CORRELATION & SIMPLE LINEAR REGRESSION

Dr. Aanchal A. Awasthi

A Thought Full Starter.....

S.No.	Statements
1	Is there a relationship between the height of a father and the height of a son
2	A relationship between anxiety leading to depression
3	Relationship between length of hospital stay and pain score
4	How age influences bone density in osteoporosis patients
5	or simply relationship between hours of study and exam score

Correlation & Regression Analysis

Content

- Concept of correlation
- Measures of correlation
- Interpretation of correlation coefficient
- Limitations of Correlation Analysis
- Concept and definition of simple linear regression
- Interepretation of Intercept and coefficient
- R Square
- Summary

Correlation

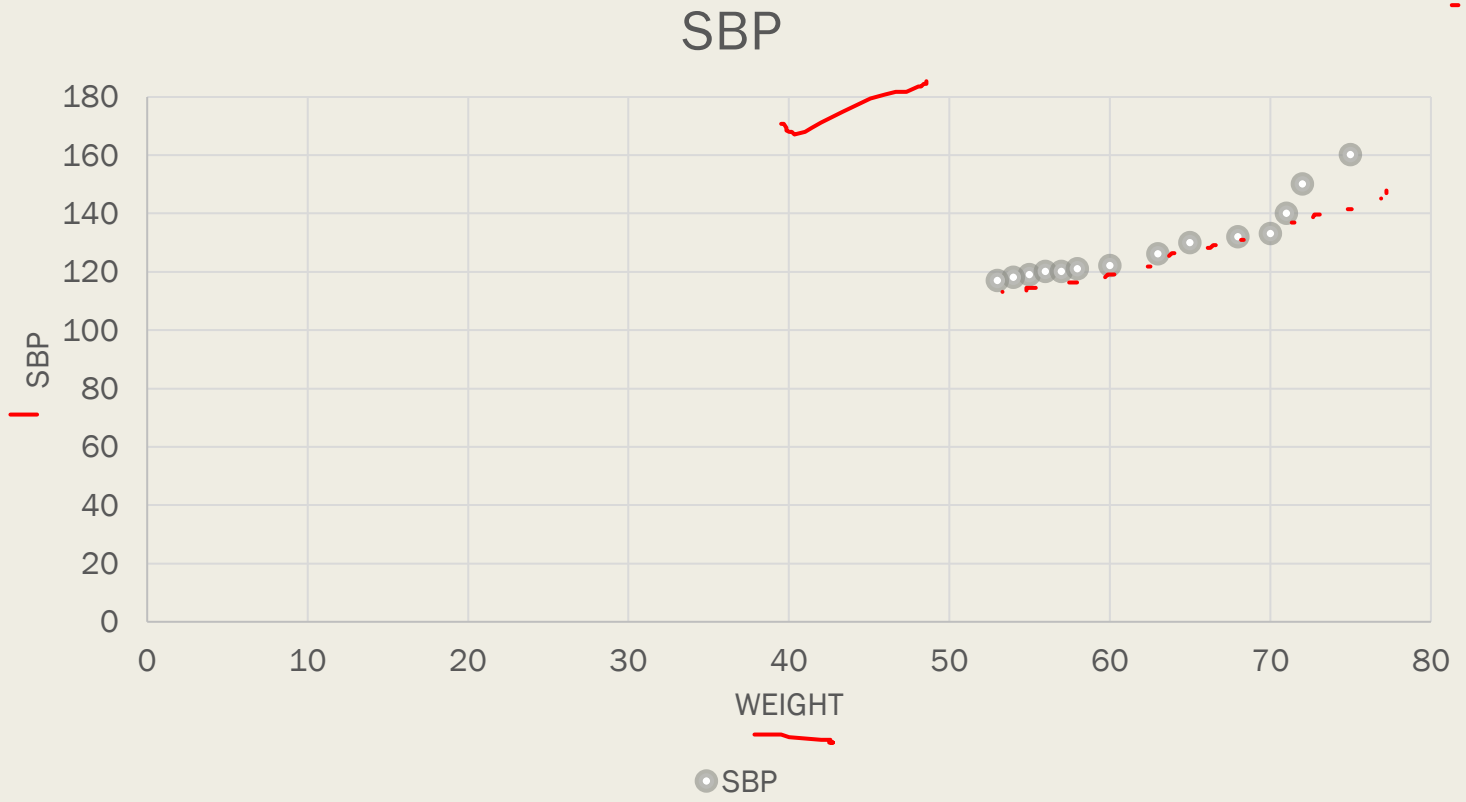
- Finding the relationship between two variables without being able to infer causal relationships

- Correlation is a statistical technique used to determine the degree to which two variables are related

Scatter diagram

- Rectangular coordinate
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined

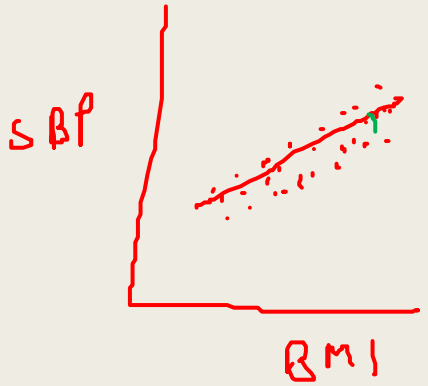
Figure 01: Scatterplot for representing relationship between weight & Systolic Blood Pressure



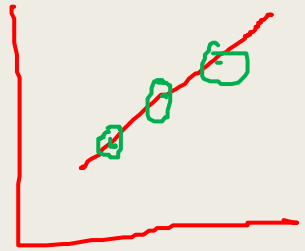
Pattern of Correlation

- Positive Relationship
- Negative Relationship
- Perfectly Positive Relationship
- Perfectly Negative Relationship
- No Linear Relationship

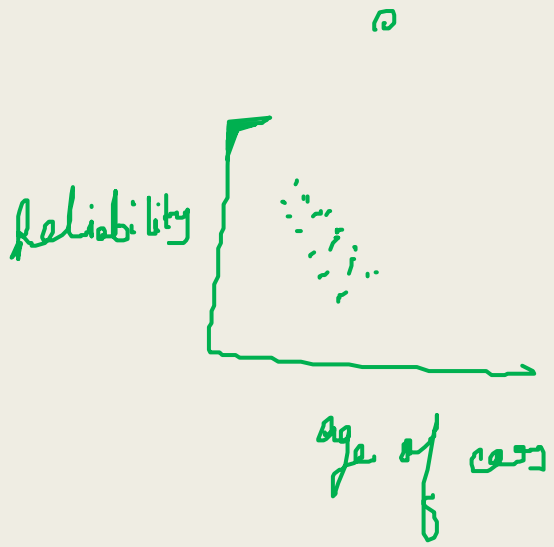
Positive Relationship



Perfectly Positive Relationship



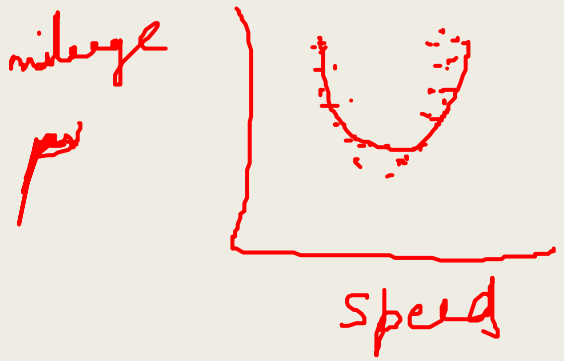
Negative Relationship



Perfectly Negative Relationship

No Relationship

No Linear Relationship



Measures of Correlation Coefficient

- Karl Pearson's correlation coefficient
- Spearman's Correlation Coefficient

Simple Correlation coefficient (r)

Pearson's *correlation coefficient or* ***product moment*** *correlation coefficient*

- It measures the **nature** and **strength** between two variables of the **quantitative** type

Assumption of Pearson Correlation Coefficient

- X and Y are linearly related — Scatter plot
- X & Y are assumed to come from a Normal Population (Sampling from “Bi-variate Normal Population)

Components of Correlation Coefficient

- The sign of r denotes the nature of association

  - *If the sign is +ve this means the relation is direct*
  - *While if the sign is -ve this means an inverse or indirect relationship*

- The value of r denotes the strength of association.

Range of Correlation Coefficient

Magnitude of 'r'	Interpretation
0	No Linear Relationship
0-0.25	Weak
0.25-0.75	Intermediate
0.75-1	Strong

Table 01: Data set 01

Weight	SBP
55	119
54	118
53	117
56	120
57	120
58	121
60	122
63	126
65	130
68	132
70	133
71	140
72	150
75	160

	Weight	SBP
Weight	1	
SBP	+0.925568	1

Strong Positive Correlation

$Y = a + b X$

Example 2

- Relationship between depression scores and test Scores

r = - 0.71 ???

0.2 – 0.7

moderate Negative Correl

Negative Intermediate correlation

Spearman Rank Correlation Coefficient (r_s)

Height (cm.) – quantitative

- It is a non-parametric measure of correlation

- This procedure makes use of the two sets of ranks that may be assigned to the sample values of X and Y

<u>Interpretation</u>

- Same interpretation as simple r.

Limitations of Correlation Analysis

Correlation does not imply causality

- *Two variables might be associated because they share a common cause.*
- *For example, NEET PG scores and College Grade(marks) are highly associated, but probably not because scoring well on the NEET causes a student to get high grades in college*
- *Study hours would be the common cause of the NEET PG scores and the grades*

Simple Linear Regression

Introduction

- How “values in Y” change as a function of changes in “value of X”
- Use one independent variable at a time to predict Outcome Variable “*Y*”

Height of son

Outcome
Dependent
y

Height of father

Independent variable
x

Definition

Regression Analysis is a **mathematical measure of the average relationship** between two or more variables in terms of the original units of the data.

$y = \alpha + \beta X$

Y = a + bX

Where a = the Y intercept

b = the slope / coefficient

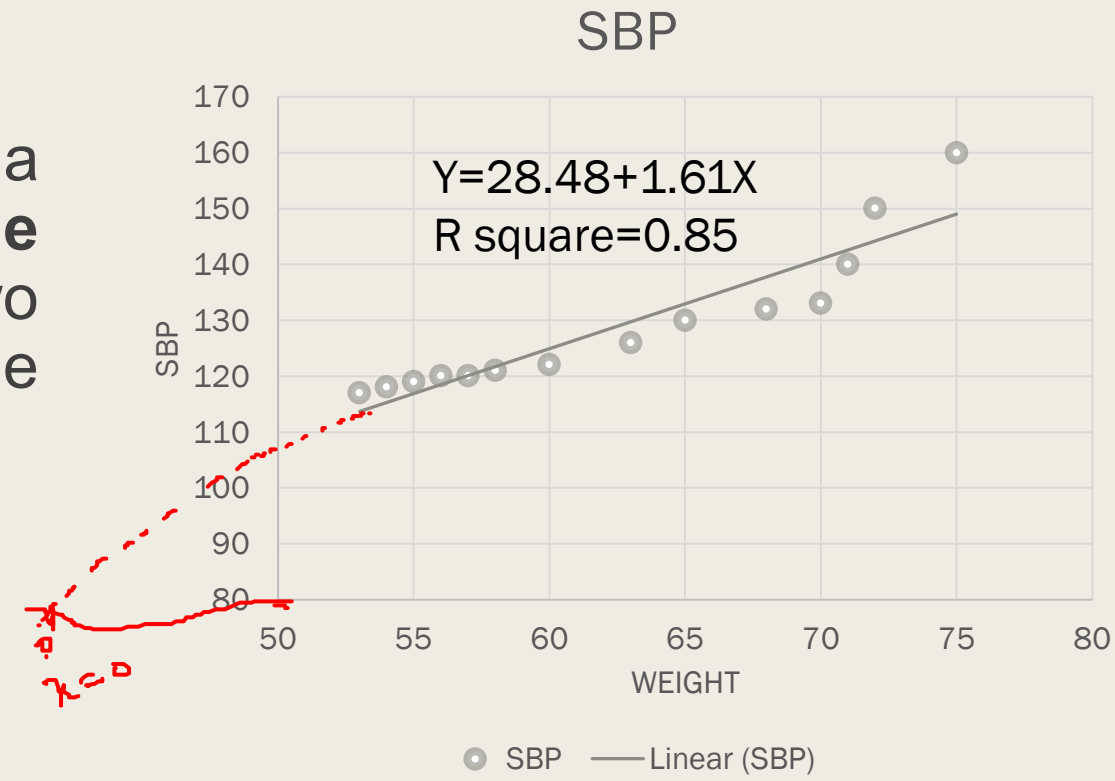


Figure 01 : Demonstration of Residuals

Line of Regression

- Line of “Best Fit”
- Obtained by Principal of “Least Squares”
- Line of best fit is obtained by Minimizing the sum of squares of residuals.
- Residual = Observed value – Predicted Value

  $= Y - \hat{Y}$

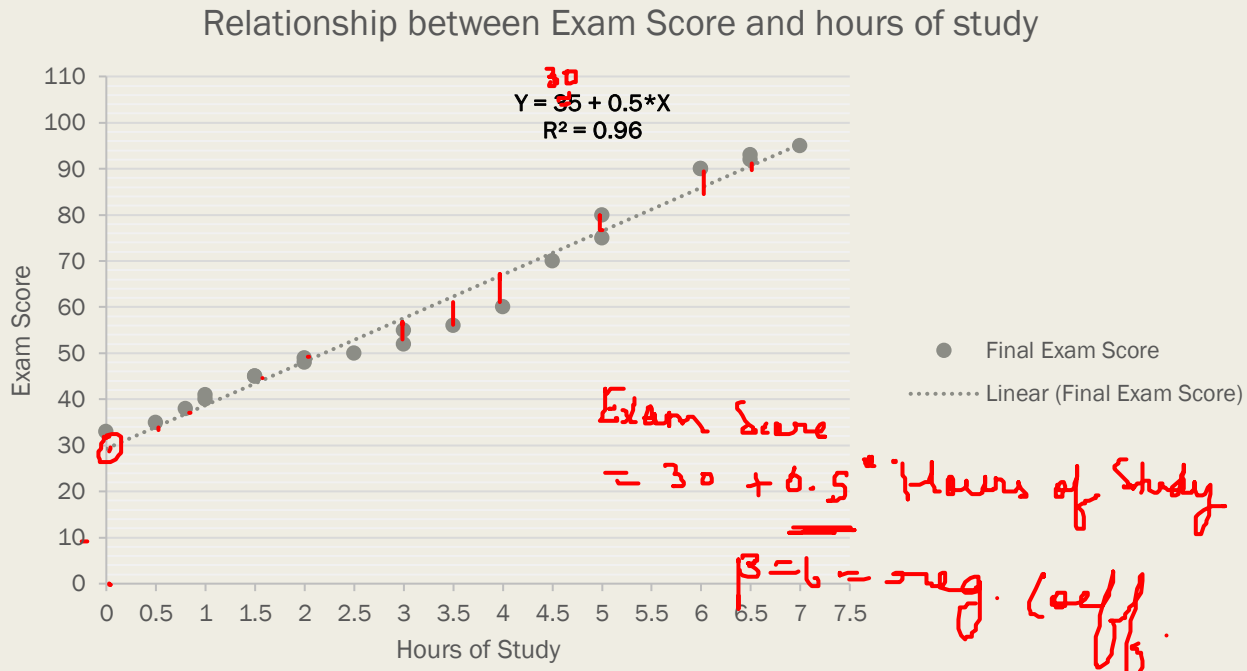
Two Components of Linear Regression

Coefficient

- Also known as "Slope"
- Theoretically, $-\infty \leq b \leq +\infty$ any value
- It tells us average change in Y for one unit change in X

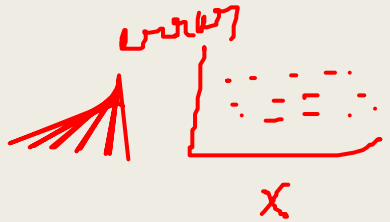
Y intercept

- Minimum Value of Y
- Value of Y at X=0



Assumptions of Simple Linear Regression

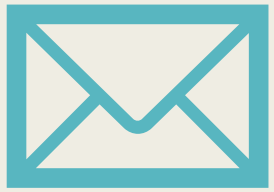
Linearity	Independence	Homoscedasticity	Normality
The relationship between the independent variable (X) and the dependent variable (Y) is linear. This means that the change in Y is directly proportional to the change in X.	The observations are independent of each other, meaning that the value of one observation does not depend on the value of another observation.	The variance of the errors (residuals) is constant across all levels of X. In other words, the spread of the residuals should be consistent as the values of X change.	The residuals are normally distributed. This means that the distribution of the residuals follows a normal distribution, with a mean of zero.



Summary of Correlation and Simple Linear Regression

Correlation	Regression
It tells us the existence of linear relationship, if any	It tells us average mathematical relationship between two variables
No concept of defining dependent and independent variables	It's mandatory to define dependent and independent variables
It tells us the direction and strength of relationship between two variables	Regression tells us change in outcome (Y) variable for one unit change in predictor (X) variable.
Correlation between two variables can be spurious	Considering domain related causality criteria, it helps in establishing cause and effect relationship
Correlation can be visualize using scatter plots	Relationship can be establish using principal of least square
Example: Correlation between height of father and height of son is 0.85	Height of Son = 145+0.5* Height of Father

Lets Connect!



draanchalawasthi@gmail.com

https://www.youtube.com/c/sscrindia

+91 750.625.0403