Correlation & Regression Analysis

The primary objective of correlation analysis is to measure the strength or degree of relationship between two or more variables. For example amount of fertilizer use and rice production, height and weight of a group of people.

If an increase in one variable corresponds to an increase in other, the correlation is said to be positive. If an increase in one variable corresponds to the decrease in other, the correlation is said to be negative. If two variables vary in such a way that their ratio is always constant then the correlation is said to be perfect.

Types of Correlation:

i. Positive or negative
ii. Simple or multiple
iii. Linear or non-linear

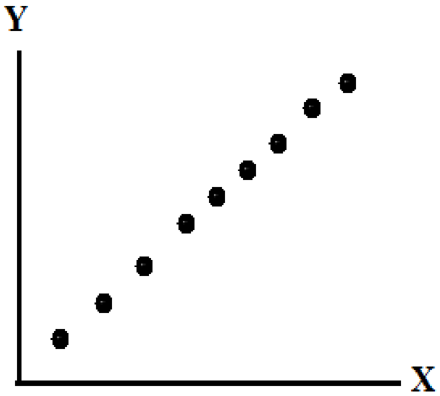
Methods of Estimating Correlation:

Correlation coefficient (r) determines a quantitative measure of the direction and strength of relationship between two or more variables.

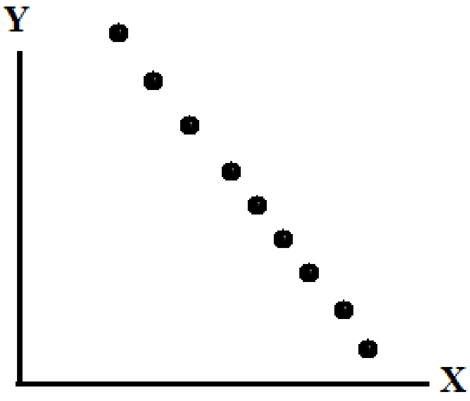
The following are the important methods of ascertaining **simple linear correlation**:

i. Scatter Diagram Method
ii. Karl Pearson's Coefficient of Correlation

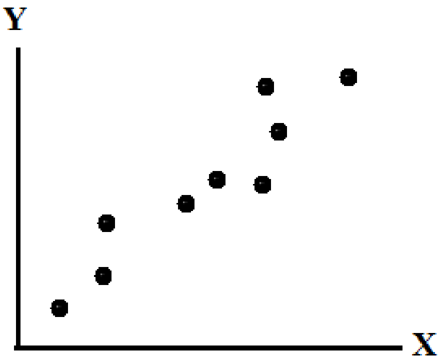
Scatter Diagram Method



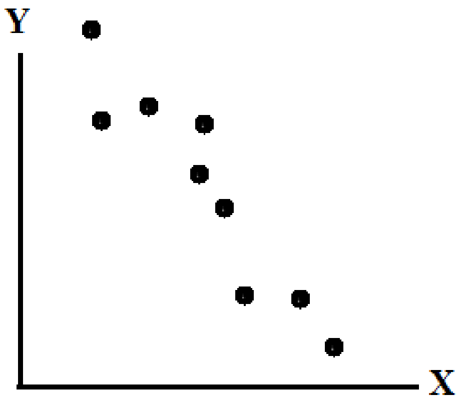
Perfect Positive Correlation



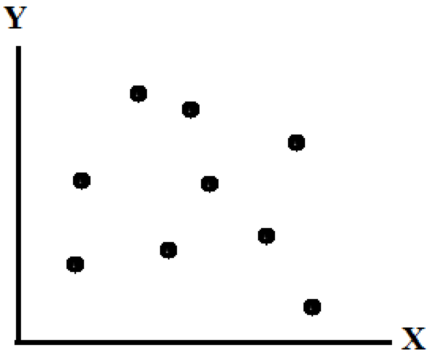
Perfect Negative Correlation



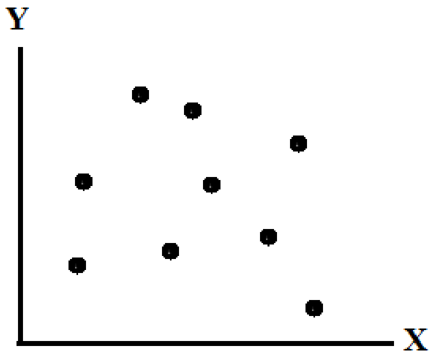
High degree of Positive Correlation



High degree of Negative Correlation



No Correlation



No Correlation

Karl Pearson's Coefficient of Correlation

If X and Y are two variables under study, then degree of relationship is measured by

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}$$

Where $\bar{X}$ and $\bar{Y}$ are the respective means of X and Y.

The above formula can be written as

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

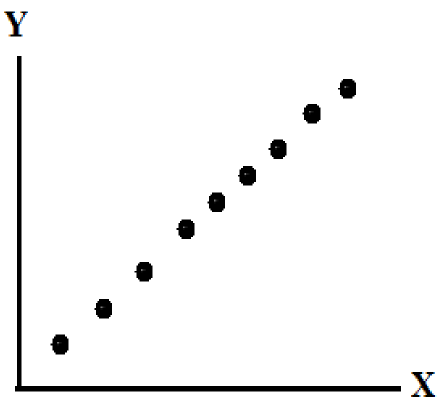
where $x = (X-\bar{X})$ and $y = (Y-\bar{Y})$

Interpretation of r:

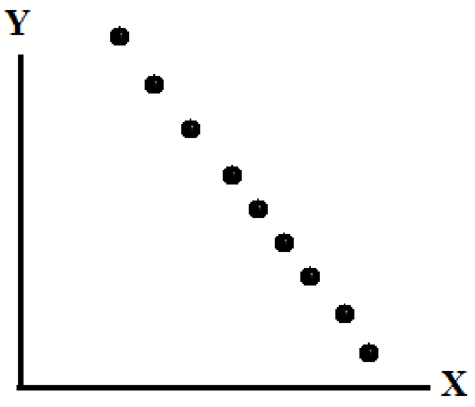
The values of the correlation coefficient lie between -1 and +1.

- A value of r = +1 indicate that X and Y perfectly related in a positive linear sense.
- A value of r = -1 indicate that X and Y perfectly related in a negative linear sense.
- Values of r close to +1 indicate a strong linear relationship between them with positive slope. Values of r close to -1 indicate a strong linear relationship between them with negative slope.

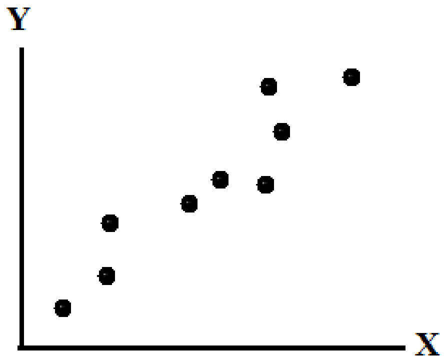
- Values of r close to 0 from positive side indicate a weak linear relationship between them with positive slope. Values of r close to 0 from negative side indicate a weak linear relationship between them with negative slope.
- Value of r =0 does not mean that X and Y are not related.



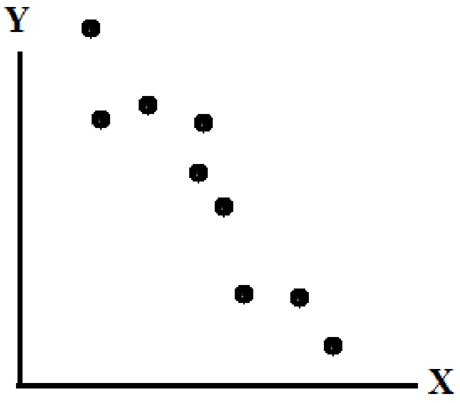
r = +1



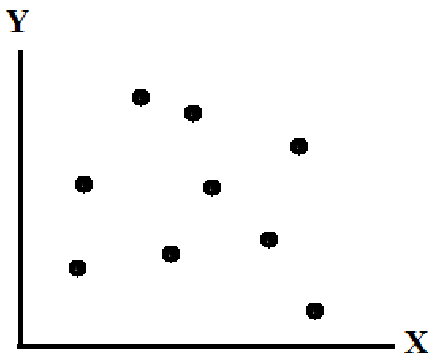
r = -1



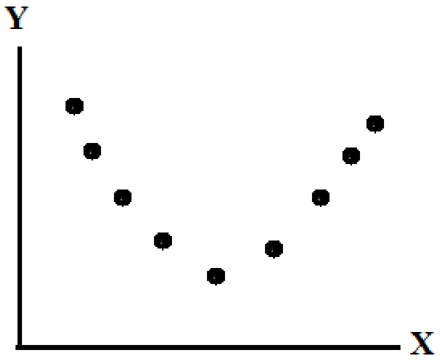
r close to +1



r close to -1



r = 0



$Y = X^2$ but r = 0

Example: Find Karl Pearson's correlation coefficient between the sales and expenses from the data given below and interpret its value:

Firm	1	2	3	4	5	6	7	8	9	10
Sales (Lakhs)	50	50	55	60	65	65	65	60	60	50
Expenses(Lakhs)	11	13	14	16	16	15	15	14	13	13

Solution:

Sales X	$(X-\overline{X})$	$(X-\overline{X})^2$	Expenses Y	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$	$(X-\overline{X})(Y-\overline{Y})$
50	-8	64	11	-3	9	+24
50	-8	64	13	-1	1	+8
55	-3	9	14	0	0	0
60	+2	4	16	+2	4	+4
65	+7	49	16	+2	4	+14
65	+7	49	15	+1	1	+7
65	+7	49	15	+1	1	+7
60	+2	4	14	0	0	0
60	+2	4	13	-1	1	-2
50	-8	64	13	-1	1	+8
$\sum X=580$	$\sum x=0$	$\sum x^2=360$	$\sum Y=140$	$\sum y=0$	$\sum y^2=22$	$\sum xy=70$

$$\overline{X}=\frac{\sum X}{N}=\frac{580}{10}=58\ \ \overline{Y}=\frac{\sum Y}{N}=\frac{140}{10}=14$$

$$r=\frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}=\frac{70}{\sqrt{360\times 22}}=0.787$$

Hence, there is a high degree of positive correlation between the two variables i.e., as the value of sales goes up, the expenses also goes up.

Example: Find Karl Pearson's correlation coefficient between the sales and expenses from the data given below and interpret its value:

Advertising Expenses(Lakhs)	10	12	15	23	20
Sales (Lakhs)	14	17	23	25	21

ANS: r = +0.865

Example: Show that the coefficient of correlation lies between -1 and +1.

Ans:

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}}$$

Let $a = \frac{(X-\bar{X})}{\sqrt{\Sigma(X-\bar{X})^2}}$ and $b = \frac{(Y-\bar{Y})}{\sqrt{\Sigma(Y-\bar{Y})^2}}$

Now

$$\sum(a+b)^2 = \sum a^2 + 2\sum a\,b + \sum b^2 = 1 + 2r + 1 = 2(1+r) \geq 0$$

i.e. $r \geq -1$

Similarly using $\sum(a-b)^2$ we get $r \leq 1$

Which concludes $-1 \leq r \leq 1$

Spearman's Rank Correlation Coefficient

Rank correlation method is applied when the rank order data are available or when each variable can be ranked in some order. The measure based on this method is known as rank correlation coefficient. This method is applied to the situation in which exact numerical measurement are not available. For instance, sincerity or honesty of each employee. Spearman's Rank Correlation Coefficient denoted by r_s and defined as

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6\sum D^2}{(N^3-N)}$$

D refers to the difference of ranks between paired items in two series.

[Proof: Link]

Example: Two managers are asked to rank a group of employee in order of potential for eventually becoming top managers. The rankings are as follows:

Employee	A	B	C	D	E	F	G	H	I	J
Ranking by Manager 1	10	2	1	4	3	6	5	8	7	9
Ranking by Manager 2	9	4	2	3	1	5	6	8	7	10

Compute the coefficient of rank correlation and comment on the value.

Solution:

Employee	R_1	R_2	$D^2 = (R_1 - R_2)^2$
A	10	9	1
B	2	4	4
C	1	2	1
D	4	3	1
E	3	1	4
F	6	5	1
G	5	6	1
H	8	8	0
I	7	7	0
J	9	10	1
N=10			$\sum D^2$=14

$$r_s = 1 - \frac{6\sum D^2}{(N^3 - N)} = 1 - \frac{6 \times 14}{990} = 0.915$$

Thus we find that there is a high degree of positive correlation in the ranks assigned by the two managers.

Example: Compute the rank correlation coefficient for the following data od 2 tests given to candidates for a critical job and comment on the value.

Preliminary test	92	89	87	86	83	77	71	63	53	50
Final test	86	83	91	77	68	85	52	82	37	57

Solution:

Preliminary test	R_1	Final test	R_2	$D^2 = (R_1 - R_2)^2$
92	10	86	9	1
89	9	83	7	4
87	8	91	10	4
86	7	77	5	4
83	6	68	4	4
77	5	85	8	9
71	4	52	2	4
63	3	82	6	9
53	2	37	1	1
50	1	57	3	4
N=10				$\sum D^2$=44

$$r_s = 1 - \frac{6\sum D^2}{(N^3 - N)} = 1 - \frac{6 \times 44}{990} = 0.733$$

Thus we find that there is a high degree of positive correlation between preliminary test and final test

Tie in Ranks:

An adjustment of the above formula is made when ranks are equal

$$r_s = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\ldots\ldots\right\}}{(N^3 - N)}$$

$m_1, m_2, \ldots \ldots ..$ each are no of the repeated numbers.

Example: An examination of eight applicants for a post was taken by a firm. From the marks obtained the applicants in the Bangla and English papers, Compute the rank correlation coefficient:

Applicant	A	B	C	D	E	F	G	H
Marks in Bangla	15	20	28	12	40	60	20	80
Marks in English	40	30	50	30	20	10	30	60

Solution:

Marks in Bangla	R_1	Final test	R_2	$D^2 = (R_1 - R_2)^2$
15	2	40	6	16
20	3.5	30	4	0.25
28	5	50	7	4
12	1	30	4	9
40	6	20	2	16
60	7	10	1	36
20	3.5	30	4	0.25
80	8	60	8	0
N=8				$\sum D^2$=81.5

Marks 20 is repeated 2 times, hence $m_1 = 2$ and Marks 30 is repeated 3 times, hence $m_2 = 3$.

$$r_s = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\ldots\ldots\right\}}{(N^3 - N)}$$

$$= 1 - \frac{6\left\{81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{(8^3 - 8)}$$

$$= 0$$

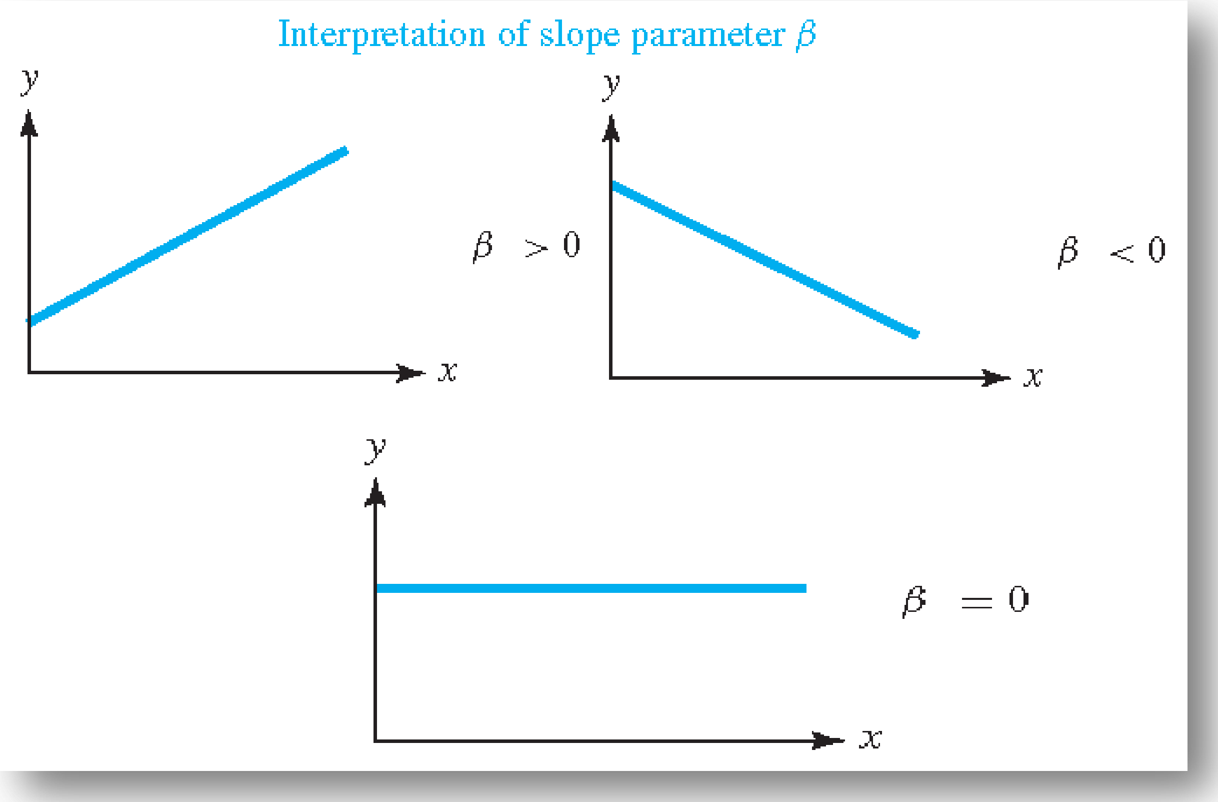
Regression Analysis

The regression analysis is a technique of studying the dependence of one variable (called dependent variable), on one or more variables (called independent variables), with a view to estimating or predicting the average value of the dependent variable in terms of the known or fixed values of the independent variables.

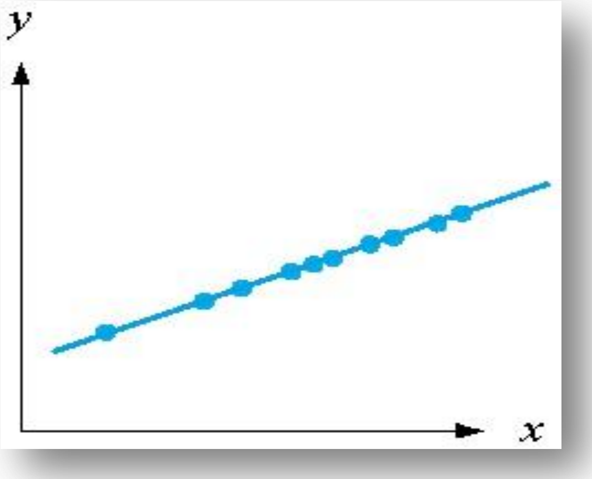
Example: Variable 1: Distance to transmitter: X
Variable 2: Wireless signal strength: Y

Let's assume a linear relationship between Y and X is reasonable as

$$\mathbf{Y = \beta X + \alpha}$$

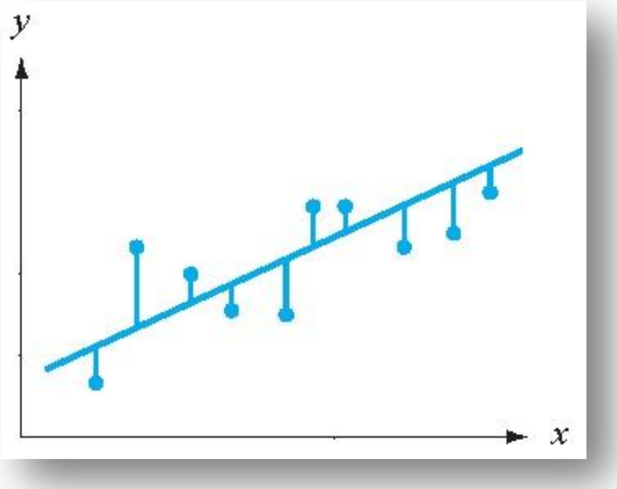


- If the relationship between Y and X is exact this is called deterministic relationship.
- If the relationship between Y and X is not exact this is called non-deterministic relationship. In real life application there are many sources of randomness. Randomness means the same value of X does not always give the same value of Y.



A deterministic relationship

$$\mathbf{Y = \beta X + \alpha}$$



A non-deterministic relationship

$$\mathbf{Y = \beta X + \alpha + \epsilon}$$

Primary Objectives of Regression Analysis:

i. To estimate the relationship that exits, on the average, between the dependent variable and independent variables.

ii. To determine the effect of each independent variable on the dependent variable, controlling the effects of the others independent variables.

iii. To predict the value of the dependent variable for a given value of the explanatory variables.

Simple Linear Regression Model

The simplest form of the regression model that displays the relation between X and Y is a straight line, which appears as follows:

$$\mathbf{\hat{Y} = bX + a}$$

Where $\hat{Y}$ denotes the predicted value of Y, a is the intercept and b is the slope of the straight line. In regression terminology, b is the regression coefficient of Y on X. This straight line is called the fitted line of Y.

In practice, the observed value of Y would almost always invariably deviate from the expectation. If this discrepancy denoted by a quantity $\boldsymbol{\epsilon}$. Then

$$\mathbf{\epsilon = Y - \hat{Y}}$$

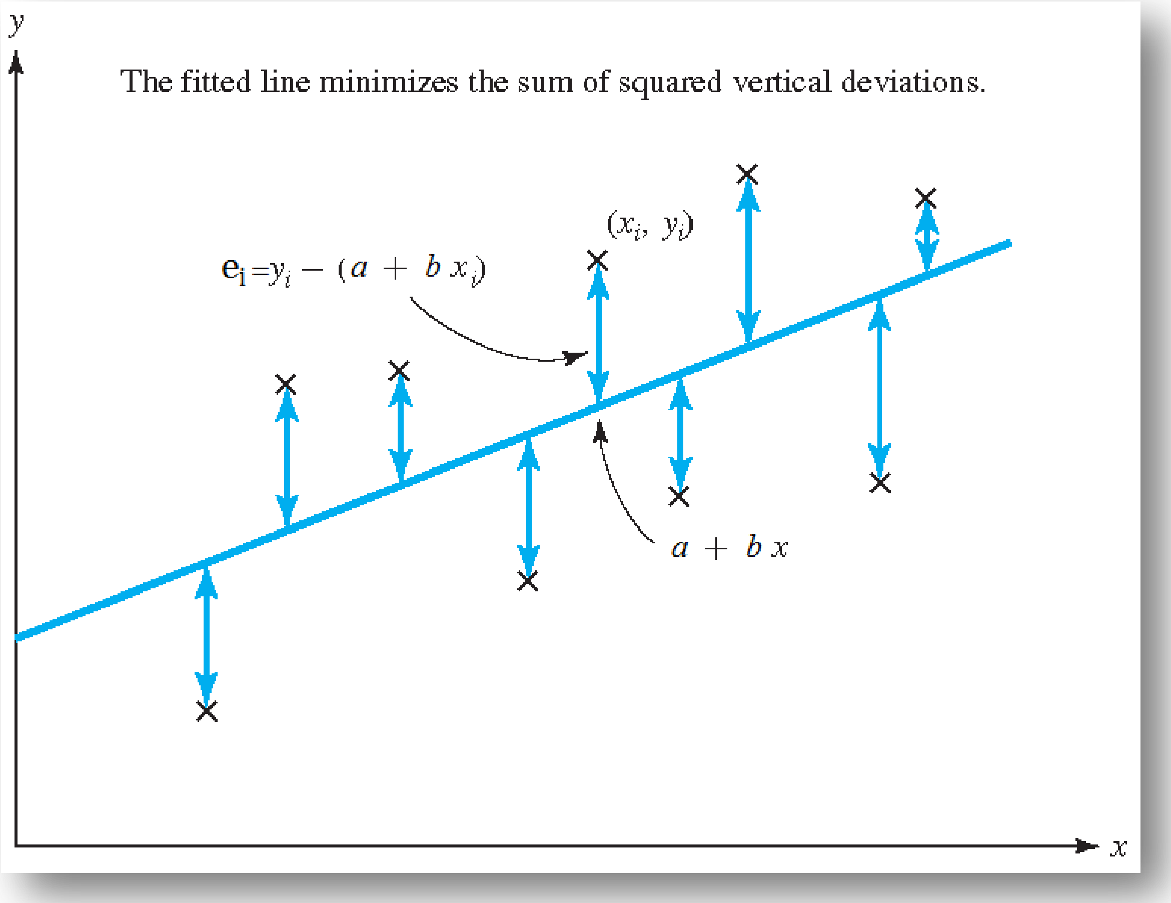
$$\mathbf{i.e.\, Y = bX + a + \epsilon}$$

The least-Squares Method:

The least-squares method is a technique for minimizing the sum of the squares of the differences between the observed values and estimated values of the dependent variable. That is the least-squares line is the line that minimizes

$$\boldsymbol{\sum e_i^2 = \sum \left(Y_i - \widehat{Y}_i\right)^2 = \sum (Y_i - bX_i - a)^2}$$

Here $e_i =$ deviation of Y_i from $\widehat{Y}_i$ and $\sum e_i^2$ is called sum of squares of errors (SSE).


y
The fitted line minimizes the sum of squared vertical deviations.
$(x_i,\ y_i)$
$e_i = y_i - (a + b\,x_i)$
$a + b\,x$
x

To minimizes SSE with respect to a and b, from calculus we know that the partial derivatives of SSE with respect to a and b must be 0. Then

$$\frac{\partial \mathrm{SSE}}{\partial a} = -2\sum (Y_i - bX_i - a) = 0$$

$$\frac{\partial \mathrm{SSE}}{\partial b} = -2\sum (Y_i - bX_i - a)\,X_i = 0$$

Which concludes

$$\sum Y_i = na + b\sum X_i$$

and

$$\sum X_iY_i = a\sum X_i + b\sum {X_i}^2$$

From the above equations we get

$$\mathbf{b = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum {X_i}^2 - (\sum X_i)^2}}$$

and $$\mathbf{a = \bar{Y} - b\bar{X}}$$

Example: The following table show distance to transmitter (X) and corresponding wireless signal strength (Y).

Distance to transmitter (m)	13	1	17	19	14	15	15	8	13	3
wireless signal strength (dB)	34.4	38.4	30.4	29.7	30.1	33.9	32.8	35.2	34.9	36.8

i. Find the regression line of Y on X.

ii. Predict what the signal strength would be if the distance was 10 meters.

Solution:

X_i	Y_i	${X_i}^2$	X_iY_i
13	34.4	169	447.2
1	38.4	1	38.4
17	30.4	289	516.8
19	29.7	361	564.3
14	30.1	196	421.4
15	33.9	225	508.5
15	32.8	225	492
8	35.2	64	281.6
13	34.9	169	453.7
3	36.8	9	110.4
$\sum X_i$=118	$\sum Y_i$=336.6	$\sum {X_i}^2$=1708	$\sum X_iY_i$=3834.3

i.

$$b = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum {X_i}^2 - (\sum X_i)^2} = \frac{10(3834.3) - (118)(336.6)}{10(1708) - (118)^2} = -0.44$$

$$a = \bar{Y} - b\bar{X} = \frac{336.6}{10} + 0.44\frac{118}{10} = 38.8$$

The regression line of Y on X is

$$\hat{\mathbf{Y}} = \mathbf{bX + a = -0.44\,X + 38.8}$$

ii. For X=10 $\hat{Y} = 34.4$ dB

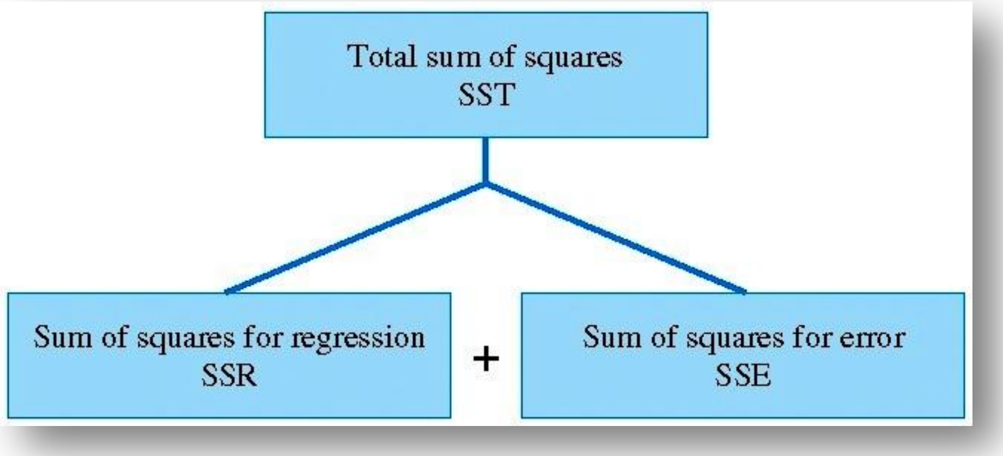
- In the case when X is assumed to be independent variable and Y as dependent variable, the regression is said to be regression of Y on X and the estimating regression line is $\hat{\mathbf{Y}} = \mathbf{bX + a.}$ When X acts as dependent variable and Y as independent variable, the regression is said to be regression of X on Y and the estimating regression line is $\hat{\mathbf{X}} = \mathbf{dY + c}$ with

$$d\,\cancel{\mathbf{b}} = \frac{\mathbf{n \sum X_i Y_i - \sum X_i \sum Y_i}}{\mathbf{n \sum {Y_i}^2 - (\sum Y_i)^2}}$$

$$\mathbf{c = \bar{X} - d\bar{Y}}$$

Goodness of Fit in Regression and Coefficient of Determination:

$Y_i - \bar{Y}$ is called the total deviation and corresponding $\sum(Y_i - \bar{Y})^2$is called total sum of squares (SST), $\hat{Y}_i - \bar{Y}$ is called explained deviation and corresponding $\sum\left(\hat{Y}_i - \bar{Y}\right)^2$ is called sum of squares for regression (SSR) and $Y_i - \hat{Y}_i$ is called unexplained deviation and corresponding $\sum\left(Y_i - \hat{Y}_i\right)^2$ is called sum of squares for error (SSE). The relation among SSE, SST, SSR is



Total sum of squares
SST

Sum of squares for regression
SSR

+

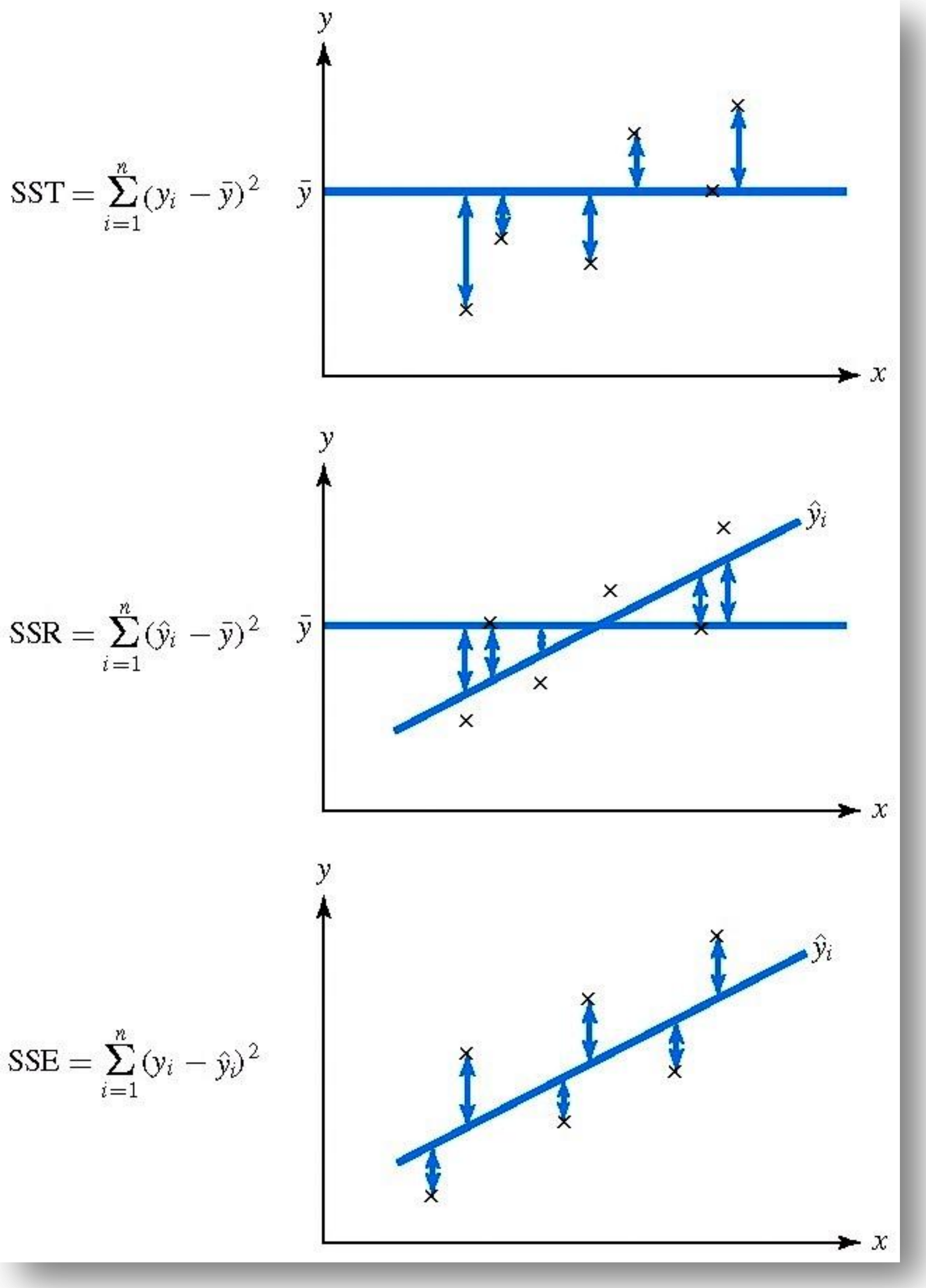
Sum of squares for error
SSE

SST = SSE + SSR

Symbolically

$$\sum(Y_i - \bar{Y})^2 = \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \bar{Y}\right)^2$$

In deterministic relationship SSE = 0 i.e. for a perfect fitting estimation line SST = SSR and hence SSR/SST=1. For the worst case of data SSR = 0 i.e. SSE = SST and hence SSR/SST=0.


$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
$\bar{y}$
y
x
$\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
$\bar{y}$
$\hat{y}_i$
y
x
$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$\hat{y}_i$
y
x

So the ratio SSR/SST evaluate how good the estimated regression line is, values of this ratio closer to 1 would imply better fitting estimated line. Thus the ratio SSR/SST is known as the coefficient of determination.

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum(Y_i - \bar{Y})^2}$$

r^2 is a non-negative value and it's limit are $0 \leq r^2 \leq 1$. Verbally, r^2 measures the percentage of the total variation in the dependent variable explained by the regression model.

Example: The following table shows the hardness (X) and tensile strength (Y) of 5 samples of metal:

X	146	152	158	164	170
Y	75	78	77	89	82

Find the regression line Y on X. Is this linear model adequate for the given data set, Justify your result?

Difference between Regression and Correlation Analysis:

i. In regression analysis, there is an asymmetry in the way the dependent and independent variables are treated.

ii. Regression analysis provides us the overall measure of the extent to which the variation in one variable determines the variation in the other.

Multiple Regression Model

If there are more than one independent variable in the regression model, then it is called the multiple regression model. The fitted plane of Y is denoted as

$$\mathbf{\hat{Y} = b_1 X_1 + b_2 X_2 + a}$$

Using least squares method we get the error equation as

$$\sum e_i^2 = \sum \left(Y - \hat{Y}_i\right)^2 = \sum (Y - b_1X_1 - b_2X_2 - a)^2$$

Taking partial derivatives with respect to a, b_1 and b_2 we get

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$$

$$\sum X_1Y = a \sum X_1 + b_1 \sum {X_1}^2 + b_2 \sum X_1X_2$$

$$\sum X_2Y = a \sum X_2 + b_1 \sum X_1X_2 + b_2 \sum {X_2}^2$$

After simplifying the above equations we get

$$b_1 = \frac{SS(X_2)\ SP(X_1Y) - SP(X_1X_2)\ SP(X_2Y)}{SS(X_1)\ SS(X_2) - \{SP(X_1X_2)\}^2}$$

$$b_2 = \frac{SS(X_1)\ SP(X_2Y) - SP(X_1X_2)\ SP(X_1Y)}{SS(X_1)\ SS(X_2) - \{SP(X_1X_2)\}^2}$$

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

Here

$SS(X_i) = \sum {X_i}^2 - \frac{(\sum X_i)^2}{n}$ and $SP(X_iY) = \sum X_iY - \frac{\sum X_i \sum Y}{n}$

Example: A researcher is interested in predicting the average value of rice production (Y) in a field, on the basis of two predictor variables, the average rainfall per day (X_1) and average urea used in per square feet (X_2). Data for 10 individuals were recorded as in table below:

Sl. no.	1	2	3	4	5	6	7	8	9	10
X_1	24	0	25	0	5	18	20	0	15	6
X_2	53	47	50	52	40	44	46	45	56	40
Y	11	22	7	26	22	15	9	23	15	24

Estimate the regression of Y on X_1 and X_2.

Solution:

Sl. no.	${X_1}^2$	${X_2}^2$	X_1Y	X_1X_2	X_2Y
1	576	2809	264	1272	583
2	0	2209	0	0	1034
3	625	2500	175	1250	350
4	0	2704	0	0	1352
5	25	1600	110	200	880
6	324	1936	270	792	660
7	400	2116	180	920	414
8	0	2025	0	0	1035
9	225	3136	225	840	840
10	36	1600	144	240	960
	$\sum {X_1}^2$=2211	$\sum {X_2}^2$=22635	$\sum X_1Y$=1368	$\sum X_1X_2$=5514	$\sum X_2Y$=8108

$$SS(X_1) = \sum X_1{}^2 - \frac{(\sum X_1)^2}{n} = 2211 - \frac{(113)^2}{10} = 934.1$$

$$SS(X_2) = \sum X_2{}^2 - \frac{(\sum X_2)^2}{n} = 22635 - \frac{(473)^2}{10} = 262.1$$

$$SP(X_1Y) = \sum X_1Y - \frac{\sum X_1 \sum Y}{n} = 1368 - \frac{113 \times 174}{10} = -598.2$$

$$SP(X_2Y) = \sum X_2Y - \frac{\sum X_2 \sum Y}{n} = 8108 - \frac{473 \times 174}{10} = -122.2$$

$$SP(X_1X_2) = \sum X_1X_2 - \frac{\sum X_1 \sum X_2}{n} = 5514 - \frac{113 \times 473}{10} = 169.1$$

$$b_1 = \frac{SS(X_2)\ SP(X_1Y) - SP(X_1X_2)\ SP(X_2Y)}{SS(X_1)\ SS(X_2) - \{SP(X_1X_2)\}^2}$$

$$= \frac{262.1 \times -598.2 - 169.1 \times -122.2}{934.1 \times 262.1 - \{169.1\}^2} = -0.630$$

$$b_2 = \frac{SS(X_1)\ SP(X_2Y) - SP(X_1X_2)\ SP(X_1Y)}{SS(X_1)\ SS(X_2) - \{SP(X_1X_2)\}^2}$$

$$= \frac{934.1 \times -122.2 - 169.1 \times -598.2}{934.1 \times 262.1 - \{169.1\}^2} = -0.060$$

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 = 17.4 + 0.630 \times 11.3 + 0.060 \times 47.3 = 27.357$$

Required regression plane of Y is

$$\widehat{Y} = b_1X_1 + b_2X_2 + a = -0.630\,X_1 - 0.060\,X_2 + 27.357$$

Polynomial Regression Model

$$\mathbf{\widehat{Y} = a + b_1X + b_2X^2}$$

Treat X as X_1 and X^2 as X_2 in multiple regression model.

Using least squares method

$$\sum Y = na + b_1 \sum X + b_2 \sum X^2$$

$$\sum XY = a \sum X + b_1 \sum X^2 + b_2 \sum X^3$$

$$\sum X^2Y = a\sum X^2 + b_1\sum X^3 + b_2\sum X^4$$

We get

$$b_1 = \frac{SS(X^2)\ SP(XY) - SP(X\,X^2)\ SP(X^2Y)}{SS(X)\ SS(X^2) - \{SP(X\,X^2)\}^2}$$

$$b_2 = \frac{SS(X)\ SP(X^2Y) - SP(X\,X^2)\ SP(X\,Y)}{SS(X)\ SS(X^2) - \{SP(X\,X^2)\}^2}$$

$$a = \overline{Y} - b_1\overline{X} - b_2\overline{X^2}$$

Example: A test was made to different doses of nitrogen (X) on rice field for observing rice production. The following data were recorded:

Nitrogen Dose	0	1	2	3	4	5
Rice Production	15	25	40	55	52	43

Compute a second degree polynomial to the data.

Solution:

N-Dose (X)	R-Production (Y)	X^2	$(X^2)^2$	$X\,X^2$	$X\,Y$	X^2Y
0	15	0	0	0	0	0
1	25	1	1	1	25	25
2	40	4	16	8	80	160
3	55	9	81	27	165	495
4	52	16	256	64	208	832
5	43	25	625	125	215	1075
15	230	55	975	225	693	2587

$SS(X) = 17.5$; $SS(X^2) = 474.83$; $SP(X\,X^2) = 87.50$; $SP(X^2\,Y) = 478.67$; $SP(X\,Y) = 118.00$

$$b_1 = 21.65 \quad b_2 = -2.98 \quad a = 11.53$$

The estimated polynomial regression is

$$\widehat{Y} = 11.53 + 21.65\,X - 2.98\,X^2$$