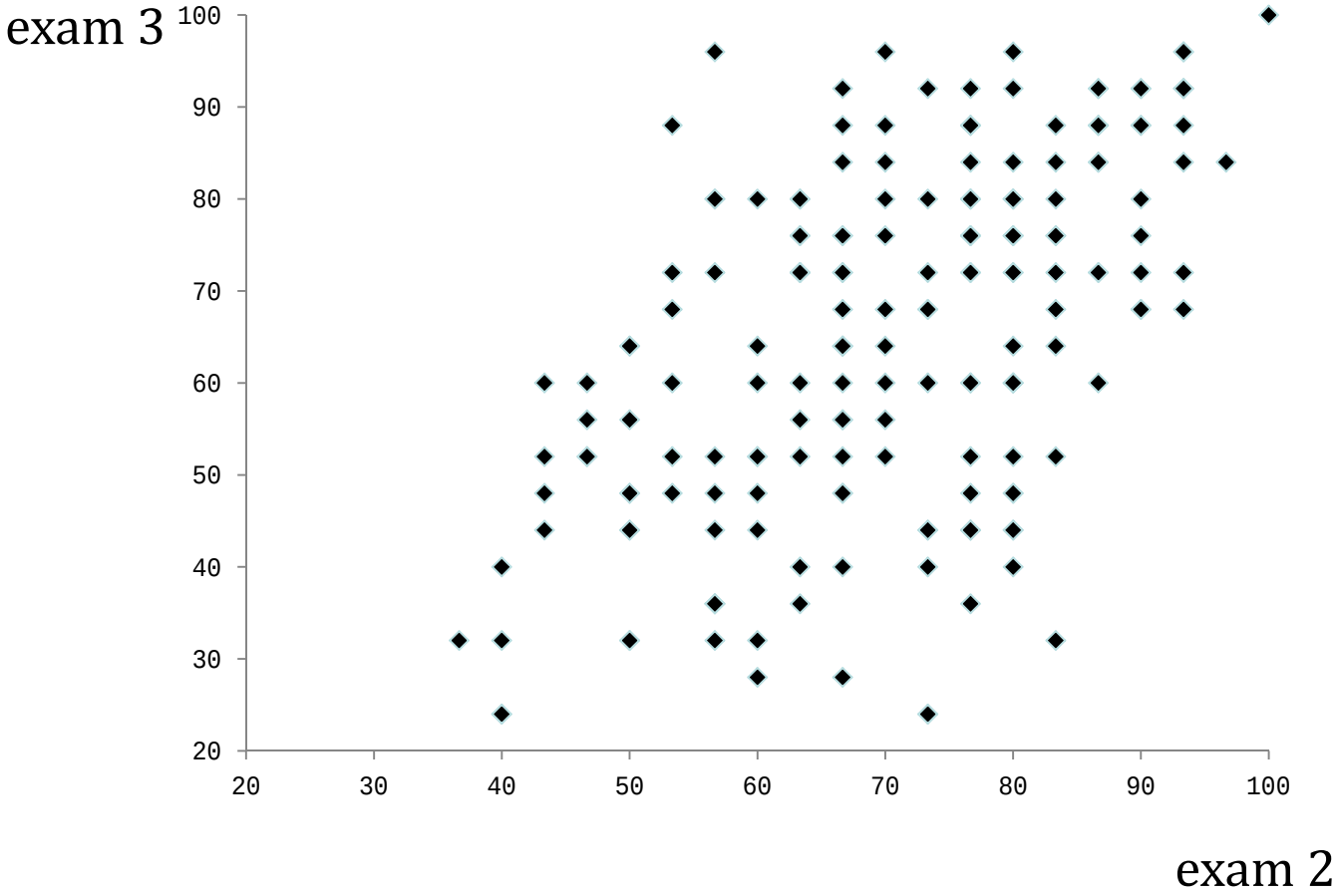
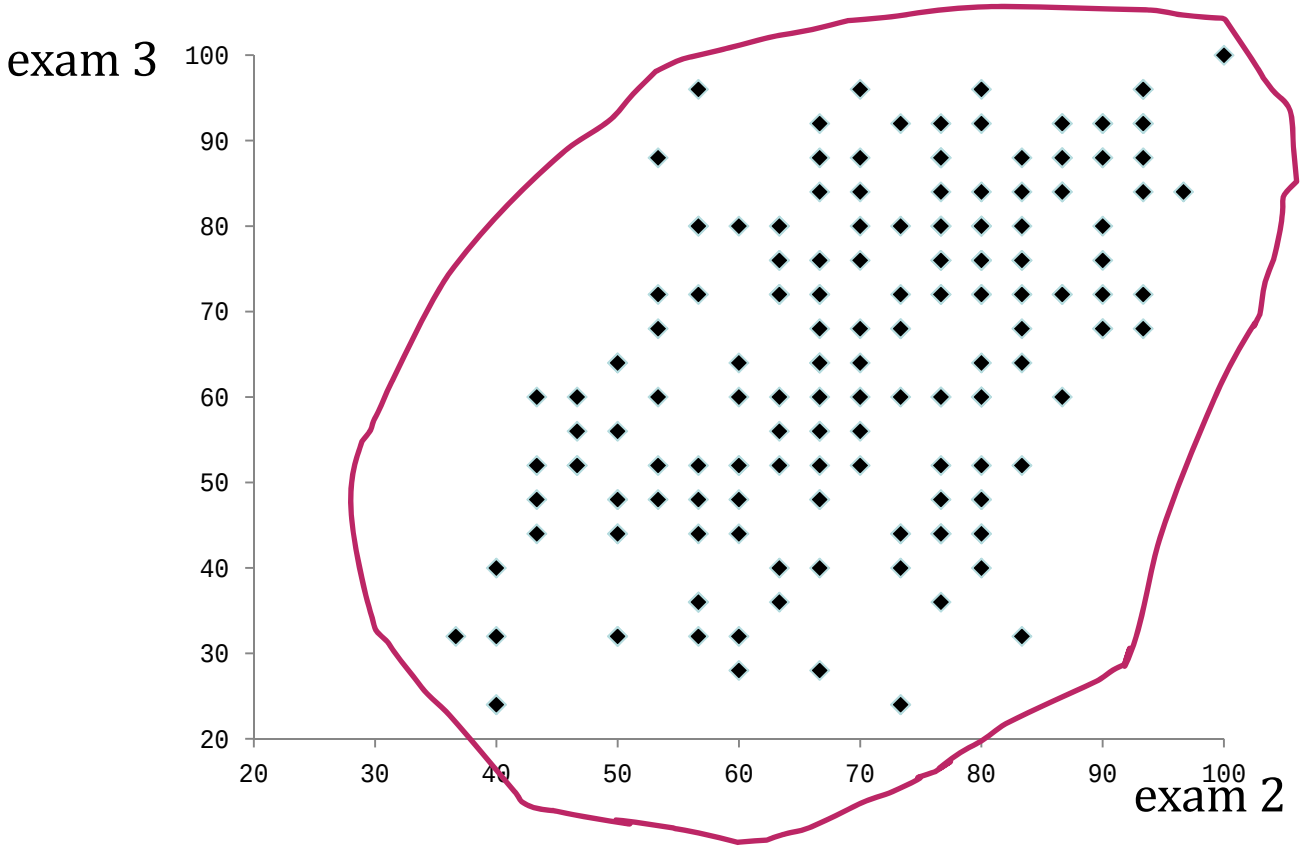
13.1 Simple Linear Regression Model

Example. Is there a relation between students' scores on exam 2 and exam3? If we plot the scores on exam 3 against the scores on exam 2, we get a **scatter diagram**.

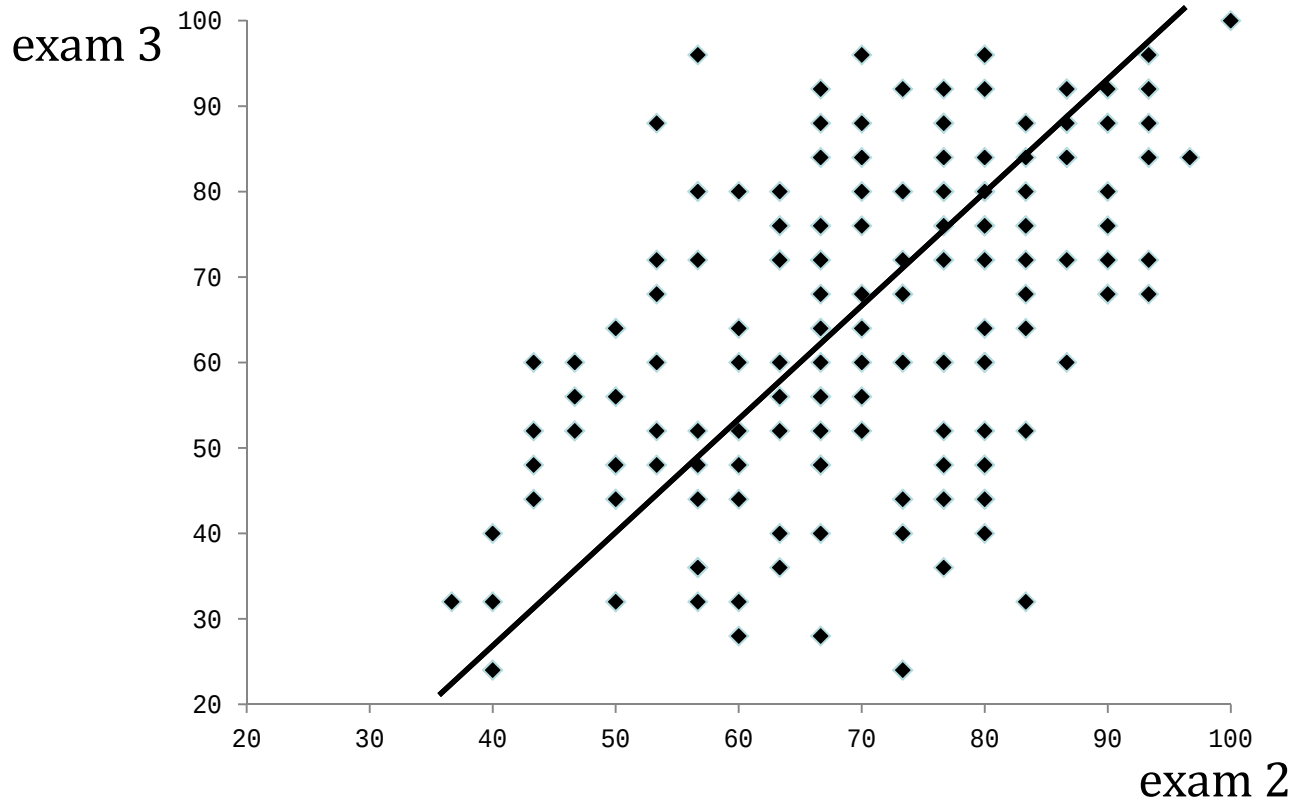
Scatter Diagram



Note that higher scores on exam 2 tend to imply higher scores on exam 3.



Note that the data are scattered around a straight line. Thus, the relation is linear (not curvilinear).

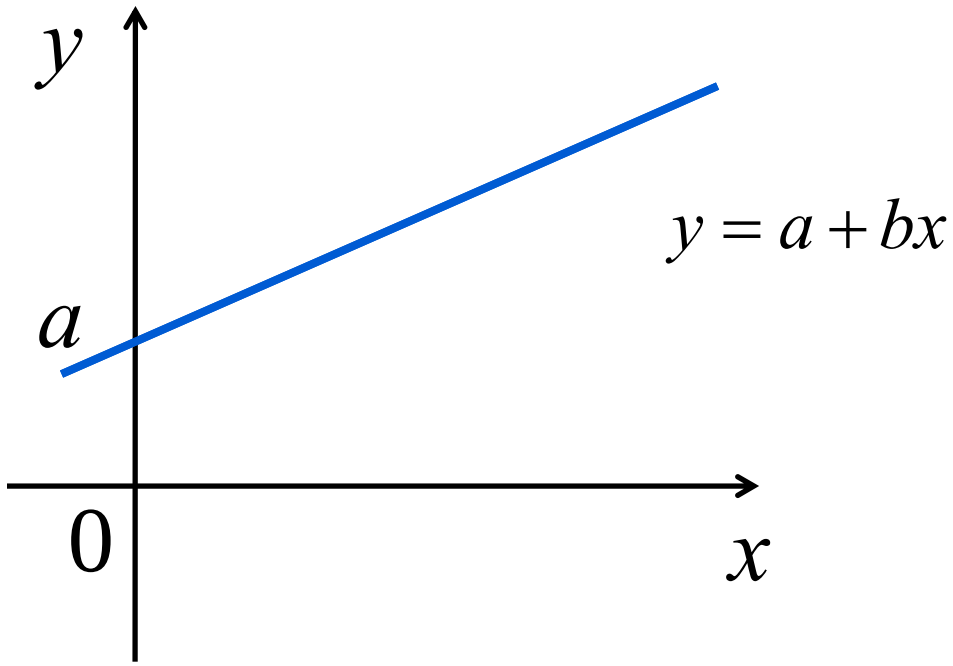


Definition. A **linear relation** between two variables x and y is described by a straight-line equation:

$$y = a + bx$$

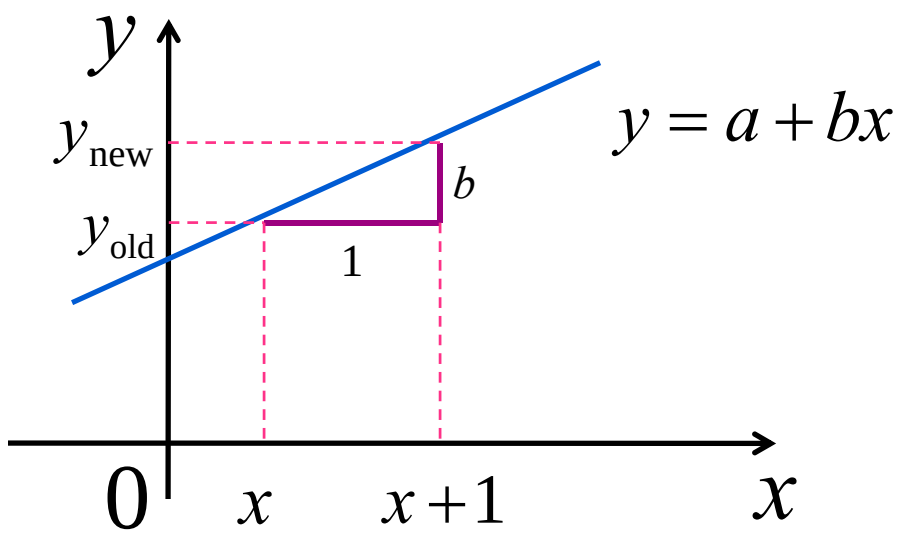
where a is the **intercept**, and b is the **slope**.

Definition. The **intercept** a is the value of y when $x = 0$, that is, where the straight line intercepts the y - axis.

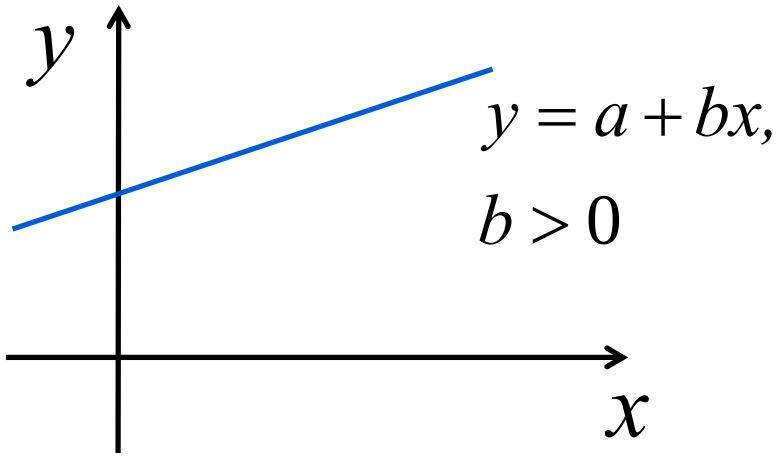


Definition. The **slope** b represents by how much y changes when x is increased by one unit. Indeed,

$$y_{\text{new}} - y_{\text{old}} = a + b(x+1) - [a + bx] = b.$$

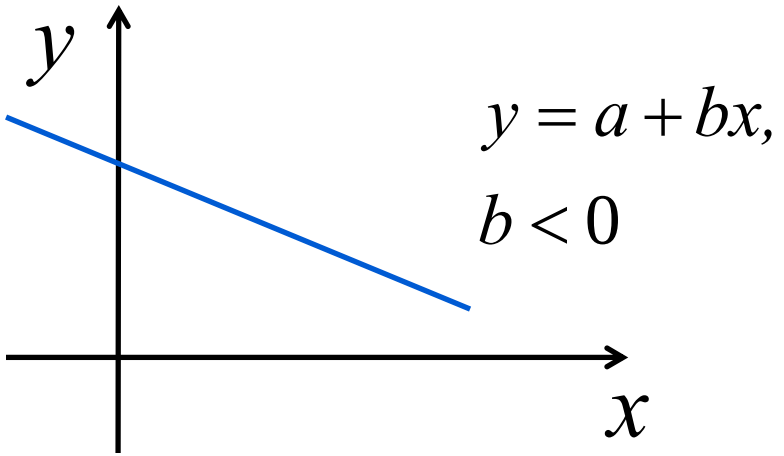


Definition. If the slope b is positive, the **linear relationship** between x and y is called **positive**.



In a positive linear relationship, as x increases, y increases. In other words, larger values of x tend to accompany larger values of y.

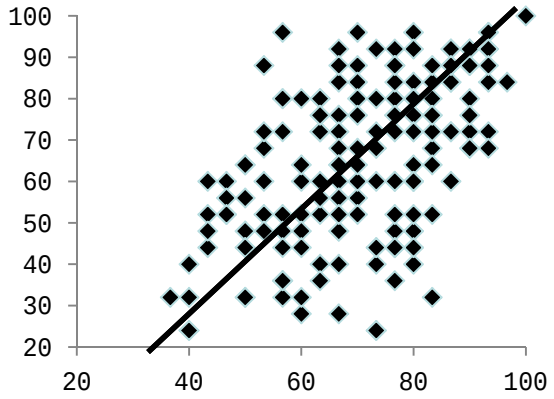
Definition. If the slope b is negative, the **linear relationship** between x and y is called **negative**.



In a negative linear relationship, as x increases, y decreases. In other words, larger values of x tend to accompany smaller values of y.

13.2 Simple Linear Regression Analysis

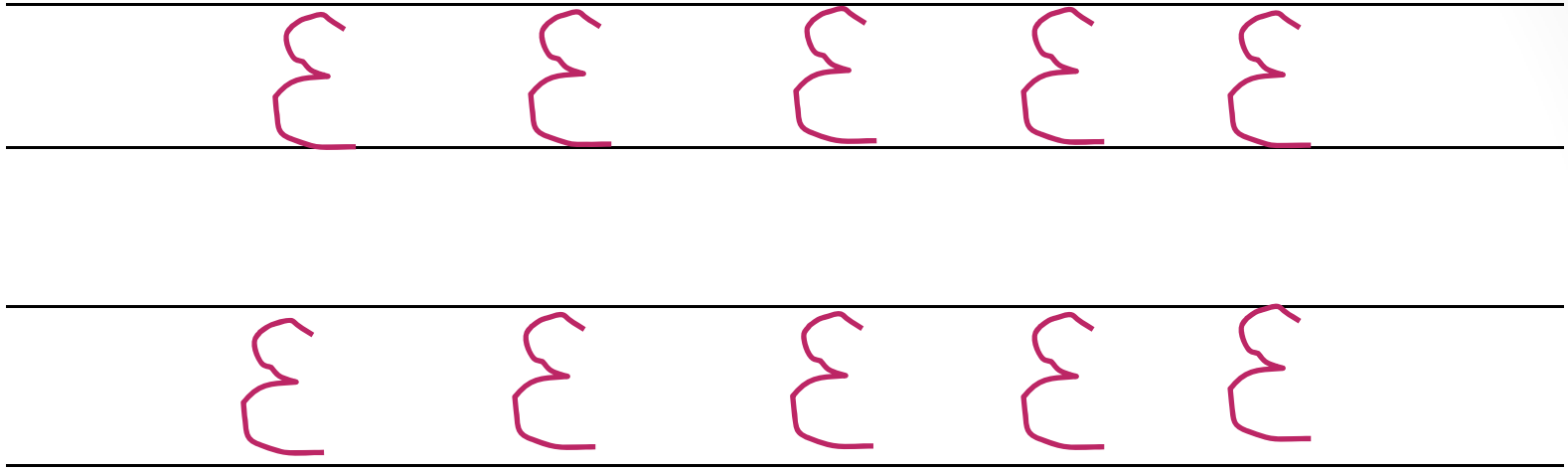
As seen on the scatter diagram, the relation between exam 2 and exam 3 scores is positive linear, but it is not a perfectly straight-line relation. The points are scattered around an imaginary straight line.



Definition. The simple linear regression model has the form

$$y = a + bx + \varepsilon$$

where y is a **dependent** variable, x is an **independent** variable, a is the **intercept**, b is the **slope**, and ε is a **random error**. It is assumed that ε a normally distributed random variable with mean zero and standard deviation σ.



epsilon /epp-sill-on/

The unknown parameters of the model are $a, b,$ and σ. They must be estimated from the data.

The straight line $\hat{y} = a + bx$ is called a **predicted** (or **fitted**) **line**, $\hat{y}$ is called a **predicted** (or **fitted**) **value**.

The predicted line is used for **prediction** of future observations.

Example. Suppose in our example with exam scores, the predicted line is

$$\hat{y} = 15 + 0.7x$$

where x is the score on exam 2, and $\hat{y}$ is the predicted score on exam 3.

The intercept of 15 hypothetically means that if the score on exam 2 was zero, the score on exam 3 is predicted to be15.

The slope of 0.7 means that if the score on exam 2 is increased by one point, the score on exam 3 will be increased by 0.7 points.

How to use the fitted line for prediction?

Example. Suppose a student got 85 on exam 2. What is his predicted score on exam 3?

Answer. $\hat{y} = 15 + (0.7)(85) = 74.5$

How to compute a and b?

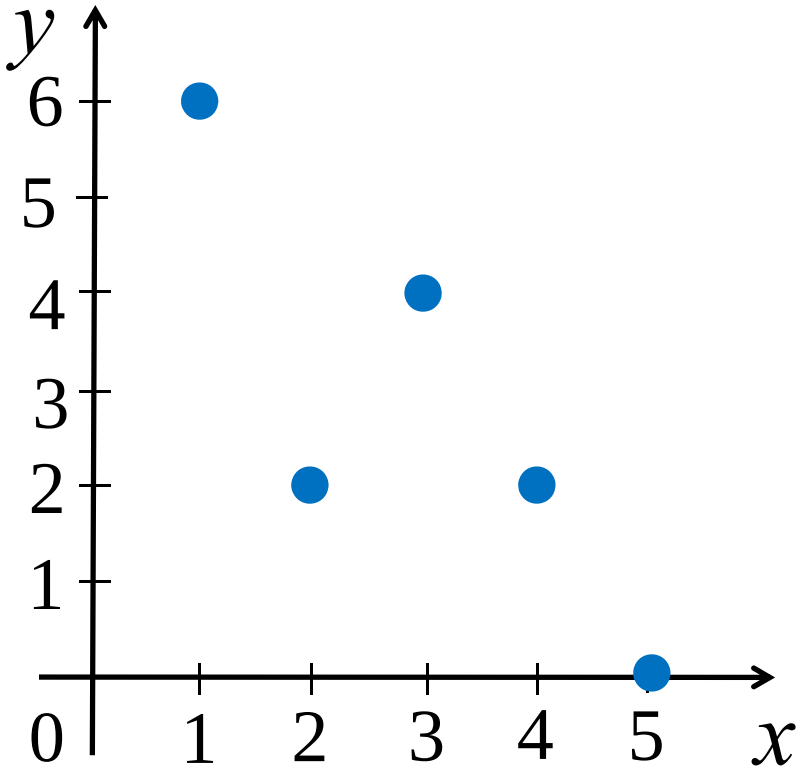
The computational formulas for a and b are:

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

and $$a = \bar{y} - b\,\bar{x}$$

Note that the predicted regression line passes through the point $(\bar{x}, \bar{y})$.

Example. Suppose points (1,6), (2,2), (3,4), (4,2), and (5,0) are observed. Here is the scatter diagram. Find the fitted regression line.



Solution. We compute the regression parameters as follows:

$$\sum xy = (1)(6)+(2)(2)+(3)(4)+(4)(2)+(5)(0) = 30,$$

$$\sum x = 1+2+3+4+5 = 15,\ n = 5,\ \bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3,$$

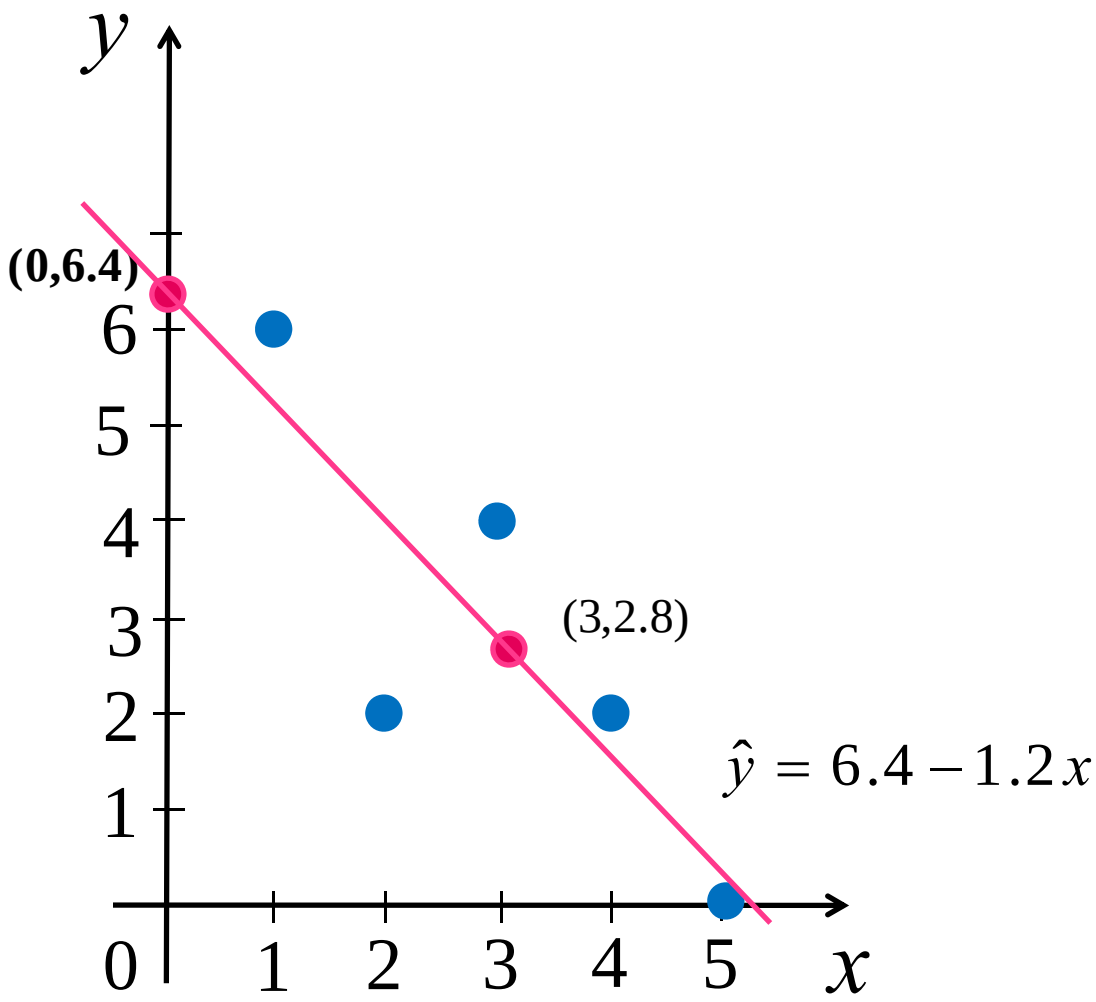
$$\sum y = 6+2+4+2+0 = 14,\ \bar{y} = \frac{\sum y}{n} = \frac{14}{5} = 2.8,$$

$$\sum x^2 = 1^2+2^2+3^2+4^2+5^2 = 1+4+9+16+25 = 55,$$

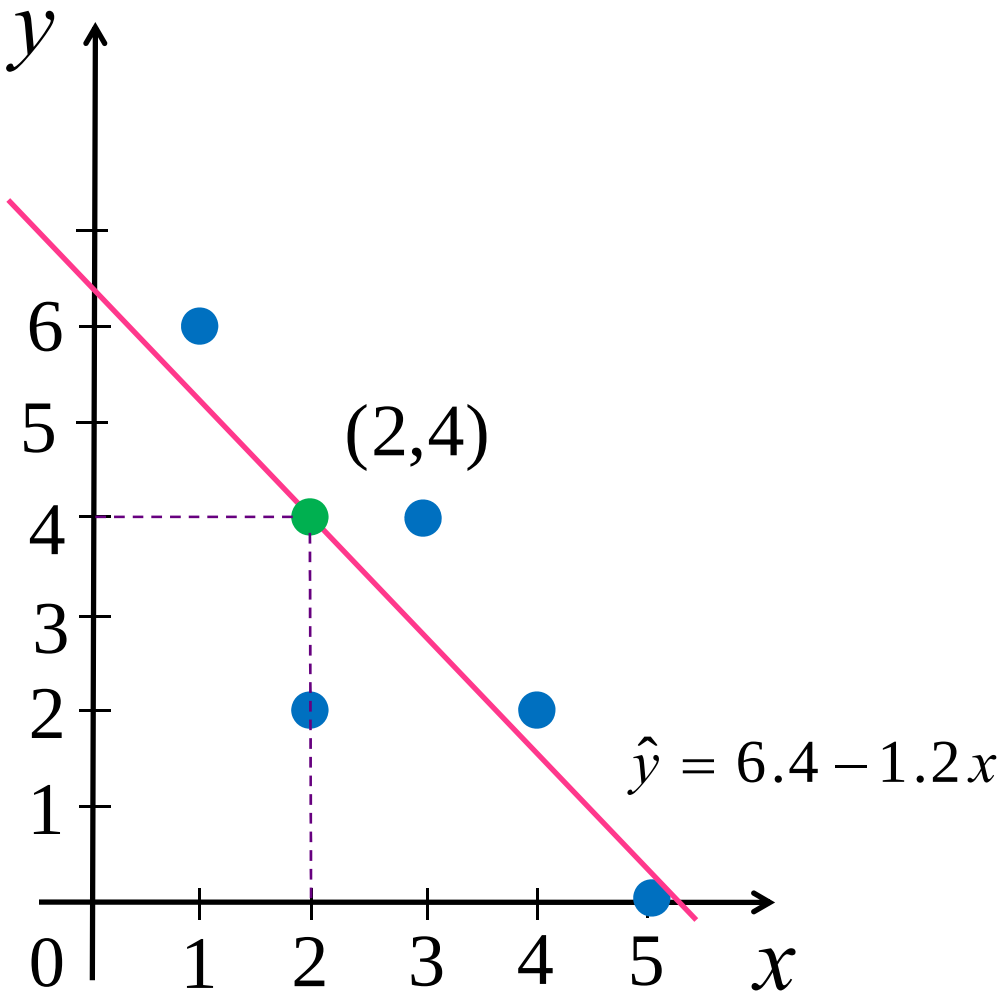
$$b = \frac{\sum xy - n\,\bar{x}\,\bar{y}}{\sum x^2 - n\,\bar{x}^2} = \frac{30-(5)(3)(2.8)}{55-(5)(3)^2} = \frac{30-42}{55-45} = \frac{-12}{10} = -1.2,$$

$$a = \bar{y} - b\,\bar{x} = 2.8-(-1.2)(3) = 2.8+3.6 = 6.4,$$

The fitted regression line is $\hat{y} = 6.4 - 1.2x.$


y
(0,6.4)
6
5
4
3
2
1
0
1
2
3
4
5
x
(3,2.8)
$\hat{y} = 6.4 - 1.2x$

Find predicted value of y for $x = 2$.



The predicted value is $\hat{y} = 6.4 - (1.2)(2) = 6.4 - 2.4 = 4.$