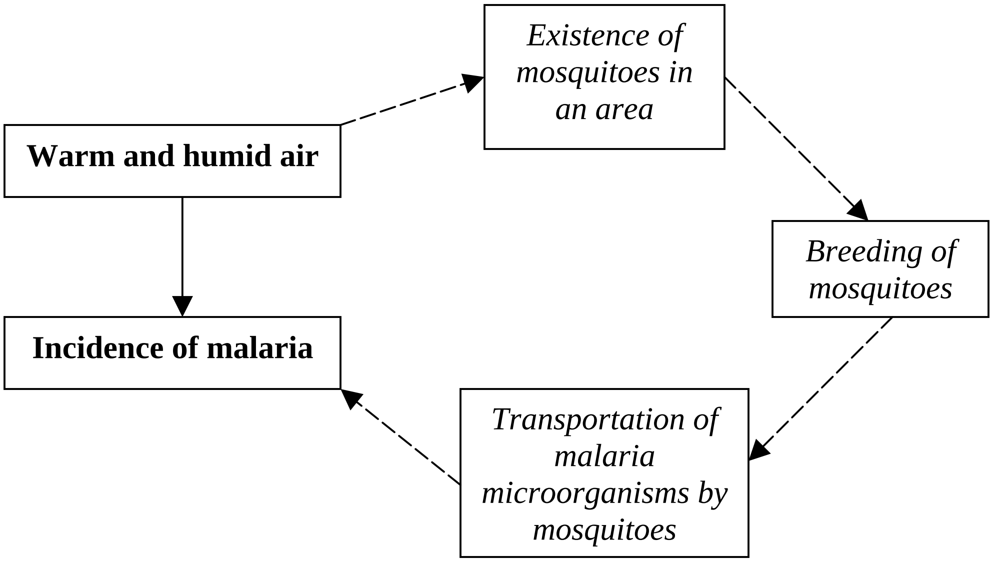

Existence of mosquitoes in an area
Warm and humid air
Breeding of mosquitoes
Incidence of malaria
Transportation of malaria microorganisms by mosquitoes

Fig. 3.1 Correlation does not imply causality

Thus, the existence of warm and humid climatic conditions is not itself the cause of malaria.

4 Spurious correlation

Spurious correlation occurs between two variables that are supposed to be mutually independent. It must be conceded that the correlation coefficient can be readily calculated for any given set of paired data, the names of the variables being totally irrelevant.

We may think about the variables *X* and *Y* being respectively given by

X = "number of bags of potatoes sold daily at the Quatre-Bornes market"
Y = "number of accidents daily in the United States of America"

With a given set of data, we may well find that the correlation between *X* and *Y* is highly significant (for example, 0.89). But does that mean that these two variables are strongly correlated? Certainly not! Not even through the longest and most complex cause-effect chain. That is what spurious correlation is all about.

5 Coefficient of determination

The *coefficient of determination* is much more useful than the correlation coefficient in the sense that it gives a more plausible statistical explanation of the relationship between two variables X and Y. It is denoted by r^2 and is simply the square of the correlation coefficient. While the correlation coefficient only describes the strength of the relationship in terms of a carefully chosen adjective, the coefficient of determination gives the variability in Y explained by the variability in X.

For instance, although $r = 0.8$ (high positive correlation), r^2 is only 0.64. We say that the variability in X account for 64% of the variability in Y. In terms of regression, it simply means that, apart from the predictor X, there are other factors which also influence the response variable Y (the remaining 36%) but which have not been considered in the analysis.

The following example should help the reader to have a better understanding of this concept.

Example

An investigation was carried out in order to find the correlation between the marks obtained by students in July mock exams and November final exams. The main objective of this investigation was to determine the *degree of predictability* of mock exam marks. A Pearson's correlation coefficient of 0.72 was obtained and it was concluded that the two sets of marks were significantly correlated. The result meant that mock exam marks could be used to predict final exam marks *to a reasonable extent*.

However, the coefficient of determination, 0.52, revealed that mock exam marks explained only 52% of the variability in final exam marks. Thus, a *simple regression model* was inadequate since it did not include enough *predictor variables*, that is, 'mock exam marks' is not enough to predict 'final exam marks' accurately.

The model could be improved by adding a few more significant predictors like

1. Attendance rate
2. Number of hours of study
3. Number of hours of sleep
4. Number of hours of leisure

The adequacy of this new model (multiple linear regression) can be checked by running it in SPSS (or any other statistical software). If the abovementioned predictors do influence (positively or negatively) 'final exam marks' to the slightest extent, the value of the coefficient of determination will increase.