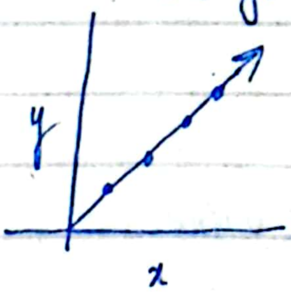
★ Scatterplots

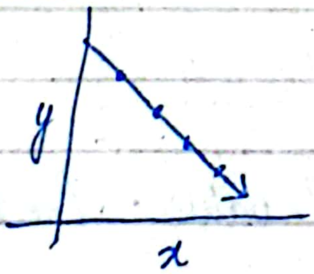
* b/w two numerical data to see relationship status

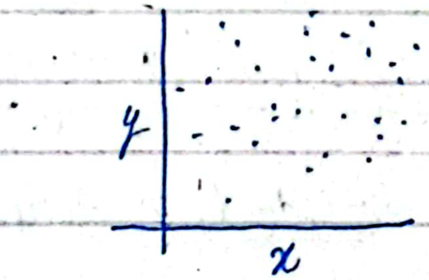
- Linearity (+ive, -ive) or non-linear scenario
- Strength of relationship (how close points (data) are present)
- Outlier
- Clusters (Distinguished regions of closely related datapoints)

★ Correlation Coefficient (Explain the relationship between two numerical variables)

Range = (+1 → -1)





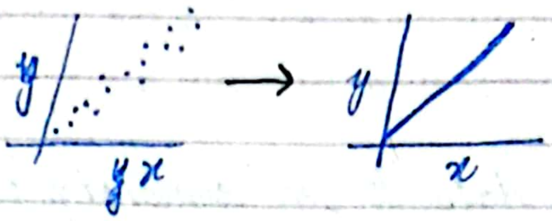


$r = +1$ $r = -1$ $r = 0$

- Calculation of "r" ($n \to$ data points)

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

★ Introduction to trend lines (line graph)



$$y = mx + c$$

→ Now from given data, we can predict future data by just manipulation the x any y values.

→ This whole concept is known as linear regression

- The line of best fit (!)

- Estimation with linear models ($y = mx + c$)

  ↳ It can be done through Eq as well as by simply extrapolatine the line.

★ Least Squares regression Eq.

residuals = Actual value − Predicted value

+ive value → The predicted value is less (Actual high)

−ive value → The predicted value is high (Actual below)

(The aim of regression is to minimise the sum of residuals)

⇒ "A residual is a measure of how well a line fits an individual data point"

- Calculating the Eq. of a regression line

For given data set we have ($\bar{x}, \bar{y}$ and s_x, s_y)
linear regression Eq. is

$$y = mx + c$$

Here,

$$m = r\frac{s_y}{s_x}$$

And,

$$\bar{y} = m\bar{x} + c$$

$$c = \bar{y} - m\bar{x}$$

Thus,

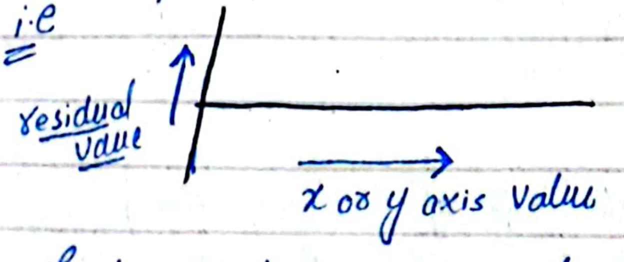
$$\hat{y} = mx + c \quad \text{—} \quad (1)$$

* Assessing the fit in least-squares regression.

- Residual plots

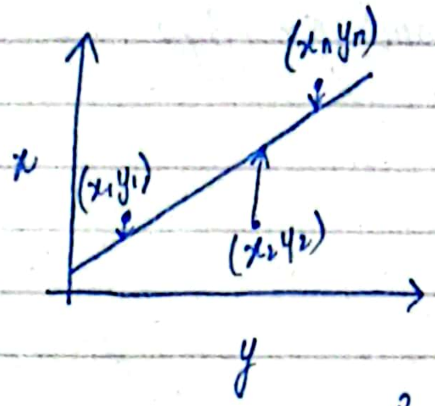
→ Calculate the residual value of each data point and then polot its for x and y axis values.

i.e



If the points on residual plots are evenly scattered then relation is authentic (linear) if not then some non-linear relation should be suggested.

r (Correlation coefficient) vs r^2 (coefficient of determination)



$$SE_{LINE} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 \cdots (y_n - (mx_n + b))^2$$

Q What %age of total variation in y is described by the variation in x?
to answere this:

$$\text{Total variation in } y = SE \text{ from } \bar{y} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_n - \bar{y})^2$$

Q: What %age of variation is not described by the variation in x or by the regression line

$$\frac{SE_{LINE}}{SE_{\bar{y}}}$$

$r^2 \Rightarrow 1 - \dfrac{SE_{LINE}}{SE_{\bar{y}}} \Rightarrow$ (What %age of total variation is described by the line or by the variation in x)

coefficient of determination

if SE_{LINE} → small → line is a good fit (r^2 value close to 1)

⇒ It will mean a lot of total variation in y is described by the variation in x (which is good in relation terms)

if SE_{LINE} → large (opposite to above)

- Std of residuals / Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{(residual_1)^2 + (r_2)^2 + (r_3)^3 \cdots (r_n)^2}{n-1}}$$

The lower this value is, the better fits the model !

- Impact of removing outliers on regression lines !

	(Positive)	(Negative)
r →	Increase (close to 1)	decrease
Slope →	Increase	decrease
r^2 →	Increase (close to 1)	~~decrease~~ increase
y-intercept →		

→ All depends on what is the location of outlier !

- SE on regression line ⇒ (Sum is done because otherwise +ive and -ive residuals with counter effect each other)

"The whole idea of perfect regression line is to minimize the sum of squared error"

The line with least error !

From $\underline{\underline{SE_{LINE}}}$ we get,

$$\boxed{m = \frac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}} \left[m = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \right]$$

$$\boxed{b = \bar{y} - m\bar{x}}$$

Covariance and regression line

$$Cov(x,y) = E[(x - E(x)) \cdot (y - E(y))]$$

i.e $x \rightarrow 1$ $y \rightarrow 2$

$E(x) \rightarrow 3$ $E(y) \rightarrow 1$

$$\boxed{E[E(x)] \Rightarrow E(x)}$$

$$Cov(x,y) = (1-3)(2-1)$$

$$\Rightarrow (-2)(1)$$

$$\Rightarrow -2$$

$$\Rightarrow Cov(x,y) = E[XY] - E[Y][E[X]]$$

$$Cov(x,y) = E[XY] - \bar{Y}\bar{X}$$

$$Cov(x,y) = \overline{XY} - \bar{X}\bar{Y}$$

$$m = \frac{Cov(X,Y)}{\overline{X^2} - (\bar{X})^2} = \frac{Cov(X,Y)}{\overline{XX} - \bar{X}\cdot\bar{X}} \Rightarrow \boxed{\frac{Cov(X,Y)}{Var(X)}}$$