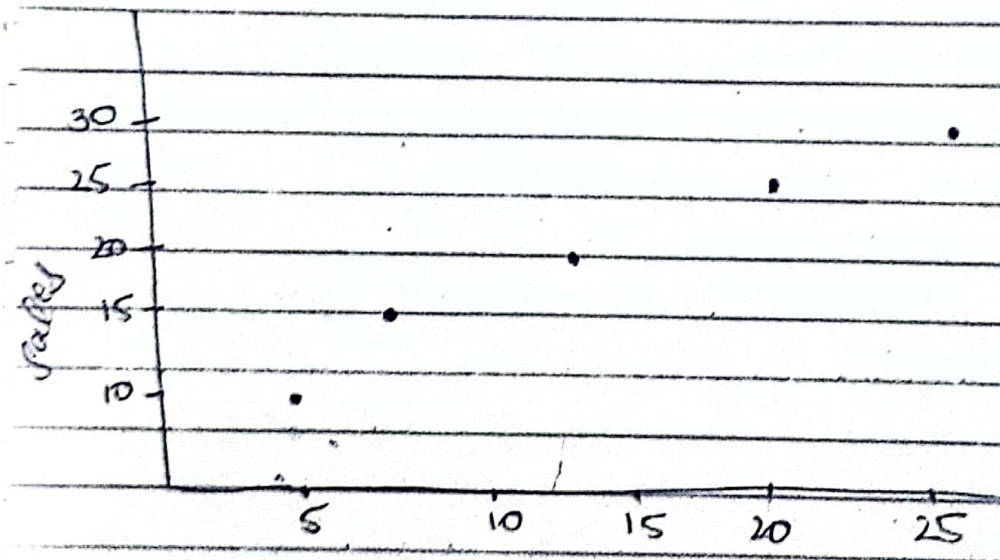
Correlation Analysis

i) Scatter Diagram & it's uses
ii) Covariance & its limitations
iii) Pearson's coefficient of correlation
iv) Interpretation of (ii) & (iii)

Ex:

Sales (X) ($)	Adv. Exp (Y) ($)	XY	X^2	Y^2
10	5	50	100	25
15	7	105	225	49
20	12	240	400	144
25	20	500	625	400
30	25	750	900	625
100	69	1645	1243	2250



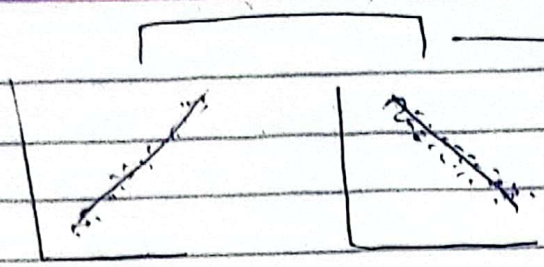
$$Cov(x, y) = S_{xy}$$

$$S_{xy} = \frac{1}{n}\left[\Sigma XY - \frac{\Sigma X \Sigma Y}{n}\right]$$

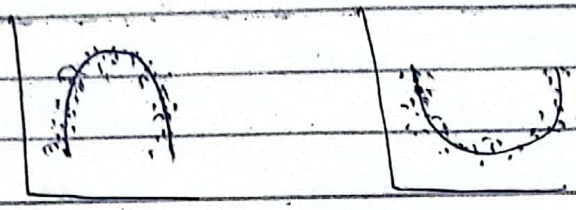
$$= \frac{1}{5}\left[1645 - \frac{(69)(100)}{5}\right]$$

$$= 53 \ (\$^2)$$

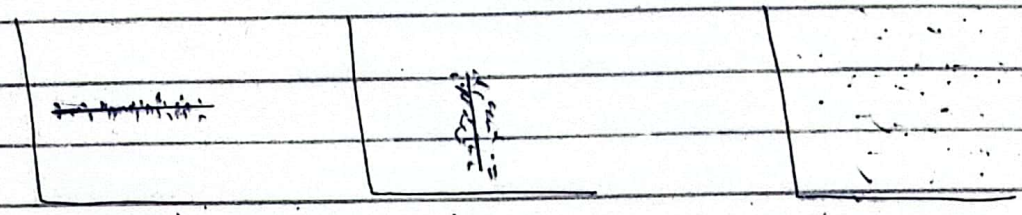
Scanned with CamScanner



→ Linear Relationship



→ Non-Linear Relationship



↓

No Relation

→ Covariance :

$\rightarrow cov(x,y) = S_{xy}$

$S_{xy} = \frac{1}{n}\left[\Sigma xy - \frac{\Sigma x \Sigma y}{n}\right]$

(i) $cov(x,y) = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{n}$

(ii) $cov(x,y) = \frac{1}{n}\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]$

- If $S_{xy} = 0$ (NO Linear Relationship)
- $S_{xy} < 0$ (Negative Relationship)
- $S_{xy} > 0$ (Positive Relationship)
- $-\infty < S_{xy} < +\infty$ (Range)

Disadvantages.

- can not identify state (Moderate, Strong, Weak) etc

Scanned with CamScanner

•) we do not have an interpretated value.

→ Pearson Product Moment co-efficient of Correlation

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x) \cdot var(y)}}$$

$$var(x) = S_x^2$$

•) $cov(x,y) = S_{xy}$

•) $corr(x,y) = r_{xy}$

i) $S_x^2 = \frac{\Sigma(X-\bar{X})^2}{n}$

ii) $S_x^2 = \frac{1}{n}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]$

$$var(Y) = S_y^2$$

i) $S_y^2 = \frac{\Sigma(Y-\bar{Y})^2}{n}$

ii) $S_y^2 = \frac{1}{2}\left[\Sigma Y^2 - \frac{(\Sigma Y^2)}{n}\right]$

$$S_x^2 = \frac{1}{5}\left[1243 - \frac{(69)^2}{5}\right] = 38.16$$

$$S_y^2 = \frac{1}{5}\left[2250 - \frac{(100)^2}{5}\right] = 50$$

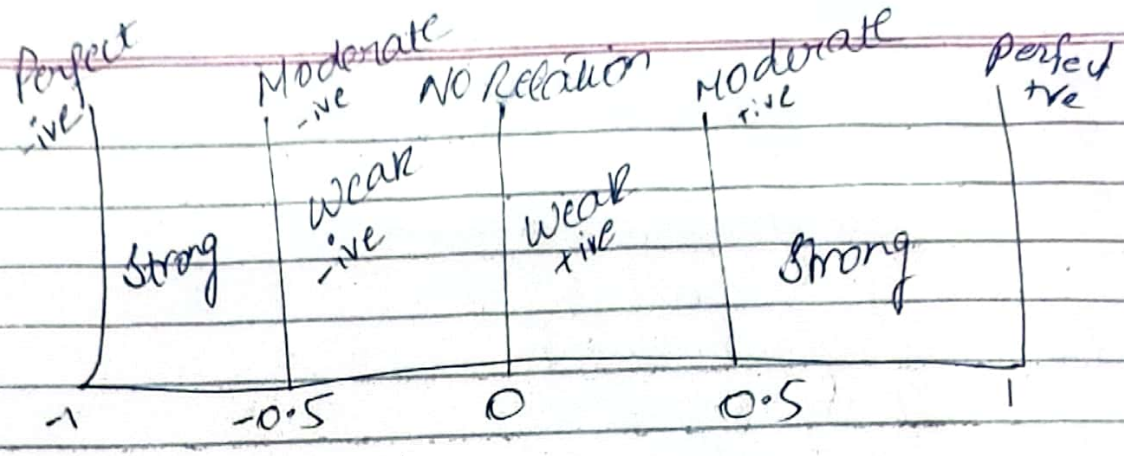
$$r_{xy} = \frac{53}{\sqrt{(38.16)(50)}}$$

$$= 0.98$$

→ Interpretation: There is a strong +ive correlation b/w Sales & Adv. Exp

→ Range : $-1 \leq r_{xy} \leq +1$

Scanned with CamScanner



→ R-code for correlation

```
x = c( )
y = c( )
plot(x, y)
corr(x, y)
```

<u>Using Calculator</u>

① Mode → Stat → A+Bx

x	y

Press AC

② Shift

Σx^2 Σxy

Σy^2 Σy

C_{xy}

Sum

r_{xy}

Scanned with CamScanner

1. Testing of Hypothesis about population co-efficient of correlation
2. Linear Regression Model
3. Simple Linear Regression Model (SRM)
4. Estimating the parameters of SRM by least square method
5. Interpretation of estimating regression model

Correlation Analysis :-

- Pearson's Product Moment correlation co-efficient
- Spearman's Rank Correlation
- Kendall's co-efficient of correlation

1. Two Tailed

i) $H_0 : \rho = 0$ (No Linear Relationship)

$H_1 : \rho \neq 0$ (Linear Relationship)

$\alpha = 0.05$

Test - Statistic :

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Critical Region:

$$|t| \geq t_{\alpha/2}(n-2)$$

Conclusion.

One tailed

ii) $H_0 : \rho = 0$

$H_1 : \rho > 0$ (+ive Relationship)

$\alpha = 0.05$

Scanned with CamScanner

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

CR

$$t > t_{\alpha(n-2)}$$

Conclusion

iii) One tailed.

$H_0 : \rho = 0$

$H_1 : \rho > 0$

$\alpha = 0.05$

Test statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

CR

$$t < -t_{\alpha(n-2)}$$

Conclusion

→ For Spearman's r will be replaced by r_s

→ For Kendall's r will be replaced by τ (Taa).

ii) Testing hypothesis about the population co-efficient of correlation equal to zero at 5% level of significance

$r = 0.6 \qquad n = 6.$

$H_0 : \rho = 0$

$H_1 : \rho \neq 0 \qquad \alpha = 0.05$

Scanned with CamScanner

Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{(0.6)\sqrt{6-2}}{\sqrt{1-(0.6)^2}}$$

$$= 1.5$$

CR

$$|t| \geq t_{\alpha/2}(n-2)$$

$$|t| \geq t_{0.02}(4)$$

Conclusion:

[Linear Regression Model]

↳ Regression Analysis

income consumption

~~any~~ consumption = f(income)

↳ a+b income

↳ a+b income

function can be form of anything and determined after plotting.

consumption = $\beta_0 + \beta_1$ Income

$Y = \beta_0 + \beta_1 X$

↳ Deterministic (Exact) Model

Scanned with CamScanner

parameters epsilon error term.

residual

$Y = \beta_0 + \beta_1 X_1 + \epsilon$

independent

Probabilistic Model

(inclusion of residual)

dependent

explanatory/predicted.

$Y = \beta_0 X^{\beta_1}$

$Y = e^{X\beta_1}$

$Y = \beta_0 \beta_1^{X}$

non-Linear Model (Exponential Function)/

The power of vario parameter is not 1)

Simple Linear Regression Model:

$Y = \beta_0 + \beta_1 X_1 + \epsilon$

only 1 independent variable.

Multiple Linear Regression Model:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

multiple independent variable.

Scanned with CamScanner

Mean of residual = Sum 0 = 0

consumption → Income

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{--- (1)}$$

Estimating the eq (1)

Difference b/w estimated & true values is less

β_0, β_1 → Method of Least Square.

$\hat{\beta}_0$ ✓ estimated value

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \cancel{n\Sigma XY =} \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

Estimated Simple Linear Regression Model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Ex:

Fit the simple linear Regression Model by Least Square method. For the following data

consumption	Income
8	7
15	13
25	20
29	25
28	31

$$\hat{Y} = -1.42 + 0.98X$$

Scanned with CamScanner

① [Mode] [Stat] [2-a+bx]

| x | y |

[2-a+bx]

② Press AC after entering all obs.

③ [Shift] [1] [Reg]

using calculator

[a] → β_0

[b] → β_1

i) Predict the consumption for a household when its income is 15

ii) $\hat{Y} = -1.42 + 0.98X$

$X = 15$

$\hat{Y} = -1.42 + 0.98(15)$

$\hat{Y} = 13.3$

. Goodness of fit ?

↓

forecast is good/not.

Scanned with CamScanner

Goodness of fit for the Regression Model

i) Standard error of estimate $= \frac{\Sigma(Y-\hat{Y})^2}{n-2}$

ii) Co-efficient of Determination (R^2)

a) $R^2 = \frac{RSS}{TSS}$

b) $R^2 = 1 - \frac{ESS}{TSS}$

$0 \le R^2 \le 1$

Ex:

Consider the simple linear regression model

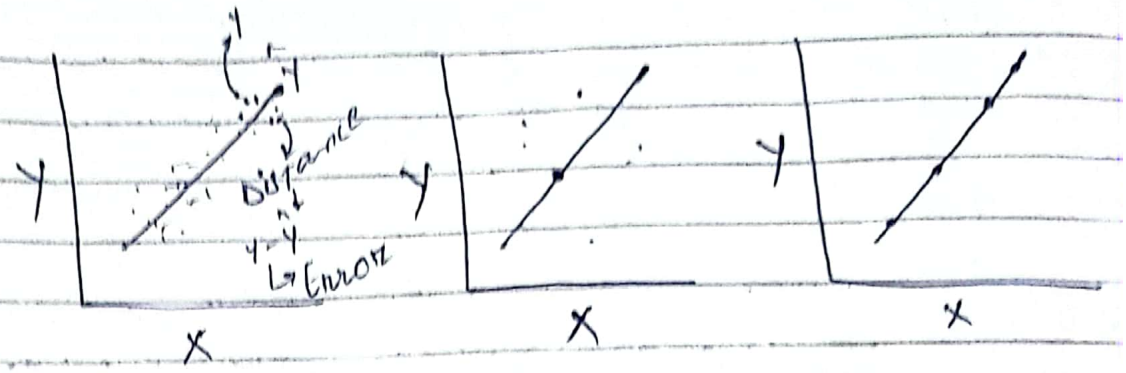
consumption $= \beta_0 + \beta_1$ income $+ \epsilon$

The estimated model is

$$\hat{Y} = -1.42 + 0.982X$$

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
7	8			
13	15			
20	25			
25	29			
31	28			

Scanned with CamScanner



Standard Deviation = 0 = Good fit

~~Standard Error $\propto \frac{1}{\infty}$~~

Standard Error $\propto \frac{1}{\text{co-efficient of determination}(R)}$

TSS	RSS	ESS
↓	↓	↓
Total sum of squares	Regression sum of square	Error sum of square

$$TSS = RSS + TSS$$

① $R^2 = \frac{RSS}{TSS}$

② $R^2 = 1 - \frac{ESS}{TSS}$

$0 \le R^2 \le 1$

i) $R^2 \ge 0.70$ (good fit)

Scanned with CamScanner

ii) R^2 b/w 0.5 - 0.65 (moderate fit)
iii) less than 0.5 (poor fit)

→ Violation of any assumption of ϵ can cause the value of R^2 to be greater, which concludes the model to be a good fit which is not correct.

Residual Analysis:

Assumptions of SLR model:

$$Y = \beta_0 + \beta_1 X + \epsilon_i$$

i) Model is linear
ii) ϵ terms ($\epsilon_i \epsilon_j$) are independent
iii) ϵ term have same variance
iv) ϵ and X are independent
v) ϵ follow normal distribution

When the above assumption violates
ii) $E(\epsilon_i \epsilon_j) \neq 0$ (Auto correlation)

↓

Relation within the values of one variable.

Scanned with CamScanner

Test for Auto-Correlation

i) Durbin-Watson Test

H_0: There is no auto correlation

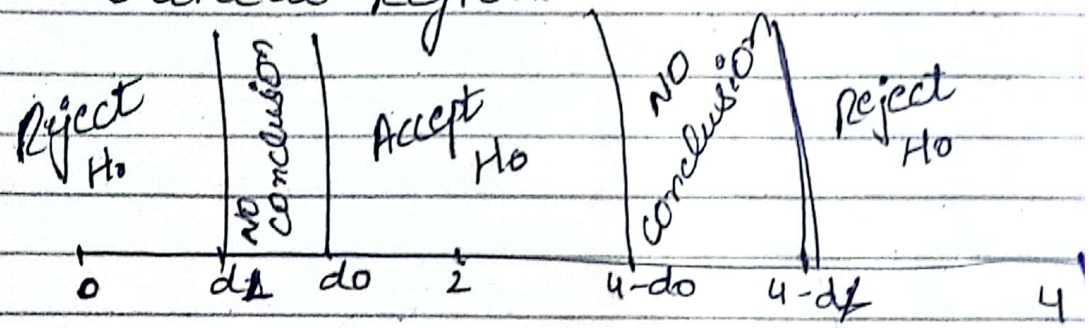
H_1: There is correlation

$\alpha = 0.05$

Test-Statistic:

$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$$

Critical Region:



conclusion.

Q.

For the following data test whether there is auto correlation or not at 5% level of significance.

Income X	Consumption Y	$\hat{Y} = 3.5454 + 0.91X$	$e_t = Y - \hat{Y}$	e_{t-1}
7	8	9.91	-1.91	-
13	15	15.36	-0.36	-1.91
20	25	21.73	3.27	-0.36
25	29	26.72	2.73	3.27
31	28	31.72	-3.72	2.73

Scanned with CamScanner

no. of values $n = 5$
no. of independent variable / variables $k = 1$) → d_L, d_U
$\alpha = 0.05$

$(e_t - e_{t-1})^2$	e_t^2
–	3.6481
5.1529	0.1296
13.177	10.6929
0.2916	7.4529
~~6.45~~	
41.6025	13.8384
60.224	35.7619

Durbin-Watson Test :

H_0 : ____

H_1 : ____

$\alpha = 0.05$

Test-Statistic :

$$d = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2} = \frac{60.224}{35.7619}$$

$$= 1.6841$$

CR ____

$n = 5$ (no. of values)

$k = 1$ (no. of independent variable)

$\alpha = 0.05$

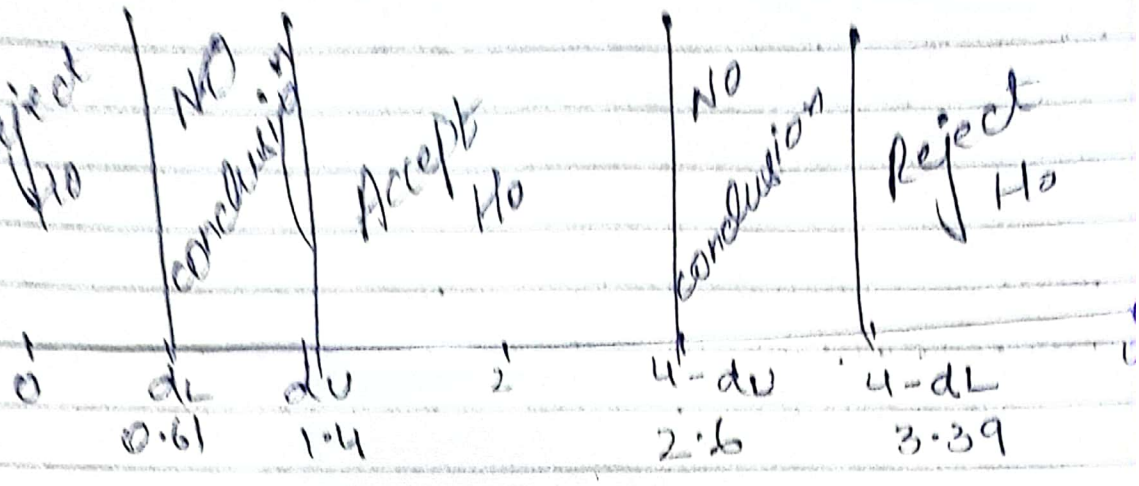
$d_L = 0.610$

$d_U = 1.400$

$4 - d_L = 4 - 0.610 = 3.39$

$4 - d_U = 4 - 1.400 = 2.6$

Scanned with CamScanner



Conclusion:

Accept Ho $\cancel{(1.4 \leq 1.6 \leq 2.6)}$

$(1.4 \leq 1.6 \leq 2.6)$

Scanned with CamScanner