14-11-2020

Unit-5

Explore bivariate numerical data

⇒ Introduction to scatterplot

Using scatterplot to access bivariate relationship.

- Discription of scatterplot.

i) Form ii) Direction iii) Strength iv) Outliers

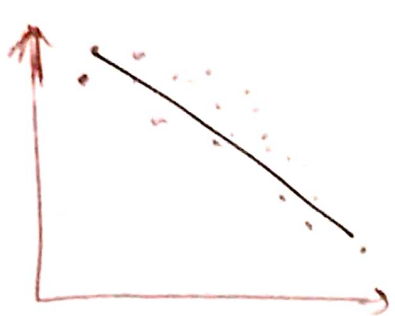
i) Form - Linear / non-Linear ?

ii) Direction - +ve / -ve

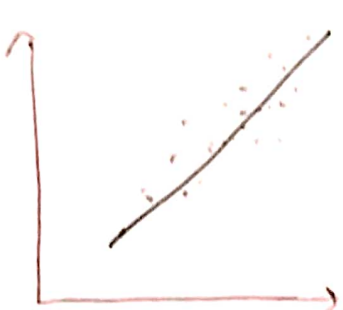
iii) Strength - Strong / moderately strong / weak

iv) Outliers → outliers or not ?

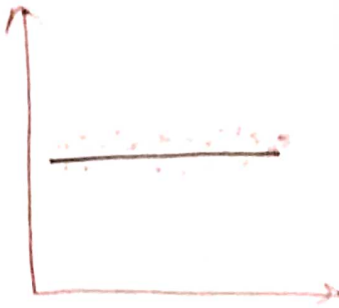
Correlation coefficient



$r \approx -1$



$r \approx +1$



$r \approx 0$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

we are given set of (x_i, y_i), hence we can calculate $\bar{x}, s_x$ and $\bar{y}, s_y$.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$ (Avg of product of z-scores)

$$-1 \leq r \leq +1$$

• An intuition about how r works :

• If for a given (x_i, y_i), 2 sets of both are of same sign. (+,+) or (-,-) Then ~~the~~ more relationship between them is +ve and they moves r toward +1.

• Else if 2 sets of (x_i, y_i) are of different sign, (+,-) or (-,+), then the product will be -ve and they move r towards -1 and is relationship is -ve.

Defination :

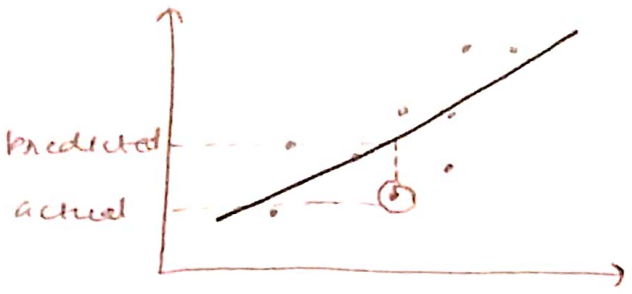
The correlation coefficient measures the direction and strength of a linear relationship.

⇒ Introduction to trend lines ⇒

Linear Regression :

When we see a relationship in a scatterplot we can use a line to summarize the data. We can also use that line to prediction. This process is called linear Regression.

⇒ Least square regression equation :



Residual = ~~predicted~~
Actual - predicted .

if line is $y = mx + c$,
for a point (x_i, y_i)

$residual_i = y_i - (mx_i + c)$

least square regression :

estimated line is choosen such that it minimizes $\sum (y_i - mx_i + c)^2$ | $m = \frac{s_y}{s_x} \cdot r$

⇒ Assessing the fit in least squares regression

⇒ Residual plot

→ should be random (without any trend)

⇒ R-Squared

r^2 → how much prediction error is eliminated when we used least square

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$= 1 -$ % of variance not explained
(proportion)

$R^2 \rightarrow$ coefficient of determination

⇒ Root mean Squared Error (RMSE)

OR

Standard deviation of residuals.

- Calculate the residuals and note their distribution
- Calculate their standard deviation.

This can also be written as :-

$$RMSE = \sqrt{\frac{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2}{n-1}}$$