Descriptive Statistics With R Software

Fitting of Linear Models
::
Least Squares Method – One Variable

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Relationship Between Variables

Relationship exists between two variables.

Output of a variable is affected by one or more than one variables.

Example:

- **Yield of crop increases with an increase in quantity of fertilizer.**
- **Speed of electric fan (rotations per minute) increases as voltage increases.**
- **People drink more water as weather temperature increases.**
- **Yield of a crop depends upon other variables like quantity of fertilizer, rainfall, weather temperature, irrigation etc.**

Relationship Between Variables

Relationships are expressed through models.

Model:

Relationship among the variables depicting the phenomenon.

Relationship is characterized by variables and parameters.

What type of relationships ?

Relationship can be linear or nonlinear.

Input and Output Variables

Usually any phenomenon has two types of variables

- input variables and

- output variables.

- **Marks depend upon number of hours a student studies**

  or

  Number of hours of study depends upon the marks obtained by student.

- **Yield of a crop depends upon the rainfall and weather temperature**

  or

  Rainfall and weather temperature depends upon yield of crop.

Variables and Parameters

Example

Equation of a straight line

$$y = mx + c$$

c **: Intercept term**

m **: Slope of line**

x **: Values on x - axis**

y **: Values on y - axis**

Variables and Parameters

Example

<u>Option 1</u>: Knowing the values of (x, y), say $x = 4$, $y = 2$, can we know all the information about the line?

For example, $2 = 4m + c$

<u>Option 2</u>: Knowing the values of (m, c), say $m = 5$, $c = 6$, can we know all the information about the line?

For example, $y = 5x + 6$

Option 1 : Incorrect

Option 2 : Correct

Variables and Parameters

(m, c) : parameters

(x, y) : variables

Knowing the parameters is equivalent to knowing the line

$$y = mx + c$$

What We Have?

Suppose *X* denotes the quantity of fertilizer (in Kg.) and *Y* denotes the yield of a crop (in Kg.)

We want to find the relationship between *X* and *Y*.

Conduct an experiment and collect observations

x_1 = 1 Kg of fertilizer,	y_1 = 6 kg of yield is obtained
x_2 = 2 Kg of fertilizer,	y_2 = 7 kg of yield is obtained
x_3 = 3 Kg of fertilizer,	y_3 = 6 kg of yield is obtained

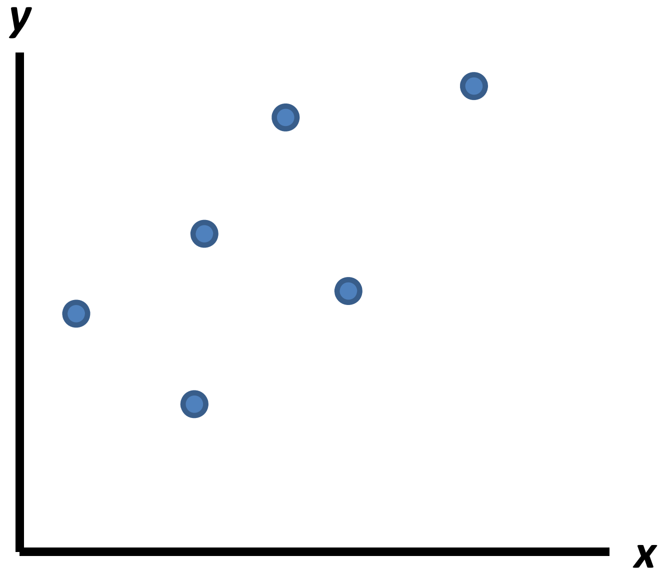
and so on.

What We Have?

Suppose we collect such n pairs of observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Then we plot the data



Using the graphical and analytical procedures, we find the equation of the curve representing the population on the basis of given data.

What We Want?

Example

Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

We know from experience that marks obtained by students increase as the number of hours increase.

Marks	337	316	327	340	374	330	352	353	370	380
Number of hours per week	23	25	26	27	30	26	29	32	33	34

Marks	384	398	413	428	430	438	439	479	460	450
Number of hours per week	35	38	39	42	43	44	45	46	44	41

What We Want?

Example

```
marks =
c(337,316,327,340,374,330,352,353,370,380,384,
398,413,428,430,438,439,479,460,450)

hours =
c(23,25,26,27,30,26,29,32,33,34,35,38,39,42,43
,44,45,46,44,41)
```

Representation

```
hours = c(23,25,26,…) marks = c(337,316,327,…)
```

x_1 = 23 hours, y_1 = 337 marks

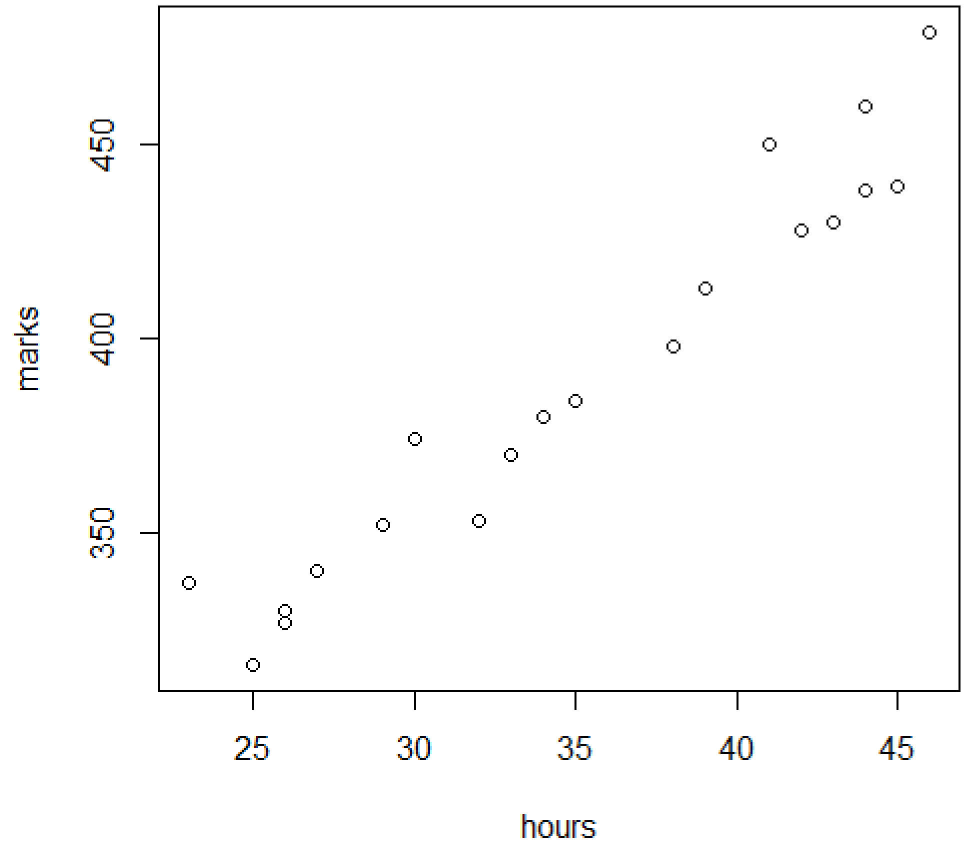
x_2 = 25 hours, y_2 = 316 marks

x_3 = 26 hours, y_3 = 327 marks and so on.

Scatter Plot

Example

```
plot(hours, marks)
```



Scatter Plots with Line

Example

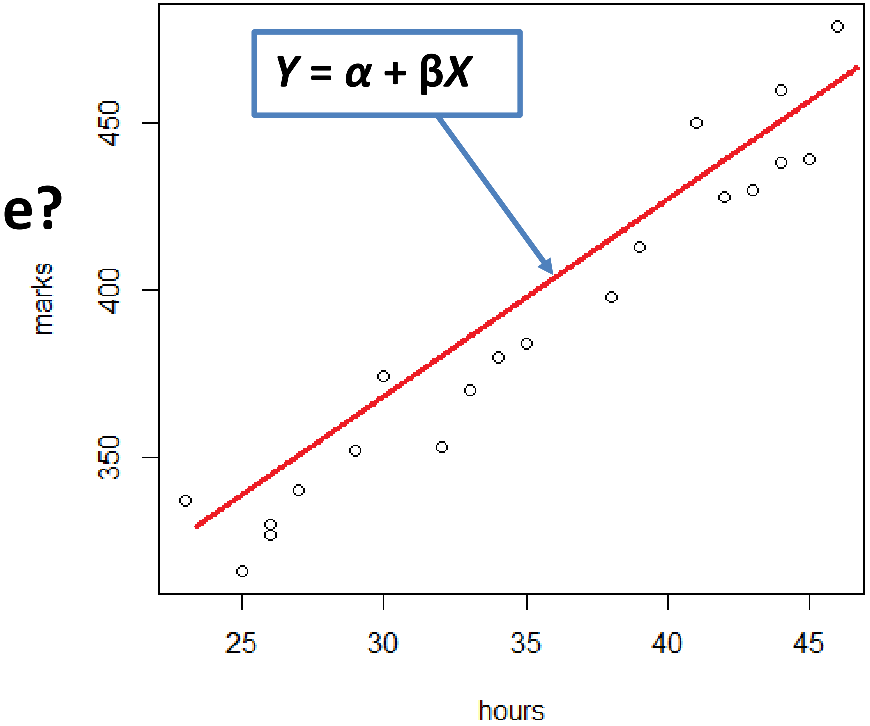
Next question?

What is the equation of this red line?

Let the equation of line be

$Y = \alpha + \beta X$

X : Hours, Y : Marks



We want to find the relationship between X and Y in terms of

$Y = \alpha + \beta X$

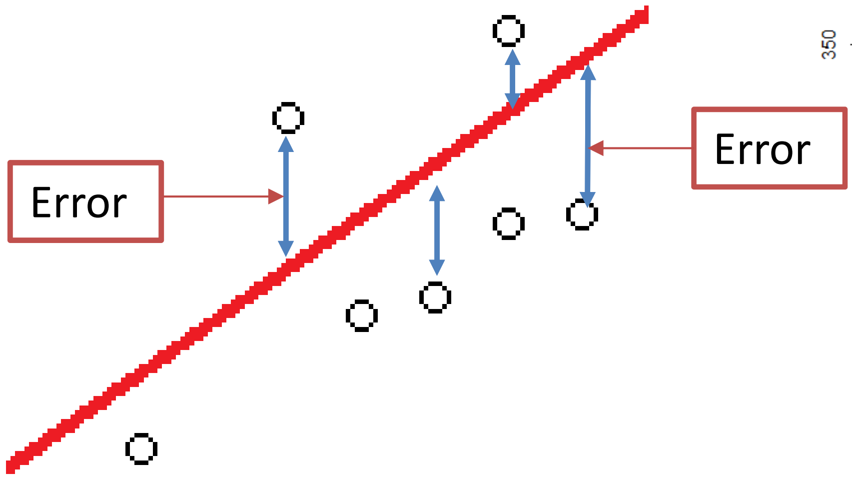
If we know α and β, the equation will be known.

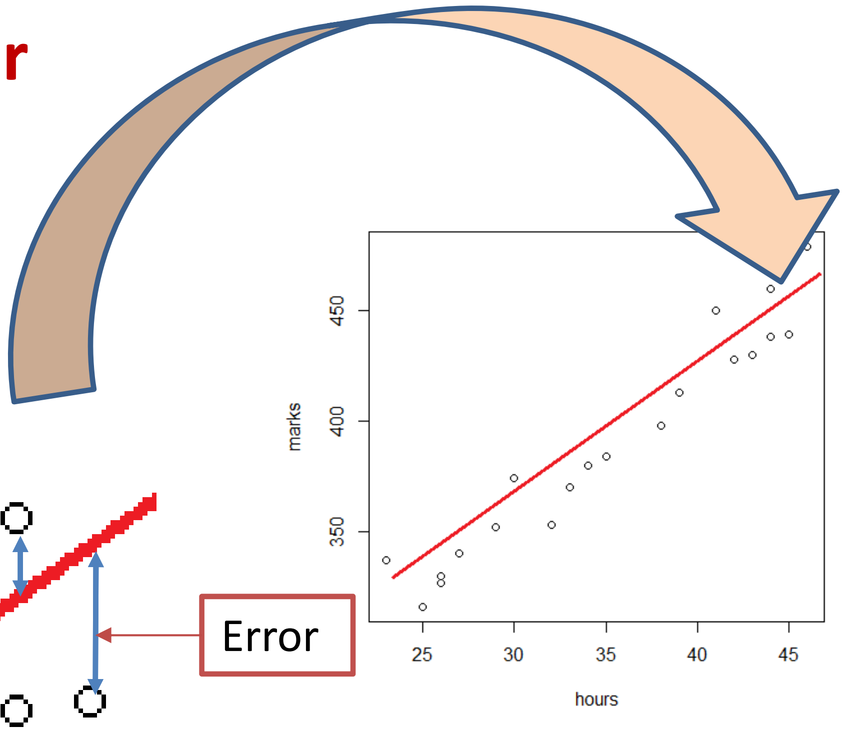
How to know α and β?

Scatter Plots with Line and Error

Example

Observations are not lying exactly on the line.





There is deviation between the observed points and the corresponding points lying over the line.

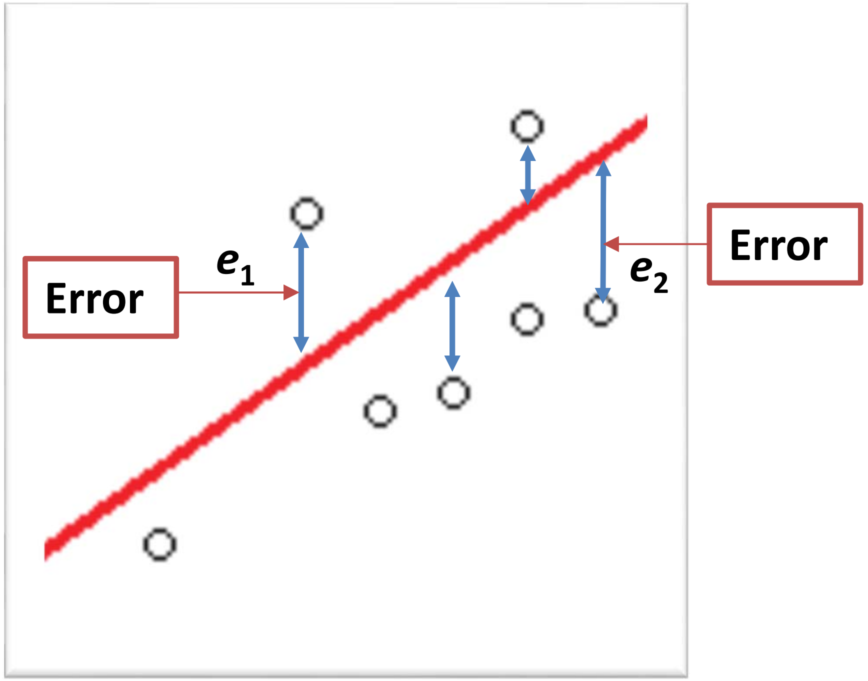
This is error.

What We Need?

Example

We find α and β such that the errors are minimum.

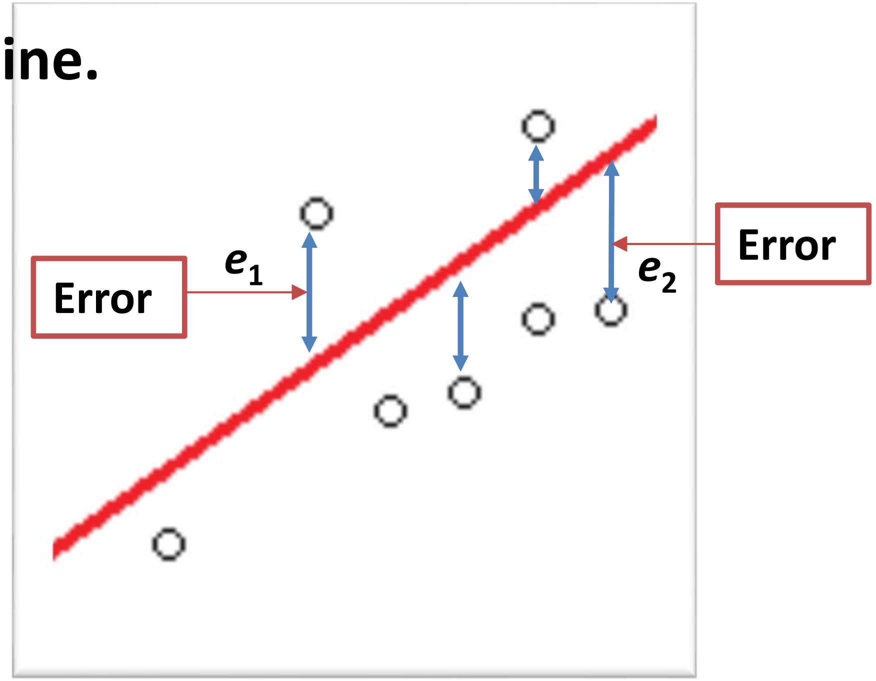
So minimize sum of such errors e_i's.



What to Do?

Example

Some errors/deviations are in positive direction and some errors are in negative direction with respect to line.



Hence the sum of errors/deviations may be close to zero indicating that there is no error or very small error.

Better option is to minimize the sum of squared errors.

How to Find the Line?

Example

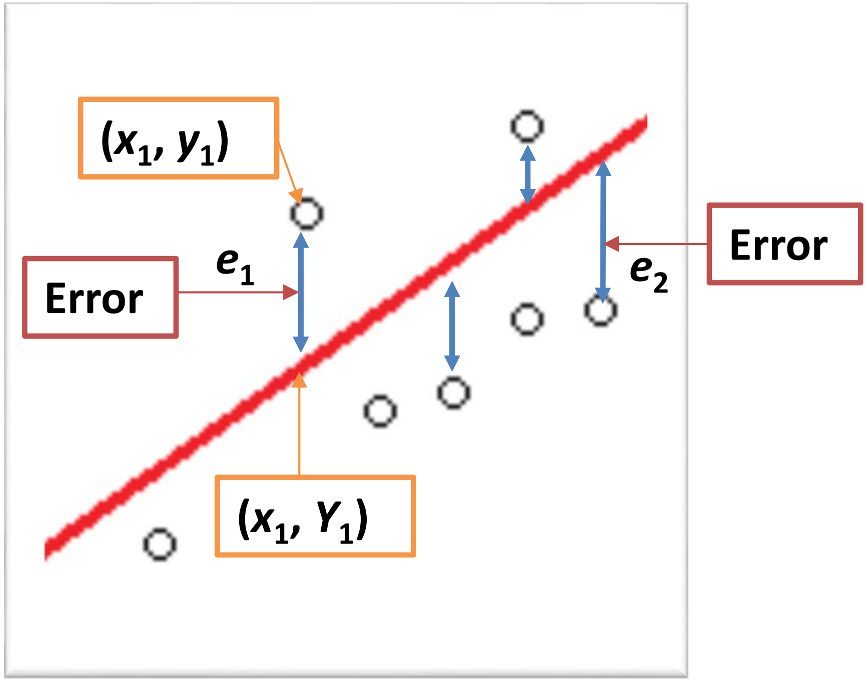
Suppose we collect such n pairs

of observations

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

and every pair (x_i, y_i) satisfies

$y_i = \alpha + \beta x_i + e_i$, $i = 1,2,\dots,n$



Find a line using the data set (x_i, y_i), $i = 1,2,\dots,n$ such that

- It passes through with maximum number of points
- The deviations of points with the fitted line are minimum.

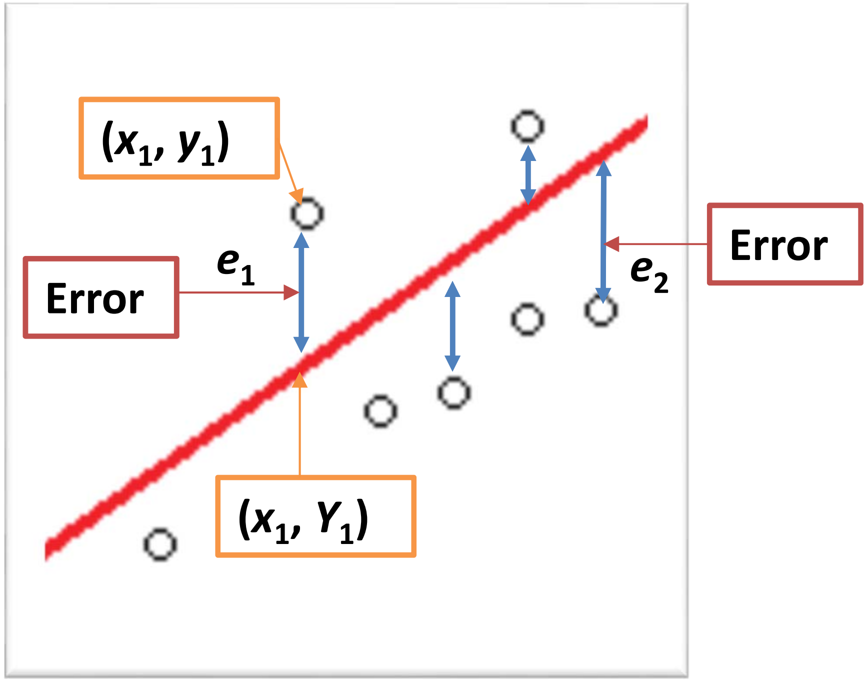
What are Errors/Deviations?

Example

Errors are the differences between

y_i and Y_i as

$e_i = y_i \sim Y_i$, $i = 1,2,\ldots,n$



We find α and β such that the sum of square of errors/deviations e_i's is minimum.

Method of Least Squares

Find the values of parameters such that the line passes through maximum number of given data points and the sum of squared errors/deviations from the line is minimum.

Use principle of maxima and minima to minimize $S = \sum_{i=1}^{n} e_i^2$

Method of Least Squares

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Find α and β such that S is minimum.

$$\frac{\partial S}{\partial \alpha} = 0 \quad => -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$\Rightarrow \alpha = \overline{y} - \beta \overline{x} \text{ provided } \beta \text{ is known, } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Method of Least Squares

$$\frac{\partial S}{\partial \beta} = 0 \quad => -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0$$

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \hat{\beta}, \text{ (denoted as } \hat{\beta})$$

$$\alpha = \overline{y} - \hat{\beta}\overline{x} = \hat{\alpha} \text{(denoted as } \hat{\alpha})$$

Method of Least Squares

$$\left.\frac{\partial^2 S}{\partial \alpha^2}\right|_{\alpha=\hat{\alpha}} > 0,$$

$$\left.\frac{\partial^2 S}{\partial \beta^2}\right|_{\beta=\hat{\beta}} > 0.$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Least squares estimate of β

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

Least squares estimate of *α*

Example

Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

We know from experience that marks obtained by students increase as the number of hours increase.

Marks	337	316	327	340	374	330	352	353	370	380
Number of hours per week	23	25	26	27	30	26	29	32	33	34

Marks	384	398	413	428	430	438	439	479	460	450
Number of hours per week	35	38	39	42	43	44	45	46	44	41

Method of Least Squares
Example

Solving it for the given data on `marks` and `hours`, we get the values of α and β as follows:

$$\bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 389.9, \qquad \bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 35.1$$

$$\hat{\beta} = \frac{\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{20}(x_i - \bar{x})^2} = 6.3,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 168.65$$

Model: marks = 168.65 + 6.3*hours