Regression Lines

Prerequisites

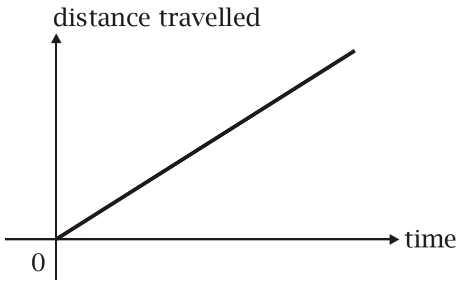
You should be familiar with direct proportionality and the equation of the straight-line.

Relationships between variables

A variable represents a physical quantity that can take more than one value. For example, time and distance are variables.

Example (1)
A man is running along a straight track away from a fixed point *O* at a constant speed. Sketch a graph of the distance he has travelled (vertical axis) against the time he has been running (horizontal axis).

Solution
This is a straight-line graph through the origin. The gradient of this graph is the speed at which the man is running.



In example (1) there is a relationship of *direct proportionality* between two variables – time (horizontal axis) and distance travelled (vertical axis). The relationship is represented graphically by a straight-line through the origin. Letting *s* stand for distance and *t* for time, the equation becomes

$s = kt$

where k is the constant of proportionality, here the speed the man is running. Any graph involving a straight-line represents a *linear relationship*. In direct proportionality the graph of the relationship between the two variables is a straight-line through the origin. There may be relationships between variables that are linear but not directly proportional. In such cases the graph is a straight-line but does not pass through the origin. It has the general equation of a straight-line, that is

$y = mx + c$

where m is the gradient and c is the intercept.

Independent and dependent variables

Generally, in science we think of one variable as changing as a consequence of some change in the other variable. We also conduct experiments in order to discover such relationships. The variable that changes as a consequence of changes in the other variable is called the *dependent* variable. The variable that does not change, but rather causes the change, is called the *independent* variable. For example, the volume of a gas (dependent variable) is related to its temperature (independent variable).

Example (2)

It is known that there is a linear relationship between the volume of a gas V (litres) and its temperature T (degrees Celsius). (The pressure of the gas being held constant.)

$V = \alpha + \beta T$

where α and β are constants. In an experiment the volume of a gas was measured at two different temperatures. At $T = 25$ °C the volume was $V = 24.775$ litres. At $T = 100$ °C the volume was $V = 31.0$ litres. Determine α and β. Find to the nearest degree the value of T for which the volume V is zero.[1] Sketch the graph of V against T.

Solution

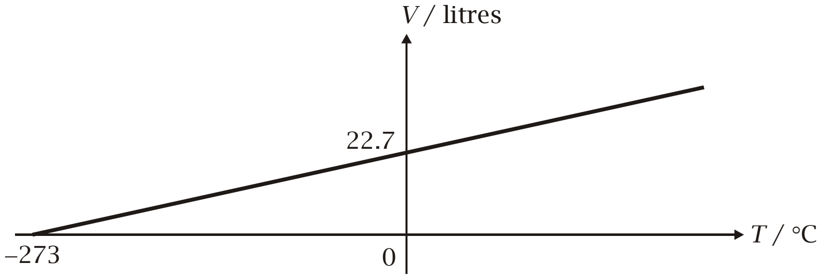
Substituting twice into $V = \alpha + \beta T$ we obtain

$$24.775 = \alpha + 25\beta \qquad (1)$$
$$31.0 = \alpha + 100\beta \qquad (2)$$

Solving these simultaneously

[1] This is a theoretical value. Before the volume becomes zero the gas at low temperatures condenses to a liquid and then to a solid.

$$24.775 = \alpha + 25\beta \qquad (1)$$
$$31.0 = \alpha + 100\beta \qquad (2)$$
$$75\beta = 6.225 \qquad \beta = 0.083 \qquad \alpha = 22.7$$
$$V = 0 \quad \Rightarrow \quad T = -\frac{22.7}{0.083} = -273 \text{ °C (nearest °C)}$$

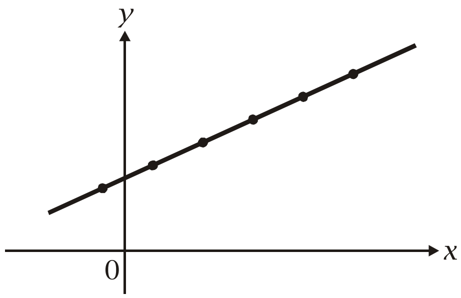


Correlation

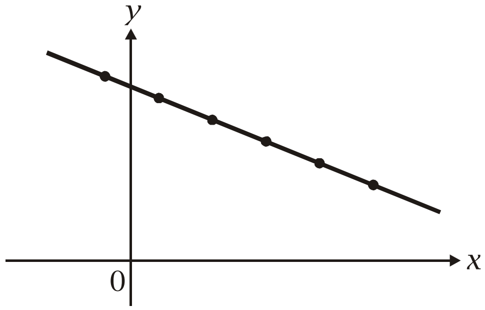
Correlation is concerned with the relation between two variables, say x and y. If y goes up as x goes up, then we say that x and y exhibit positive correlation. If y goes down as x goes up, then x and y exhibit negative correlation; there is an inverse relation between them. The question of correlation usually arises in the context of data obtained by observation, or more precisely, by experiment. The question is, does the physical quantity represented by one variable x, correlate with the physical quantity represented by the other variable y? An example might be an investigation of the relationship between micrograms of nicotine in the blood stream (x) and blood pressure (y). In an experiment we collect data. The data comprise pairs of values. Each pair of values is called a *data point*. If n data points are collected let (x_i, y_i) represent the ith data point. It is often convenient to tabulate data.

x	x_1	x_2	...	x_i	...	x_n
y	y_1	y_2	...	y_i	...	y_n

A perfect correlation between two variables x and y exists when the data points can be fitted **exactly** to a straight-line.

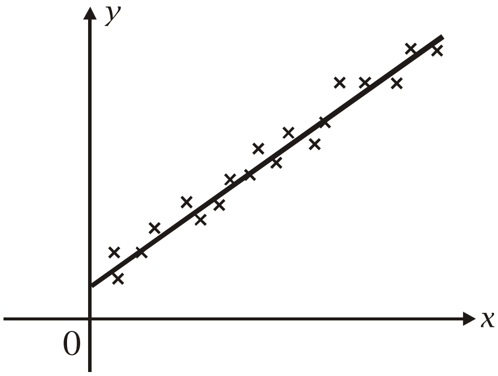


Perfect positive correlation



Perfect negative correlation

Recognising perfect correlation does not present a problem. But in practice data points do not fit exact lines. However, for several reasons the lack of exact fit is not taken automatically to entail the absence of a correlation; in fact, variation from exact correlation is to be expected in observed and experimentally determined data. This is because, firstly, observed physical quantities are often subject to the influence of extraneous variables that are beyond the control of the observer. For example, whilst it is now accepted that smoking causes cancer, most people know at least one person who has smoked all his life and is still alive at some ripe old age. This could possibly be attributed to some genetic factor giving that individual special resistance to the effects of smoking. From the point-of-view of correlation, we only expect that in the majority of cases incidence of lung cancer will be associated with increased levels of smoking. We do not expect an exact correlation between the average number of cigarettes smoked each day over a certain time period and the subsequent incidence of lung cancer. Secondly, experimental situations, whilst presenting better opportunities to the experimenter for controlling the influence of other variables, are still subject to random error. Each time the experimenter performs a routine, such as starting a stopwatch, there is some slight variation in the procedure, which shows up in the data as random fluctuations from an exact relation. Hence, deviations from exact correlation are expected. A graphical representation of the data may help to visualise the problem. Such graphs are called scatter diagrams.



The data points do not exactly fit a straight line, but appear to be very close to one.
However, arguing by appearances is a loose way of proceeding, and does not give a measure of the degree of correlation, or of the likelihood of the observed values arising by chance. What are required are precise mathematical determinations of the degree of correlation and the probability that the relationship could arise by chance. In addition, when data has been collected and it is believed that a correlation exists between two variables, x and y, what will be required is a "line of best fit" – this is the line that will fit the observed data as best as may be. Such a line is called *a regression line*. The question of the degree of correlation is taken up in another chapter. In this chapter we are concerned with determining regression lines fitted to sample data.

Regression lines

We are assuming that there is a linear relationship between two variables x and y such that

$y = \alpha + \beta x$ α, β constants

We do not in fact know α and β and seek estimates from them on the basis of sample data. The sample data comprise a set of data points giving for each value of the variable x an observed value of the variable y.

x	x_1	x_2	...	x_i	...	x_n
y	y_1	y_2	...	y_i	...	y_n

The regression line is a line that is the "best fit" to these data points. This is the line that gets "closest to the mostest" – that is, ensures that the total sum of the differences between the point on the line and the corresponding data point is a little as possible. A line fitted to a set of data points that is based on this idea is called the *least squares regression line*. Here we shall not give a precise description of this idea, which is reserved to another chapter.[2] From the set of data points (x_i, y_i) we can calculate the mean $(\bar{x}, \bar{y})$.

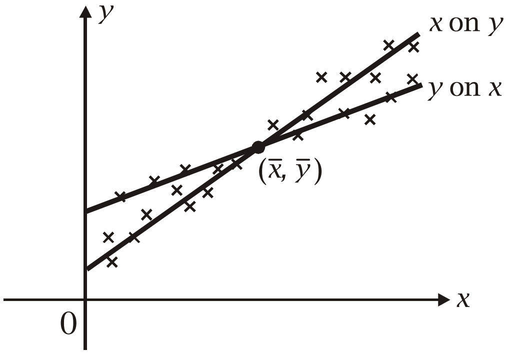
$$\bar{x} = \frac{\sum x}{n} \qquad \bar{y} = \frac{\sum y}{n}$$

There are two ways a regression line can be fitted to a given set of data. In both cases it is assumed that the regression line passes through the mean $(\bar{x}, \bar{y})$ of the *x* and *y* values.

(1) The regression line of *y* on *x*.
We assume the values of *x* are accurate and determine the regression line that gets the measured values of the variable *y* as close to the line as possible.

(2) The regression line of *x* on *y*
We assume the values of *y* are accurate and draw the regression line that gets the measured values of the variable *x* as close to the line as possible.

The following diagram illustrates how the two lines might be different.



In this chapter we shall primarily assume that the *x* values are accurate and that the *y* values are not. This makes the *y* values subject to variance and the *x* values not. We say that the *x* values are **exact**. The variable *y* shall be a function of the variable *x* and we assume a relationship between *x a*nd *y s*uch that

$y = \alpha + \beta x \qquad \alpha, \beta$ constants

However, since the *y*-values are assumed to be subject to variance, the **true** values of the coefficients α and β cannot be determined. What we find are **estimates** of these values. These shall be constants *a* and *b* such that

$y = a + bx$

[2] See the chapter at blacksacademy.net entitled *Proof of the Least Squares Regression Formula*

The values a and b are called the *least squares estimates* of α and β respectively. We now proceed to describe how we may obtain those estimates.

The regression line of y on x

The regression line of y on x, which finds y as a function of x, is given by

$$y - \bar{y} = b\left(x - \bar{x}\right) \quad \text{where} \quad b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

This line passes through the point $\left(\bar{x}, \bar{y}\right)$. From the equation $y - \bar{y} = m\left(x - \bar{x}\right)$ we may derive the constant a in the equation $y = a + bx$. This is the *least squares estimate* of the true relationship between x and y denoted by $y = \alpha + \beta x$. Also a and b are the least squares estimates of α and β respectively. In the above formula the expression $\sum xy$ stands for the sum of the products of the x and y values.

Example (3)

An investigation between two variables x and y produced the following data.

x	0	1	2	3	4	5	6	7	8
y	0	2.8	6.7	10.2	12.8	16.2	19.6	21.5	22.9

(a) Evaluate $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$

(b) Assuming a linear relationship $y = \alpha + \beta x$, calculate a and b, the least squares estimates of α and β.

Solution

(a) $\sum x = 36,\ \sum y = 112.7,\ \sum xy = 630.3$ and $\sum x^2 = 204$

Note that

$$\begin{aligned}\sum xy &= \left(0 \times 0\right) + \left(1 \times 2.8\right) + \left(2 \times 6.7\right) + \left(3 \times 10.2\right) + \left(4 \times 12.8\right) + \left(5 \times 16.2\right) \\ &\qquad + \left(6 \times 19.6\right) + \left(7 \times 21.5\right) + \left(8 \times 22.9\right) \\ &= 630.3\end{aligned}$$

(b)

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad n = 9$$

$$= \frac{9 \times 630.3 - 36 \times 112.7}{9 \times 204 - \left(36\right)^2}$$

$$= 2.9916...$$

$$\bar{x} = \frac{\sum x}{n} = \frac{36}{9} = 4.0$$

$$\bar{y} = \frac{\sum y}{n} = \frac{112.7}{9} = 12.5222...$$

The equation of the regression line is

$$y-\bar{y}=m\left(x-\bar{x}\right)$$
$$y-12.522...=2.9916...\left(x-4.0\right)$$
$$y-12.522...=2.9916...\times x-11.96$$
$$y=2.992x+0.556 \quad (3\text{ d.p.})$$

So $a=0.556$ and $b=2.991$

Covariance

The above formula for finding the regression line is based on the definition of *covariance*. For n pairs of data $\left(x_1,y_1\right),\left(x_{2,}y_2\right),\ ...\ ,\left(x_n y_n\right)$ …… the covariance, S_{xy} is

$$S_{xy}=\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$$

The covariance $S_x^2=S_{xx}$ is given by

$$S_x^2=S_{xx}=\sum\left(x-\bar{x}\right)\left(x-\bar{x}\right)=\sum\left(x-\bar{x}\right)^2$$

Likewise

$$S_{y^2}=S_{yy}=\sum\left(y-\bar{y}\right)\left(y-\bar{y}\right)=\sum(y-\bar{y})^2$$

We can show that

$$S_x^2=\sum x^2-\frac{\left(\sum x\right)^2}{n} \qquad S_y^2=\sum y^2-\frac{\left(\sum y\right)^2}{n}$$

Similarly, we can show

$$S_{xy}=\sum xy-\frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

Then the least squares estimate of the regression line y on x is

$$y=a+bx \ \text{ where } \ b=\frac{S_{xy}}{S_x^{\ 2}} \ \text{ and } a=\bar{y}-b\bar{x}$$

Here b is called the *coefficient of regression* of y on x.

Example (4)

In a chemical experiment the concentration of an acid x $\left(\text{mol dm}^{-3}\right)$ is related to the rate of reaction y $\left(\text{cm}^3\ \text{s}^{-1}\right)$. The following data were obtained.

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y	0	5.8	11.4	19.5	23.5	32.4	37.0	41.6	46.1	55.9	61.8

(*a*) Evaluate S_{xx} and S_{xy}.

(*b*) Assuming a linear relationship $y = \alpha + \beta x$, calculate *a* and *b*, the least squares estimates of α and β.

Solution

x_i	y_i	$(x_i)^2$	$x_i y_i$
0	0	0	0
0.1	5.8	0.01	0.58
0.2	11.4	0.04	2.28
0.3	19.5	0.09	5.85
0.4	23.5	0.16	9.40
0.5	32.4	0.25	16.20
0.6	37.0	0.36	22.20
0.7	41.6	0.49	29.12
0.8	46.1	0.64	36.88
0.9	55.9	0.81	50.31
1.0	61.8	1.00	61.80
$\sum x_i = 5.5$	$\sum y_i = 335$	$\sum\left(x_i^2\right) = 3.85$	$\sum x_i y_i = 234.62$

$$S_{xy} = \sum\left(x_i y_i\right) - \frac{\sum x_i \sum y_i}{n} = 234.62 - \frac{5.5 \times 335}{11} = 67.12$$

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 3.85 - \frac{(5.5)^2}{11} = 1.1$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{67.12}{1.1} = 61.02$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5.5}{11} = 0.5$$
$$\bar{y} = \frac{\sum y_i}{n} = \frac{335}{11} = 30.4545$$
$$a = \bar{y} - b\bar{x}$$
$$a = 30.4545 - 61.02 \times 0.5 = -0.0555$$
$$y = -0.0555 + 61.02x \;\; (4 \text{ s.f.})$$

Appendix

The regression line of *x* on *y*

The regression line of *x* on *y*, which finds *x* as a function of *y* where the *y* values are exact, is given by

$$x - \bar{x} = d(y - \bar{y}) \quad \text{where} \quad d = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2}$$

This line passes through the point $(\bar{x}, \bar{y})$. From the equation $x - \bar{x} = d(y - \bar{y})$ we may derive the constant *c* in the equation

$x = c + dy$

This is the *least squares estimate* of the true relationship between *x* and *y* denoted by

$x = \gamma + \delta y$

Also *c* and *d* are the least squares estimates of γ and δ respectively.

Covariant form

The regression line of *x* on *y*, which finds *x* as a function of *y* where the *y* values are exact, is given by

$x = c + dy$

Where $d = \frac{S_{xy}}{S_y{}^2}$ is the coefficient of regression of *x* on *y* and

$$S_y^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \qquad S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}.$$

The least squares estimate of the true relationship between x and y can be found from substitution into $x - \overline{x} = d\left(y - \overline{y}\right)$.