Welcome to

INTERNSHIP STUDIO

Module 04 | Lesson 03

Linear regression

Concept of Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

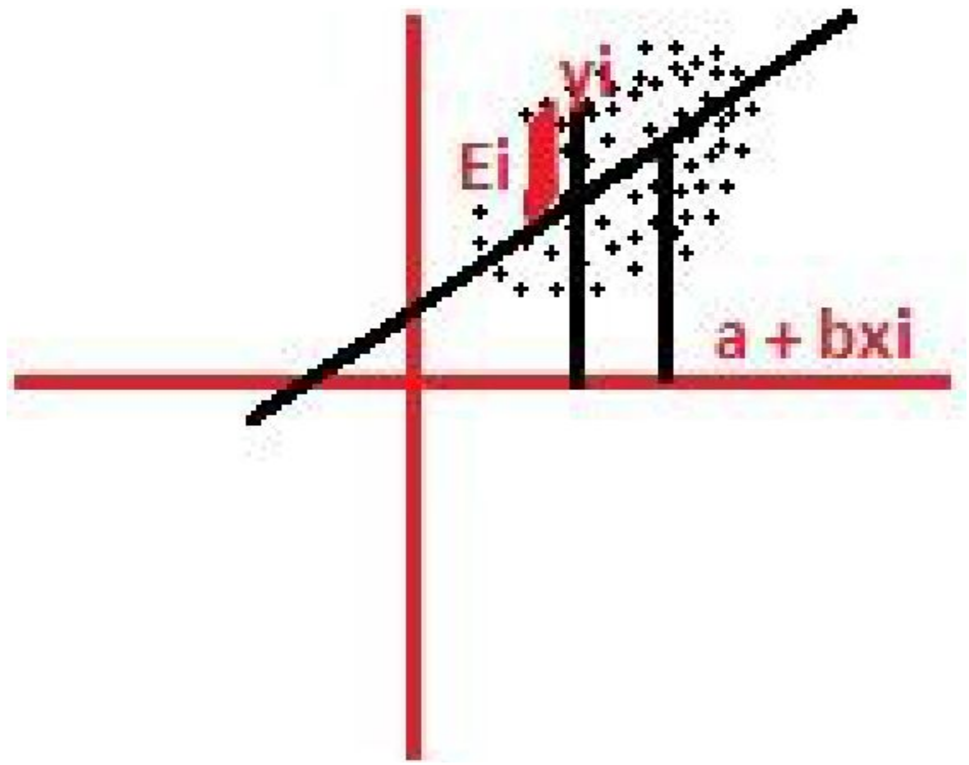
Concept of Linear Regression

Suppose:- $(x_i, y_i)\ ; \boldsymbol{where\ i} = 1, 2,3, \dots \dots \dots \dots \text{n};$

$$y_i = a + bx_i + \epsilon_i$$

$$\epsilon_i = y_i - a - bx_i$$

$$\sum \epsilon^2 = \sum (y_i - a - bx_i)^2$$



Linear Regression analysis: Draw a straight which fits all the points , also we have to minimize the error.

Take differentiation of $\sum \epsilon^2$ with respect to *a and b;*

$$Min \sum_{i=1}^{n} \epsilon_i^{\,2} \ \ over\ a\ and\ b$$

$$w1 = \frac{\sum y(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2}$$

Best line which fits all the sample points is obtained by:

$$y_i = \bar{y} + r\frac{s_y}{s_x}(x_i - \bar{x}) + \epsilon_i$$

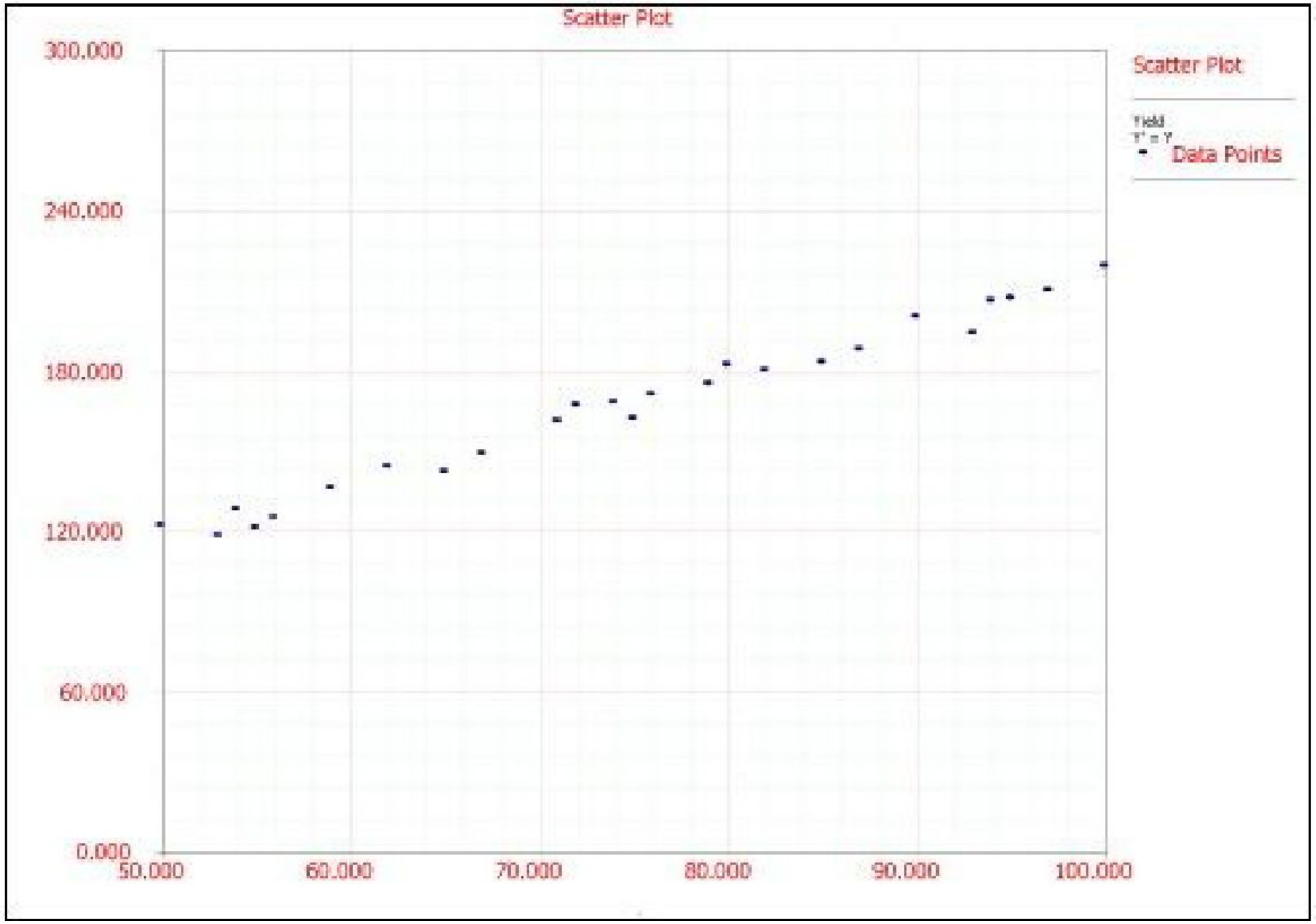
$$w2 = \frac{\sum x \sum y - n\sum xy}{(\sum x)^2 - n\sum x^2}$$

Where $r = corellation\ coefficient, s_y = \sqrt{var(y)}\ , s_x = \sqrt{var(x)}$

Given Example

Observation Number	(x_i)	(y_i)
1	50	122
2	53	118
3	54	128
4	55	121
5	56	125
6	59	136
7	62	144
8	65	142
9	67	149
10	71	161
11	72	167
12	74	168
13	75	162
14	76	171
15	79	175
16	80	182
17	82	180
18	85	183
19	87	188
20	90	200
21	93	194
22	94	206
23	95	207
24	97	210
25	100	219

Scatter Plot of given data set



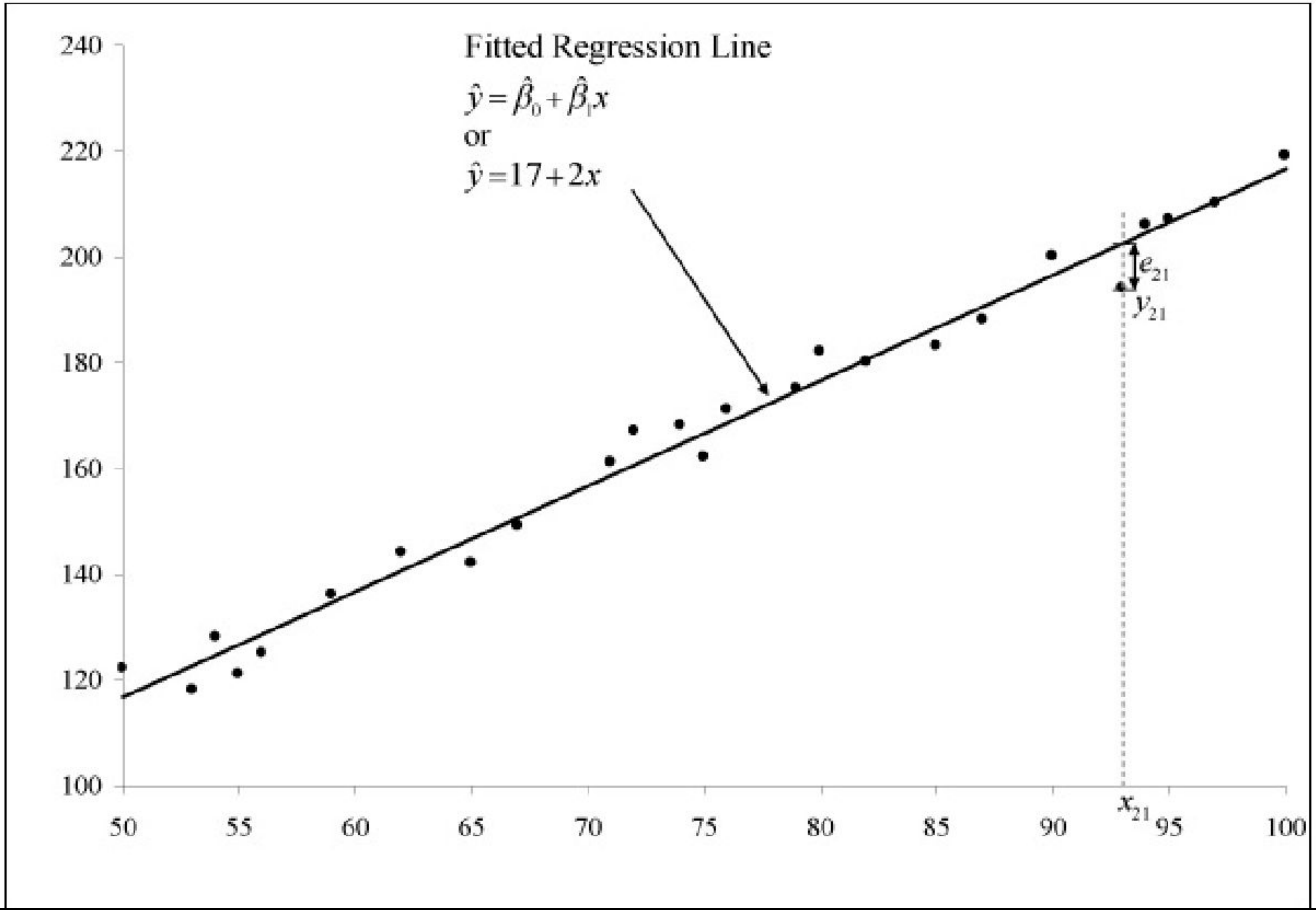
Concept of Linear Regression

No of observation	x	y	xy	x^2	y^2
1	50	122	6100	2500	14884
2	53	118	6254	2809	13924
3	54	128	6912	2916	16384
4	55	121	6655	3025	14641
5	56	125	7000	3136	15625
6	59	136	8024	3481	18496
7	62	144	8928	3844	20736
Σ	389	894	49873	21711	114690

$$w1 = \frac{\sum y(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2}$$

$$w2 = \frac{\sum x \sum y - n\sum xy}{(\sum x)^2 - n\sum x^2}$$

Fitted Regression Curve



Calculating Error

Once the fitted regression line is known, the fitted value of corresponding to any observed data point can be calculated. For example, the fitted value corresponding to the 21st observation in above Table is:

$$\hat{y}_{21} = \hat{\beta}_0 + \hat{\beta}_1 x_{21}$$
$$= (17.0016) + (1.9952) \times 93$$
$$= 202.6$$

The observed response at this point is y21 = 194 Therefore, the residual at this point is:

$$e_{21} = y_{21} - \hat{y}_{21}$$
$$= 194 - 202.6$$
$$= -8.6$$

Calculated Error Table

Standard Order	Actual Value (Y)	Fitted Value (YF)	Residual	Sta
1	122	116.76	5.24	
2	118	122.7455	-4.7455	
3	128	124.7407	3.2593	
4	121	126.7359	-5.7359	
5	125	128.731	-3.731	
6	136	134.7165	1.2835	
7	144	140.702	3.298	
8	142	146.6875	-4.6875	
9	149	150.6779	-1.6779	
10	161	158.6586	2.3414	
11	167	160.6537	6.3463	
12	168	164.6441	3.3559	
13	162	166.6392	-4.6392	
14	171	168.6344	2.3656	
15	175	174.6199	0.3801	
16	182	176.6151	5.3849	
17	180	180.6054	-0.6054	
18	183	186.5909	-3.5909	
19	188	190.5812	-2.5812	
20	200	196.5668	3.4332	
21	194	202.5523	-8.5523	
22	206	204.5474	1.4526	
23	207	206.5426	0.4574	
24	210	210.5329	-0.5329	
25	219	216.5184	2.4816	

IRIS Dataset Description

$$\underline{a}_{n\times1} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix} \quad \underline{b}_{n\times1} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$$

n dimensional column vector

```
5.1, 3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
```

```
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor
5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
```

Simple Linear Regression on IRIS data set

(x_i, y_i) where $i = 1, 2, \ldots \ldots \ldots \ldots \ldots .50$

Predictor $y_i = Mean\ (x_i) + i * standard\ deviation$

1) Let $x_1, x_2, \ldots \ldots .x_n \in \mathbb{R}$. The mean of $\underline{x}$ is defined as

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

3) variance of $x_1, x_2, \ldots \ldots .x_n \in \mathbb{R}$ is

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

Linear Regression

The value of Error is to be obtained by:-

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

Find the line and like to find the error and minimize the error;

$$(y_i, x_{1i}, x_{2i}, x_{3i}, \ldots \ldots \ldots . x_{ki})$$

$$y_i = a + b_1 x_{1i} + \ b_2 x_{2i} + \cdots \ldots \ldots \ldots \ldots \ldots \ldots . b_k x_{ki} + \epsilon_i \, ; (k+1)th\ parameter;$$

Multiple Linear Regressions ;

Linear Model : $\underline{Y} = A\underline{X} + \epsilon_i$; where A is the parameter matrix;

After linear model:- ANOVA

Reason of Experiment

ANOVA (f test)

Support Vector Regression (Kernel methods)