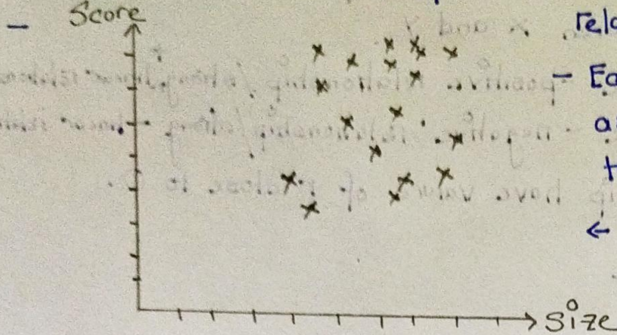
03.1 KHAN ACADEMY - STATISTICS AND PROBABILITY

03 - Exploring Bivariate numerical data

SCATTERPLOT AND CORRELATION REVIEW

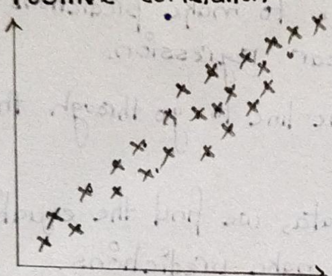
Scatterplot — A scatterplot is a type of data display that shows the relationship between two numerical variables.



- Each member of the dataset get plotted as a point whose (x,y) coordinates relates to its values for the two variable.
- ← For example, shoes quiz scores and shoes size for students in a class.

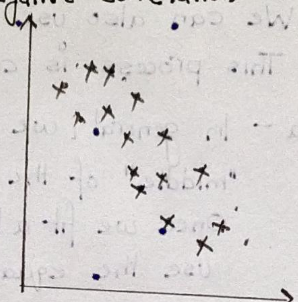
Correlation — We often see patterns or relationship in scatterplot.

i) Positive Correlation



When y variable tends to increase as the x variable increases, we say this as positive correlation.

ii) Negative Correlation

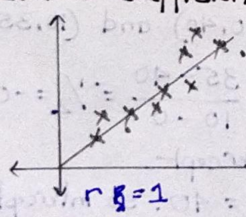


When y variable then to decrease as x variable increase is negative correlation.

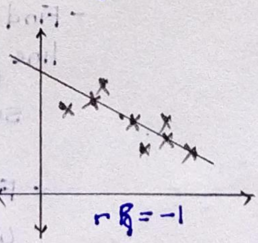
iii) No Correlation

When there is no clear relationship between two variables is no correlation.

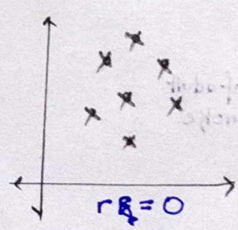
Correlation Coefficient(r) →



$r \approx 1$

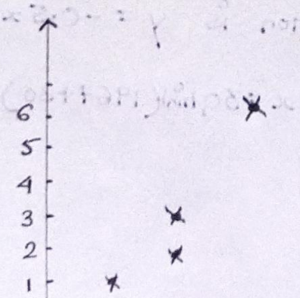


$r \approx -1$



$r \approx 0$

How to calculate correlation coefficient (r)?



points $(x,y) \to (1,1), (2,2), (2,3)\ (3,6)$

Mean of $\bar{x} = \frac{1+2+2+3}{4} = \frac{8}{2} = 4 \quad \bar{x} = 2, \ \bar{Y} = 3.$

SD of $\bar{x} = \sqrt{\frac{(1-2)^2+(2-2)^2+(2-2)^2+(3-2)^2}{2}} \Rightarrow \sigma_x = 0.81, \sigma_y = 2.1$

or $S_x = 0.81, S_y = 2.16$

$$\boxed{r = \frac{1}{n-1} \Sigma \left[\frac{x_i - \bar{x}}{S_x}\right]\left[\frac{y_i - \bar{Y}}{S_y}\right]}$$

Z score(x) . Z score(Y)

For all point, $r = \frac{1}{3}\left(\left[\left(\frac{1-2}{0.816}\right)\left(\frac{1-3}{2.160}\right)\right] + \left(\frac{2-2}{0.816}\right)\left(\frac{2-3}{2.160}\right) + \text{all pairs}\right)$

$$r = \frac{1}{3}\left(\left(\frac{1-2}{0.816}\right)\left(\frac{1-3}{2.160}\right) + \underbrace{\left(\frac{2-2}{0.816}\right)\left(\frac{2-3}{2.160}\right)}_{0} + \underbrace{\left(\frac{2-2}{0.816}\right)\left(\frac{3-3}{2.160}\right)}_{0} + \left(\frac{3-2}{0.816}\right)\left(\frac{6-3}{2.160}\right)\right)$$

$$r = \frac{1}{3}\left(\frac{2}{0.816 * 2.160} + \frac{3}{0.816 * 2.160}\right) = \frac{1}{3}\left(\frac{5}{0.816 * 2.160}\right) \cong 0.946$$

So $r = 0.946$, correlation coefficient is a measure of how well a line can describe the relationship between X and Y.

$-1 \leq r \leq 1$: If r is positive - positive relationship / strong + linear relationship
If r is negative - negative relationship / strong - linear relationship
Weak relationship have value of r close to 0.

TREND LINES / LINEAR REGRESSION

Fitting a line to data.

Linear Regression – When we see a relationship in a scatterplot, we can use a line to summarize the relationship in the data. We can also use that line to make predictions in data. This process is called linear regression.

Fitting a line to data – In general, we want the line to go through the "middle" of the points. Once we fit a line to data, we find the equation and use the equation to make predictions.

Example – The percent of adults who smoke, recorded every few year since 1967, suggest a negative linear association with no outliers. A line was fit to the data to model the relationship.

- Find the slope
  line goes through (0,40) and (10,35)
  so slope is $\frac{35-40}{10-0} = -\frac{1}{2} = -0.5$
- Find the y intercept.
  when $x = 0$, $y = 40$. so intercept $= 40$.
- Eqn of line $\Rightarrow y = mx + b$.
  so equation is $y = -0.5x + 40$

Suppose we want to estimate smokers in 1997. So $x = 30$, hint (1967+30)

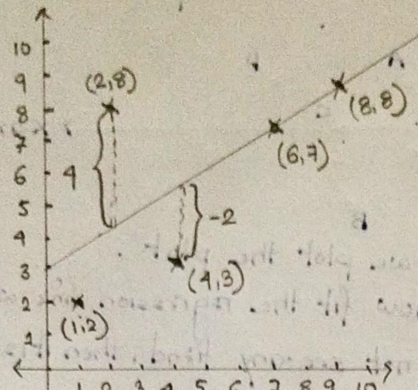
$y = -0.5x + 40$
$y = -(0.5)(30) + 40$
$y = -15 + 40$
$y = 25$.

Based on equation 25% of adults will smoke in 1997.

03.3 RESIDUALS

- We fit different line of fit, then how to decide which line fit the best?

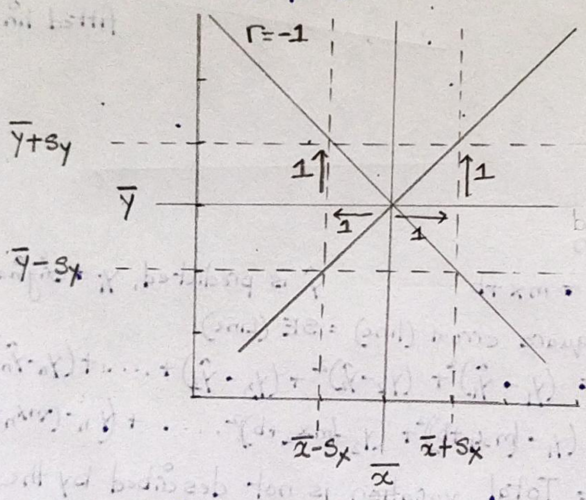
Residual → A residual is a measure of how well a line fits to an individual data point.



- Consider point (2,8) is 4 units above the line
- Vertical distance is known as a Residual. For data points above the line, residual is positive. and the data points below the line residual is negative.
- Consider point (4,3) is -2 units.
- Closer a data point residual is 0, better the fit.
- In this case (4,3) fits better than (2,8).

Calculating the equation of a regression line.

Suppose mean of $x = \bar{x}$, standard deviation of $x = Sx$.
mean of $y = \bar{y}$, standard deviation of $y = Sy$.



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

r is positive number, positive correlation
r is negative number, negative correlation

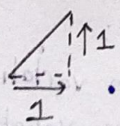
Equation of line, $\hat{y} = mx + b$.

$$\boxed{m = r\frac{Sy}{Sx}}$$

r= correlation, $\hat{y}$ = prediction
Sx, Sy = standard deviation.
$\bar{x}, \bar{y}$ = mean.

when $r=0$, m becomes 0, $\hat{y} = b$

So, when $r=1$,



→ $r\frac{Sy}{Sx}$ = when r=1, there should be unit change in x and y when we check what is the unit change then it is Sy and Sx.

when $r=-1$, there is a unit change, when found out, it is also Sy and Sx.

Suppose we have datapoints. Find the equation of line.

(x, y)
(1,1)
(2,2)
(2,3)
(3,6)

$\bar{x} = 2$, $Sx = 0.816$.
$\bar{y} = 3$, $Sy = 2.160$.

$r = 0.946$, $m = 0.946\left(\frac{2.160}{0.816}\right)$
$= 2.50$

Eqn of line, $y = mx + b$
for (2,3) $y = 3$, $x = 2$, $m = 2.50$

$3 = (2.50)(2) + b$
$3 = 5 + b$
$b = -2$.

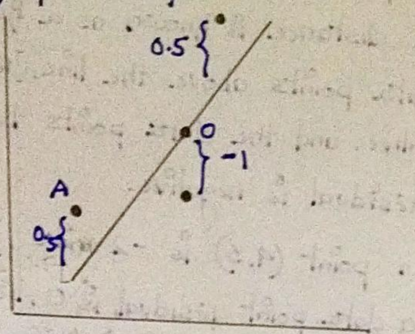
∴ eqn of line $\hat{y} = 2.50x - 2$.

RESIDUAL PLOT

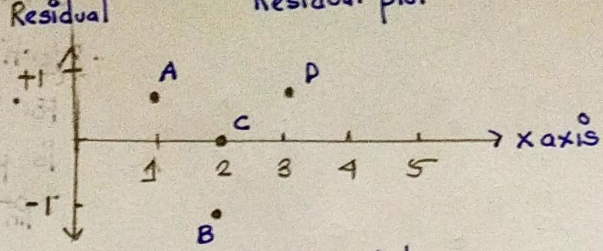
Residual = Actual − Expected

Continuing the last example → (x,y) A(1,1) B(2,2) C(2,3) D(3,6)

Eq^n of line, $\hat{y} = 2.50x + 2$ ..



Residual plot

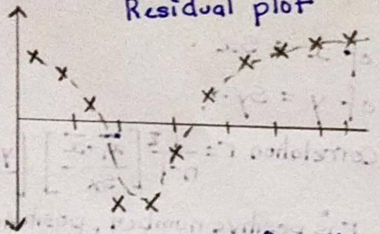


For each x, we plot the plot.

- It gives how fit the regression line was.
- If we do not see any trend, then it is good fit.

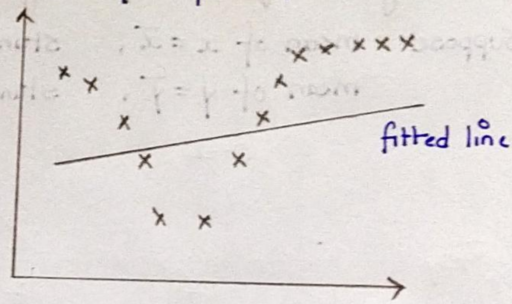
If there is some trend in the data, there may be issue.

Residual plot

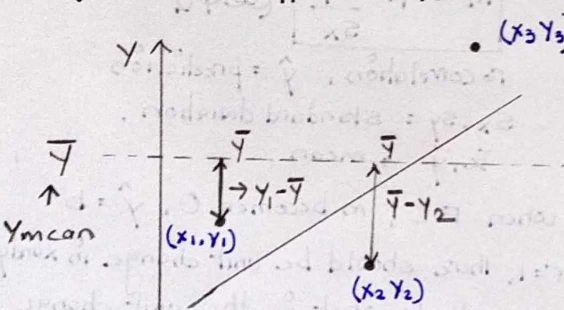


May be non linear fit should be there

Actual plot for reference



R square or Coefficient of Determination



$y = mx + b$

$\hat{y}$ is predicted, $y_i \rightarrow$ original

Square error (line) = SE (Line)

$= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3) + \dots + (y_n - \hat{y}_n)$

$= (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 \dots + (y_n - (mx_n + b))^2$

= Total variation is not described by the regression line.

So ~~how~~ how much variation (%) in y is described by the variation of x?

Total variation in $\bar{Y}$: $(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2 = SE_{(\bar{Y})}$ = Square Error $(\bar{Y})$

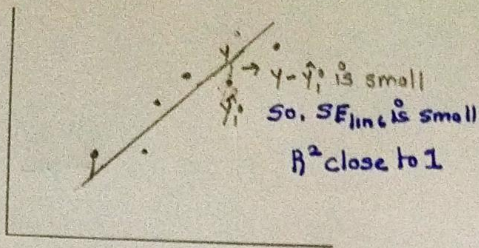
$$\frac{\text{Variation not describe by line}}{\text{Total variation}} = \frac{\text{Standard Error (line)}}{\text{Standard Error }(\bar{Y})}$$ = What % of total variation is not describe by the variation in X or by the regression line.

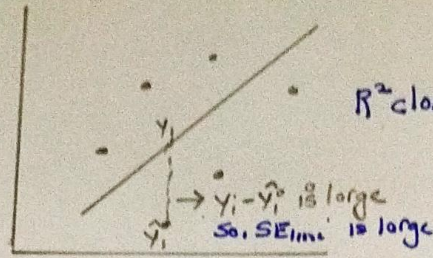
Variation explained by x = $\boxed{1 - \frac{SE_{Line}}{SE_{\bar{Y}}}}$ = What % of total variation is describe by line or variation of x.

This is known as R^2 or coefficient of Determination.

Q5

SE_{Line} is small → Line is a good fit. which is R^2 close to 1. So lots of variation in y is explained by x. And vice versa.

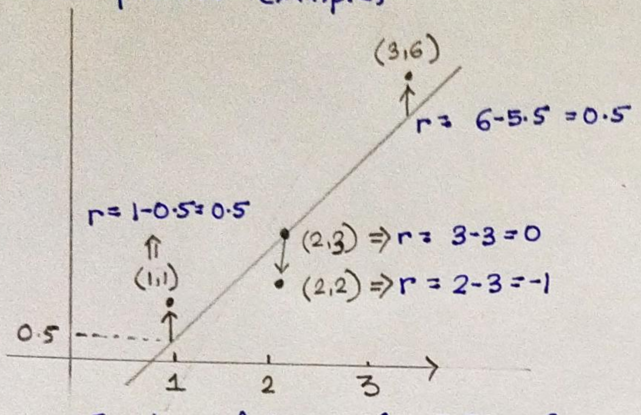




R^2 close to 0.

Standard deviation of residuals or Root mean square Error (RMSE)

Consider previous example,



Equation of line = $\hat{y} = 2.50x - 2$.

$r_i = y_i - \hat{y}$

for point (1,1). $y_i = 1$, $\hat{y}_i = 0.5$

$r(1,1) = 1 - 0.5 = 0.5$

$r(3,6)$ = use equation, y = 6

for $x = 3$, $\hat{y} = 2.50(3) - 2$
$= 7.5 - 2$
$= 5.5$.

$r(3,6) = 6 - 5.5 = 0.5$.

$r(2,3) \Rightarrow y_i = 3$, $\hat{y}_i = 3$.

$r(2,3) = 0$.

$r(2,2) \Rightarrow 2.50(2) - 2 = 5 - 2 = 3$.

$r(2,3) = 2 - 3 = -1$

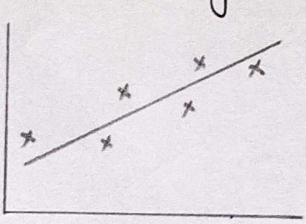
So, Standard Deviation of Residual =

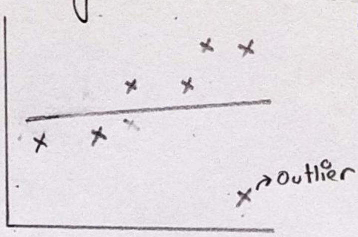
$$\sqrt{\frac{(0.5)^2 + (0)^2 + (-1)^2 + (0.5)^2}{n-1}} = \sqrt{\frac{0.25+1+0.25}{3}}$$

$$= \sqrt{\frac{1.5}{3}} = \sqrt{\frac{1}{2}} \cong 0.707$$

So, RMSE $\cong$ 0.707

Impact of removing outliers on regression line.

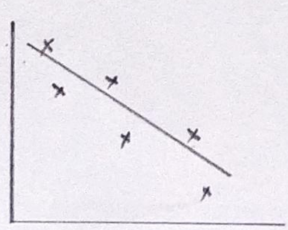


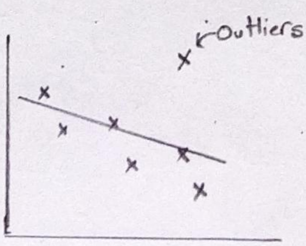


Both have similar data points but second we added outliers.

So to incorporate outlier it will tilt toward it.

- R^2 will increase if we remove outliers.
- Slope of line will increase if we remove outliers.
- Here 'r' = close to 1, correlation





- Similar data but second have outliers.
- Here slope will decrease if we remove outliers.
- 'r' goes close to -1, negative correlation