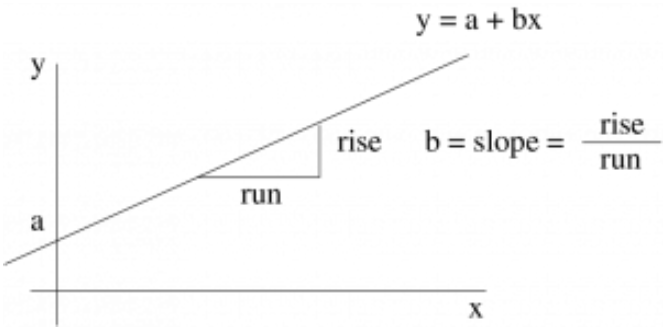
14.4: Linear Regression

Linear regression gives us the best equation of a line through the scatter plot data in terms of *least squares*. Let's begin with the equation of a line:

$$y = a + bx \tag{14.4.1}$$

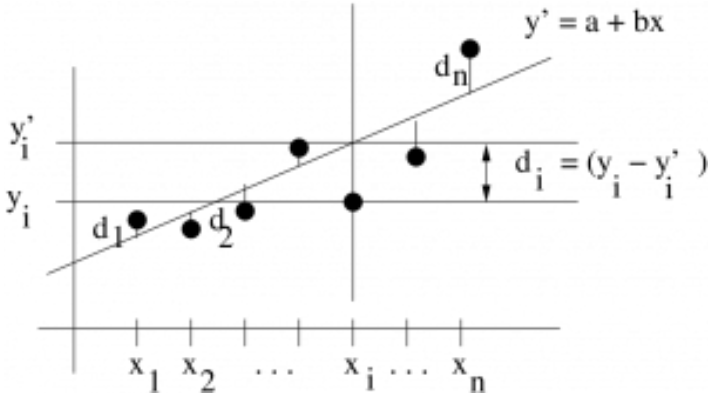
where a is the intercept and b is the slope.



The data, the collection of (x, y) points, rarely lie on a perfect straight line in a scatter plot. So we write

$$y' = a + bx \tag{14.4.2}$$

as the equation of the best fit line. The quantity y' is the predicted value of y (predicted from the value of x) and y is the measured value of y. Now consider :



The difference between the measured and predicted value at data point i, $d_i = y_i - y_i'$, is the *deviation*. The quantity

$$d_i^2 = (y_i - y_i')^2 = (y_i - (a + bx_i))^2 \tag{14.4.3}$$

is the *squared deviation*. The *sum of the squared deviations* is

$$E = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \tag{14.4.4}$$

The least squares solution for a and b is the solution that minimizes E, the sum of squares, over all possible selections of a and b. Minimization problems are easily handled with differential calculus by solving the differential equations:

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 \tag{14.4.5}$$

The solution to those two differential equations is

$$a = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{14.4.6}$$

and

$$b = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{14.4.7}$$

Example 14.3 : Continue with the data from Example 14.1 and find the best fit line. The data again are:

Subject	x	y	xy	x^2	y^2
A	6	82	492	36	6724
B	2	86	172	4	7396
C	15	43	645	225	1849
D	9	74	666	81	5476
E	12	58	696	144	3364
F	5	90	450	25	8100
G	8	78	624	64	6084
$n = 7$	$\sum x = 57$	$\sum y = 511$	$\sum xy = 3745$	$\sum x^2 = 579$	$\sum y^2 = 38993$

Using the sums of the columns, compute:

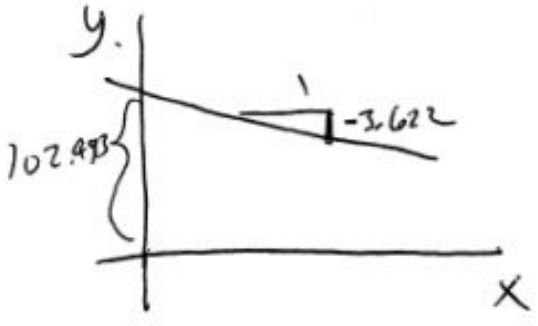
$$\begin{aligned} a &= \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \\ &= \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} \\ &= 102.493 \end{aligned}$$

and

$$\begin{aligned} b &= \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \\ &= \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} \\ &= -3.622 \end{aligned}$$

So

$$\begin{aligned} y' &= a + bx \\ y' &= 102.493 - 3.622x \end{aligned}$$



□

14.5.1: Relationship between correlation and slope

The relationship is

$$r = \frac{bs_x}{s_y} \tag{14.4.8}$$

where

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$
$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

are the standard deviations of the x and y datasets considered separately.
