Correlation & Regression Analysis

Correlation

: It measures the relationship between the variables. It may be either +ve or -ve.

A correlation is said to be +ve if values of two variables changes in same direction.
Eg:- Bandwidth & data rate, demand & supply, height & weight.

A correlation is said to be -ve if values of two variable changes in opposite direction.
Eg: Datascience & efficiency of computer.

Methods of measuring correlation.

(i) Scatter diagram
(ii) Karl Pearson's correlation coefficient.
(iii) Rank correlation coefficient

(i) Karl Pearson's correlation coefficient.

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

$$[-1 \leq r \leq 1]$$

if $r > 0$, then there is +ve condition.
if $r < 0$, then there is -ve condition.
if $r = 0$, then there is no relation.

* The following information represents experience (in year) and performance score of officer.

Experience	Score
5	43
8	50
12	60
15	80
3	40
7	49
10	54

Compute correlation coefficient between experience & score and also interpret the result.

Sol^{n}

let x = experience , y = score

x	y	xy	x^2	y^2
5	43	215	25	1849
8	50	400	64	2500
12	60	720	144	3600
15	80	1200	225	6400
3	40	120	9	1600
7	49	343	49	2401
10	54	540	100	2916
$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$
60	376	3538	586	21266

Now,

$$r = r_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{7 \times 3538 - 60 \times 376}{\sqrt{7 \times 586 - (60)^2}\sqrt{7 \times 21266 - (376)^2}}$$

$$= 0.9$$

which is $r > 0$. So, there is +ve correlation.

Rank correlation (Spearman)

$$r = 1 - \frac{6\Sigma d^2}{(n^3 - n)}$$

① The information provided below represents the ranking of 6 different brands of laptop is given by 2 students

Brand	A	B	C	D	E	F
1st student	4	5	6	1	3	2
2nd student	3	2	4	5	6	1

Compute rank correlation coefficient (Apply appropriate statistical tool to determine whether brand preference is correlated.

Here,

Given that:-

Brand	1st	2nd	d	d^2
A	4	3	1	1
B	5	2	3	9
C	6	4	2	4
D	1	5	-4	16

E	3	6	-3	9
F	2	1	1	1

$\Sigma d^2 = 46$

Now,

$$r = 1 - \frac{6\Sigma d^2}{(n^3 - n)}$$

$$= 1 - \frac{6 \times 4}{(6^3 - 6)}$$

$$= -0.14$$

② Following data set represents price and demand of product

Price	40	28	60	30	35	25
Demand	20	28	10	25	27	30

Compute rank correlation coefficient.

Here,

Since rank is not given so, ranking the given in ascending order,

Price (x)	Demand (y)	R_1	R_2	d	d^2
40	20	5	2	3	9
28	28	2	4	-2	4
60	10	6	1	5	25
30	25	3	3	0	0
35	27	4	5	-1	1
25	30	1	6	-5	25

$\Sigma D^2 = 64$

Now

$$r = 1 - \frac{6 \times 64}{(6^3 - 6)}$$

$\frac{1+2+3}{3}$

$= -0.82$

Since, $r < 0$, so it is -ve correlation.

3. From the following information compute rank correlation coefficient.

x	50	75	35	50	45	55	55	80
y	40	70	39	60	60	65	70	77

Here,

Since rank is not given so, ranking the given observation in ascending order.

x	y	R_1	R_2	d	d^2
50	40	3.5	2	1.5	2.25
75	70	7	6.5	0.5	0.25
35	39	1	1	0	0
50	60	3.5	3.5	0	0
45	60	2.	3.5	-1.5	2.25
55	65	5.5	5	0.5	0.25
55	70	5.5	6.5	-1	1
80	77	8	8	0	0

$\Sigma d^2 = 6$

$t_1 = 2$

$t_2 = 2$

$t_3 = 2$

$t_4 = 2$

Now,

$$r = 1 - 6\left[\frac{\Sigma d^2 + \frac{(t_1^3 - t_1)}{12} + \frac{(t_2^3 - t_2)}{12} + \frac{(t_3^3 - t_3)}{12} + \frac{(t_4^3 - t_4)}{12}}{(n^3 - n)}\right]$$

$= 0.86$

There is +ve correlation

✓ # Regression Analysis

→ Dependent variable (unknown variable / The variable whose value is to be predicted)
→ Independent variable (known variable / given variable)
→ ~~The~~

The main objective of Regression Analysis is to predict the value of dependent variable with the help of independent variable.

Regression equation

1. Regression equation y on x.

$$y = a + bx \quad \text{—①}$$

where y = Dependent variable
x = Independent variable
a = constant
b = Regression coefficient

where,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Now, putting the value of a & b in eq² ①, we get

$$\hat{y} = a + bx$$

2 Regression equation x on y.

$$x = a + by \quad \text{— (1)}$$

where,

$x \rightarrow$ Dependent variable

$y \rightarrow$ Independent variable

where,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

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

Now, putting the value of a & b in eq^n (1) we get,

$$\hat{x} = a + by$$

* From the following information obtained the two regression equations.

X	Y
10	20
15	30
20	42
8	10
18	39
30	52

a) Here,

Regression equations Y on X :-

$$y = a + bx \quad \text{— (1)}$$

where,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

x	y	xy	x^2	y^2
10	20	200	100	400
15	30	450	225	900
20	42	840	400	1764
8	10	80	64	100
18	39	702	324	1521
30	52	1560	900	2704
$\Sigma x = 101$	$\Sigma y = 193$	$\Sigma xy = 3832$	$\Sigma x^2 = 2013$	$\Sigma y^2 = 7389$

Then,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{6 \times 3832 - 101 \times 193}{6 \times 2013 - (101)^2}$$

$$= 1.86$$

Now,

$$a = \bar{y} - b \cdot \bar{x}$$

$$= \frac{\Sigma y}{n} - 1.86 \times \frac{\Sigma x}{n}$$

$$= \frac{193}{6} - 1.86 \times \frac{101}{6}$$

$$= 0.85$$

Now, placing the value of a & b in eq^n ①. we get.

$$\hat{y} = 0.85 + 1.86x$$

b) Regression eq^{n} X on Y.

Let the regression eq^{n} be :-

$$X = a + by \quad \text{— (1)}$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{6 \times 3832 - 101 \times 193}{6 \times 7389 - (193)^2}$$

$$= 0.49$$

Again,

$$a = \bar{x} - b \cdot \bar{y}$$

$$= \frac{\Sigma x}{n} - 0.49 \times \frac{\Sigma y}{N}$$

$$= \frac{101}{6} - 0.49 \times \frac{193}{6}$$

$$= 1.07$$

Now, putting the value of a & b in eq^{n} (1) we get,

$$\hat{x} = 1.07 + 0.49y$$

Properties of regression coefficient

i) Regression coefficient are not symmetric.
ie. $b_{yx} \neq b_{xy}$.

ii) Both regression coefficient must have same sign.

iii) If one of the regression coefficient is

r and b have same sign.

greater than 1. then another must be less than 1.

iv) $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$

b → y on x = b_{yx}
b → x on y = b_{xy}.

① Following data set represents score obtained by the student which is affected by number of absent days of student.

Abs. days	Marks
5	40
8	30
1	48
3	45
10	27
12	24
7	31
6	20

a) Identify which one is the response variable (dependent variable).
b) Obtain a simple regression equation to describe given variable.
c) Interpret the meaning of estimated regression coefficient.
d) Determine the score when number of absent days is 4.

Score ?
Abs. days = 4 so, score dependent variable.

a) Since, score is affected by number of absent days. so,
ans: Dependent variable = score
Independent variable = absent days.

b)
ans: Here,
let, y = score and x = absent days.
Let, the regression equation Y on X be:-

$$y = a + bx \quad \text{— (1)}$$

where,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

X	Y	XY	X^2
5	40	200	25
8	30	240	64
1	48	48	1
3	45	135	9
10	27	270	100
12	24	288	144
7	31	217	49
6	20	120	36

$\Sigma x = 52$ $\Sigma y = 265$ $\Sigma xy = 1518$ $\Sigma x^2 = 428$

Then,

$$b = \frac{8 \times 1518 - 52 \times 265}{8 \times [illegible] - (52)^2}$$

$$= -2.27$$

$r = +0.98$ = strength

$b = 1.24$ = rate of change when 1 unit of X change then by 1.24 it is increased on Y.

$$a = \bar{Y} - b \cdot \bar{X}$$

$$= \frac{\Sigma Y}{N} - b \cdot \frac{\Sigma X}{N}$$

$$= \frac{265}{8} + 2.27 \times \frac{52}{8}$$

$$= 47.88$$

Now, placing the value of a and b in eq^{n} (1). we get.

$$\hat{Y} = 47.88 - 2.27X$$

c)

ans: It means $b = -2.27$, if we increase absent days by 1 unit then score is decreases by 2.27 unit.

d)

ans: Here,

When $X = 4$, then

$$\hat{Y} = 47.88 - 2.27 \times 4$$

$$= 38.8$$

Coefficient of determination (r^2)

It measures percentage of variation on dependent variable is explained by independent variable Suppose, $r = 0.7$ then $r^2 = 0.49$ ie. 49%

There is 49% of total variation on dependent variable is explain by independent variable.

Standard error of the estimate

It measures average deviation of observation from the fitted regression line.

$$Se = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

2076. Examination.

2) A study was done to study the effect of temperature on electri power consumed by a chemical plant following table gives the data which are collected from an experimental pilot plant.

Temperature (°F)	27	45	72	58	31	60	34	74
Electric power	250	285	320	295	265	298	267	321

i) Identify which one is response variable, and fit a simple regression line, assuming the relationship between them is lineo

Sol:

(y)

~~let~~. power consumption = dependent (response

temperature = independent. (x)

Let, the regression equation y on x be:-

$$y = a + bx \quad \text{— (1)}$$

Where,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

X	Y	XY	X^2
27	250	6750	729
45	285	12825	2025
72	320	23040	5184
58	295	17110	3364
31	265	8215	961
60	298	17880	3600
34	267	9078	1156
74	321	23754	5476

$\Sigma x = 401$ $\Sigma y = 2301$ $\Sigma xy = 118652$ $\Sigma x^2 = 22495$

Then,

$$b = \frac{8 \times 118652 - 401 \times 2301}{8 \times 22495 - (401)^2}$$

$$= 1.38$$

Now,

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

$$= \frac{2301}{8} - 1.38 \times \frac{401}{8}$$

$$= 218.4525$$

Now,

Eqn (1) becomes :-

$$\hat{y} = 218.45 + 1.38x.$$

b. Interpret the regression coefficient with reference to problem.

ans: $b = 1.30$ ie. if we increase temp by 1 unit then electric power consumption is increases by 1.30 unit.

c, Obtain coefficient of determination, and interpret this

ans: Here,

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{8 \times 118652 - 401 \times 2301}{\sqrt{8 \times 22495 - (401)^2}\sqrt{8 \times 666509 - (2301)^2}}$$

$$= 0.98$$

Then,

$$r^2 = (0.98)^2$$

$$= 0.9604$$

ie. 96.04 % of total variation on electric power consumption is explained by temperature

d) Estimate the standard error

ans: Here,

$$Se = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

$$= \sqrt{\frac{666509 - 218.45 \times 2301 - 1.30 \times 118652}{8-2}}$$

$$= 40.01$$

Average deviation of observation from the fitted deviation line is 46.01.

e, Based on fitted model in a) predict power consumption for
ans: Here, an ambient temperature of 65°F

When $x = 65$,

$$\hat{y} = 218.45 + 1.30 \times 65$$
$$= 302.95$$

* Remarks

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

2078 Examination.

2) Write the properties of correlation coefficient The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmitted time was 0.04 seconds with the standard deviation 0.01 seconds. The correlation coefficient between the time and size was 0.86. Based on these data, fit a linear regression model and predict the time it take Sol? to transmit a 400Kbytes file.

Let, the regression ~~coefficient~~ equation y on x is:-

$$y = a + bx \quad \text{— (1)}$$

Then,

Time → dependent (y)
File size → Independent (x)

$n = 30$
$\bar{x} = 126$
$\sigma_x = 35$
$\bar{y} = 0.04$
$\sigma_y = 0.01$
$r = 0.86$

Now,

$$b = r \frac{\sigma_y}{\sigma_x}$$
$$= 0.86 \times \frac{0.01}{35}$$
$$= 0.00025$$

$$a = \bar{y} - b\bar{x}$$
$$= 0.04 - 0.00025 \times 126$$
$$= 0.0085$$

Now, putting the value of a and b, we get,

$$\hat{y} = 0.0085 + 0.00025x$$

When,

$x = 400$
$\hat{y} = 0.0085 + 0.00025 \times 400$
$= 0.10$ second.

The properties of correlation coefficient :-

i) It always lies between -1 to 1. ie. $-1 \leq r \leq 1$
ii) It is independent of unit.
iii) It is symmetric.
ie. $r_{yx} = r_{xy}$
iv) $r = \pm \sqrt{b_{yx} \cdot b_{xy}}$
v) It is independent of change of origin and scale