

Introduction to Correlation and the Least Squares Line

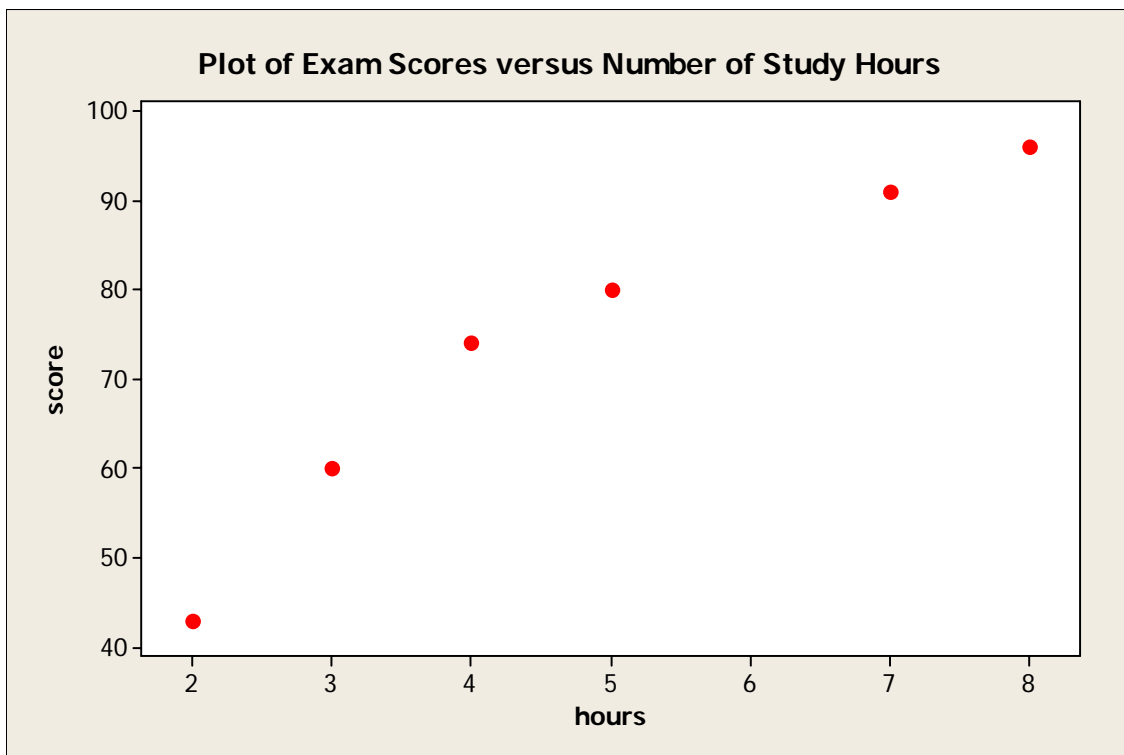
We present an example that will introduce the concepts of

- scatterplot
- correlation and linear relationship
- “least squares” line.

Example E1. The study hours (x) and exam scores (y) of six students are given in the table below.

x (hours)	5	2	8	3	4	7	
y (score)	80	43	96	60	74	91	

A scatter plot of y vs. x is given below.



The upward trend of the plot indicates a *positive correlation* between score and study hours.

- Given bivariate data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, the **sample correlation coefficient** r is defined as

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

The sample correlation coefficient r is a measure of the linear relationship between variables x and y . The value of r will always fall in the interval $[-1, 1]$.

- r^2 is called the **coefficient of determination** and it measures the proportion of the variation in the y values that can be explained by a linear relationship with the x values.

- We will find a “least squares” line that models the linear relationship between y and x . Such a line is of the form $\hat{y} = b_0 + b_1x$ and it minimizes the sum of squared residuals SSE .

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1x_i))^2.$$

The values for b_1 and b_0 that minimizes SSE are given by

$$b_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and $b_0 = \bar{y} - b_1\bar{x}$.

We let software do the number crunching. MINITAB software gives the following output :

Regression Analysis: score versus hours

The regression equation is

score = 34.193 + 8.236 hours

Predictor	Coef	SE Coef	T	P
Constant	34.193	6.150	5.56	0.005
hours	8.236	1.166	7.07	0.002

S = 6.03818 R-Sq = 92.6% R-Sq(adj) = 90.7%

From the above output we find that the least squares line is $\hat{y} = 34.193 + 8.236x$. A graph of the line is given below. It can be used to estimate expected scores for various hours of study. For example, an estimated expected exam score for a student who studies 6 hours is given by

$$\hat{y} = 34.193 + 8.236(6) = 83.609.$$

The slope coefficient 8.236 estimates the increase in score for each extra hour of study. We note that estimating an expected exam score for a number of hours that falls outside the scope of the data can be misleading and possibly disastrous.

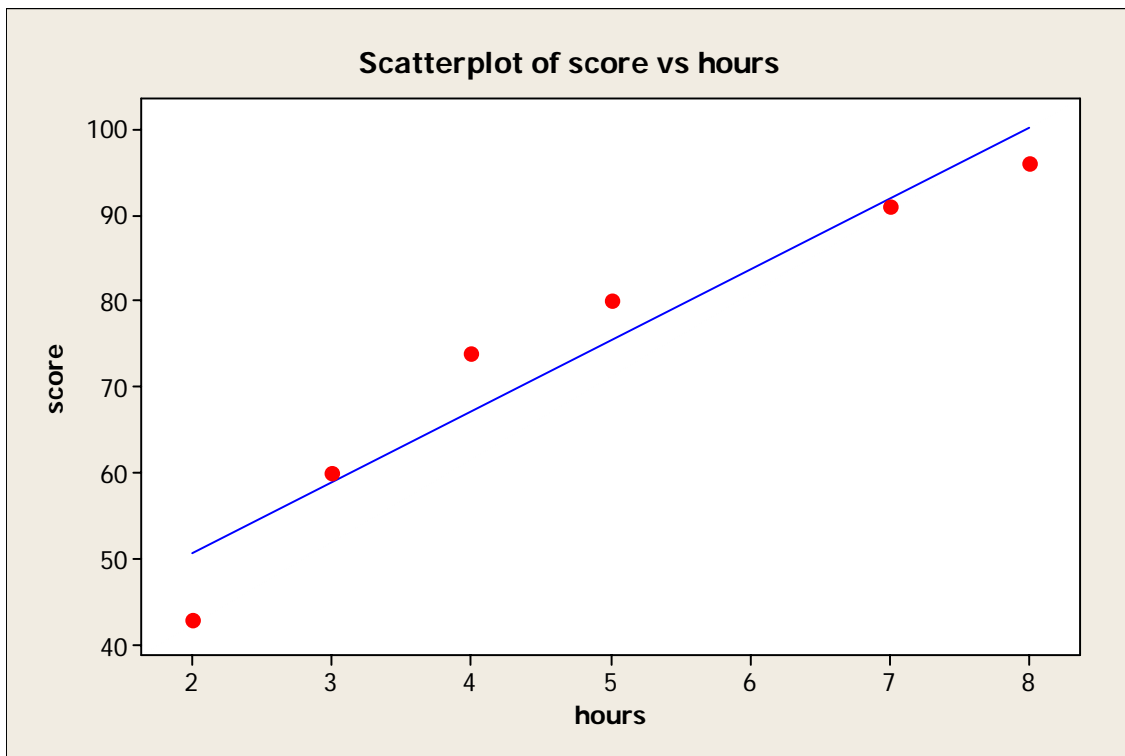


Table of Data with Fitted Values (\hat{y}) and Residuals (e)

x	5	2	8	3	4	7	
y	80	43	96	60	74	91	
\hat{y}	75.373	50.665	100.081	58.901	67.137	91.845	
e	4.2733	-7.66460	-4.08075	1.09938	6.86335	-0.84472	

A fitted value for a particular x value is the corresponding \hat{y} from the “fitted” line (the least squares line). Also note that the sum of all residuals equals zero, that is, $\sum_{i=1}^n e_i = 0$.