Introduction to Regression

- Mathematical and Statistical Equation
- Meaning of Intercept and Slope
- Error term
- Measure for Model Fit – R2,MAE, MAPE

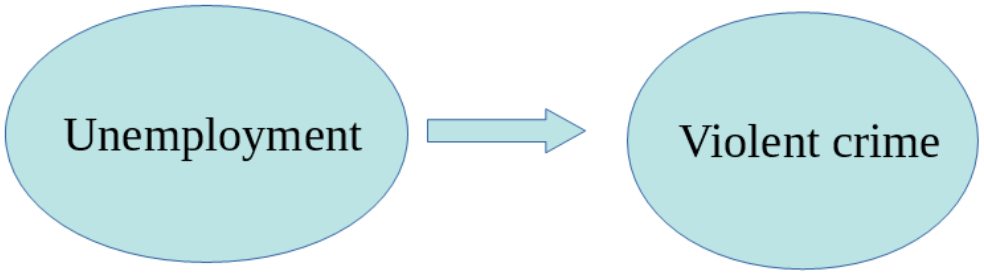
Correlation between values

- In many situations researchers and decision_x0002_makers need to consider the relationship between two or more variables

For example, the sales manager of a company may observe that the sales are not the same for each month. He/she also knows that the companyís advertising expenditure varies from year to year. This manager would be interested in knowing whether a relationship exists between sales and advertising expenditure. If the manager could successfully define the relationship, he/she might use this result to do a better job of planning and to improve predictions of yearly sales with the help of the regression technique for his/her company.

The regression

- The correlation problem considers the joint variation of two measurements neither of which is restricted by the experimenter.
- The regression problem considers the frequency distribution of one variable (dependent variable) when another variable (independent variable) is held fixed at each of several intervals



Regression

- Regression is a flexible model that allows you to “explain” or “predict” a given outcome (Y), variously called your outcome, response or dependent variable, as a function of a number of what is variously called inputs, features or independent, explanatory, or predictive variables (X1, X2, X3, etc.).

CORRELATION

- If two variables, say x and y vary or move together in the same or in the opposite directions they are said to be correlated or associated.
- Thus,correlation refers to the relationship between the variables.

- Generally, we find the relationship in certain types of variables.

Example, a relationship exists between

- income and expenditure,
- absenteesim and production,
- advertisement expenses and sales etc.

- Existence of the type of relationship may be different from one set of variables to another set of variables.

Scatter Diagram to show correlation

- When different sets of data are plotted on a graph, we obtain scatter diagrams.

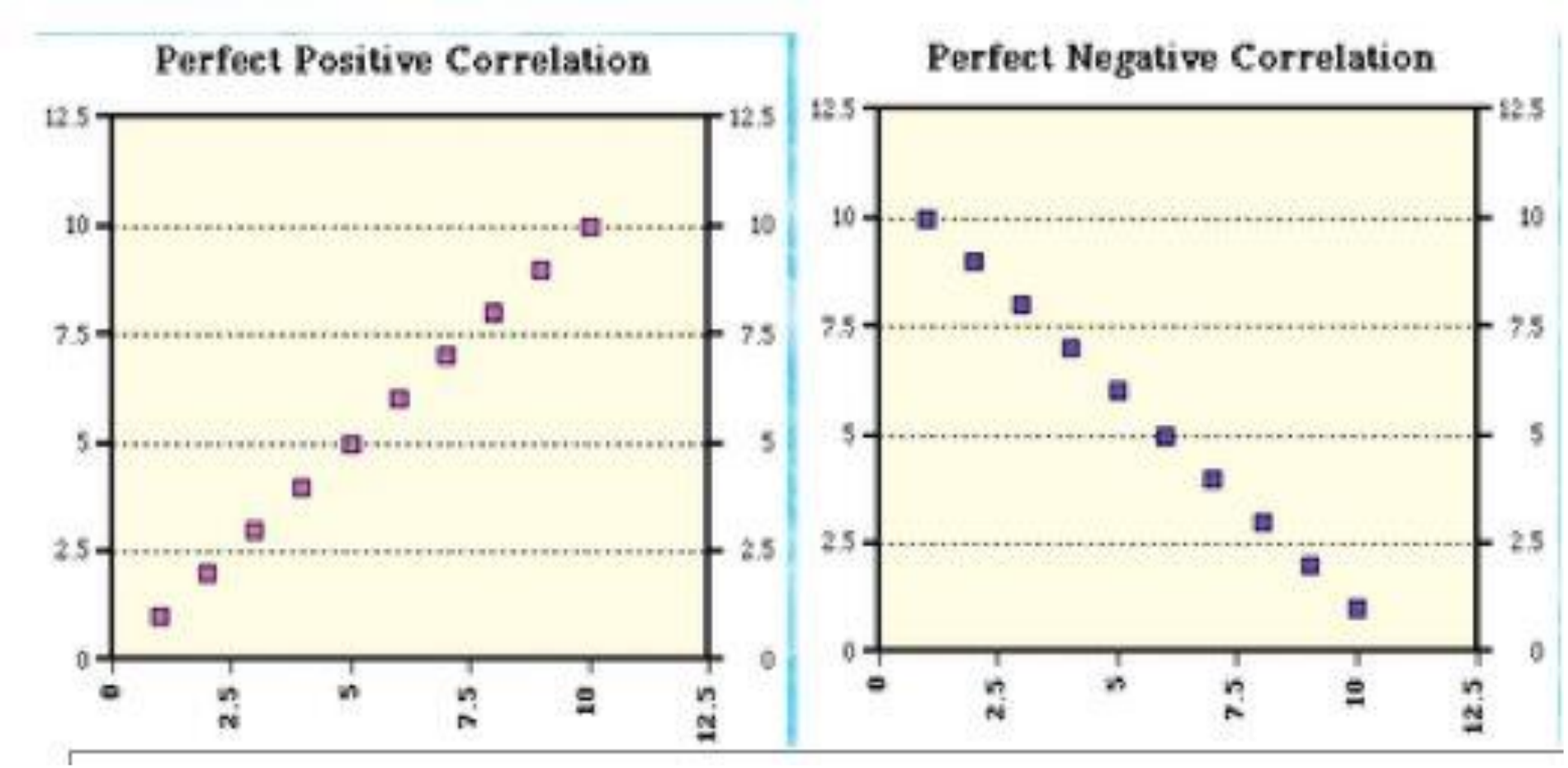
- A scatter diagram gives two very useful types of information.

(1) can observe patterns between variables that indicate whether the variables are related.

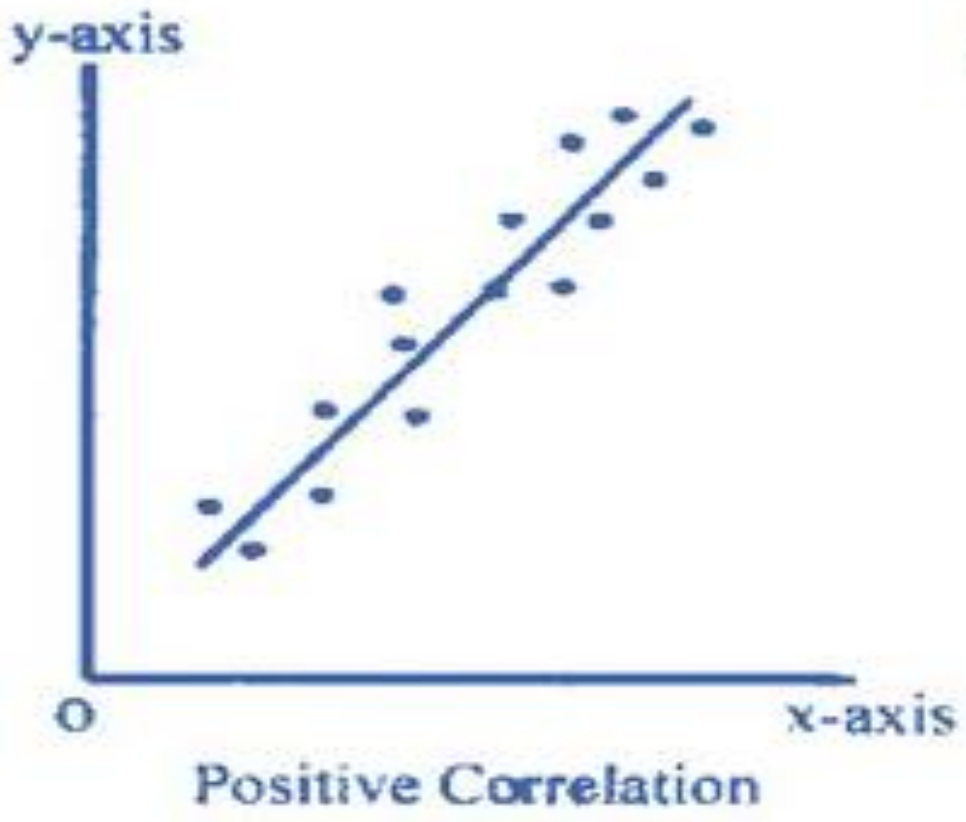
(2) if the variables are related we can get an idea of the type of relationship that exists.

The scatter diagram may exhibit different types of relationships.

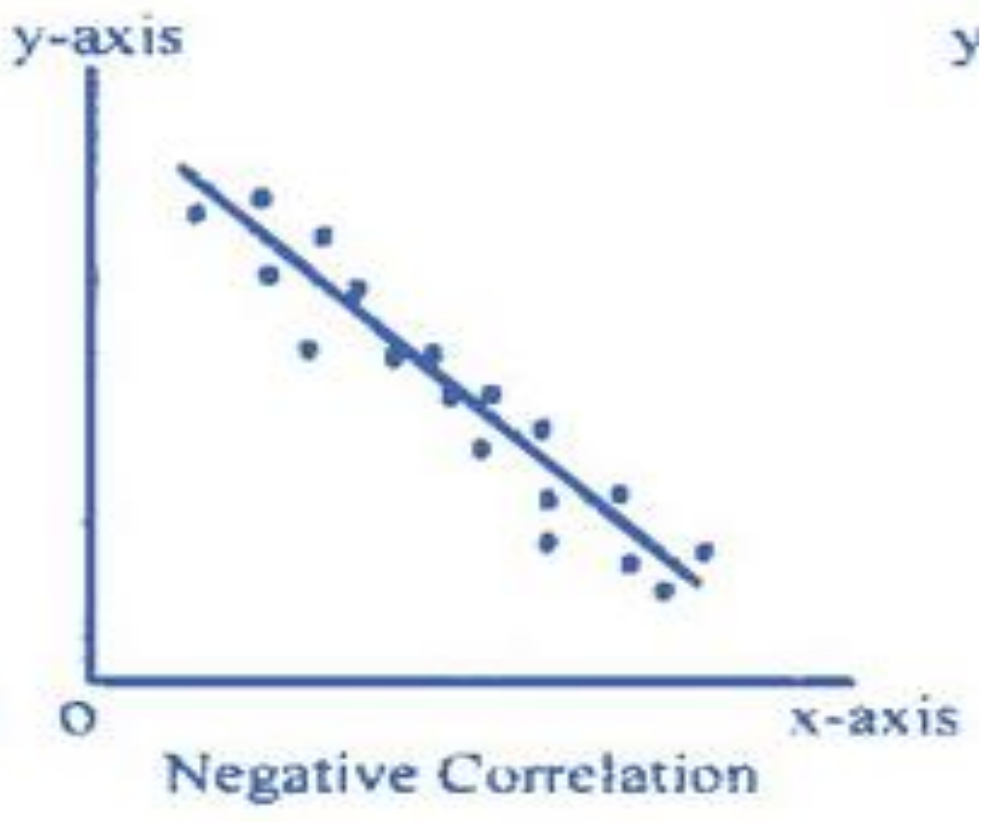
Scatter Diagram -Relation between X and Y



Positive and Negative Correlation

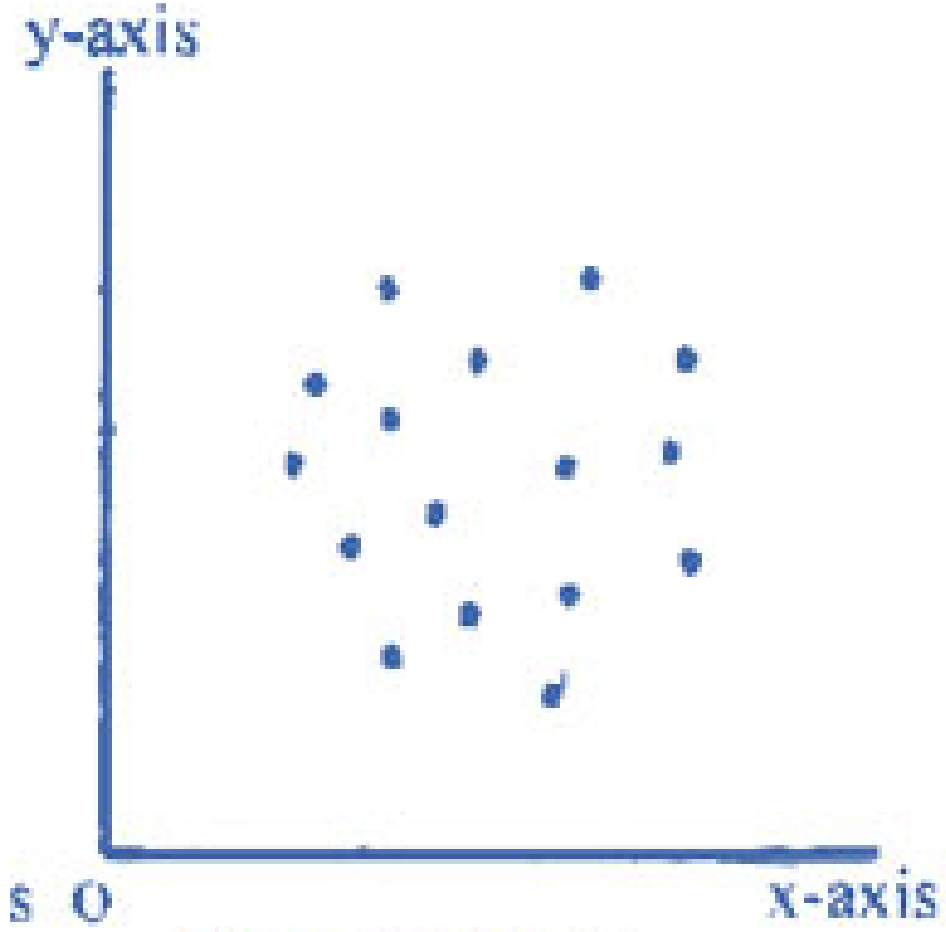


Positive Correlation

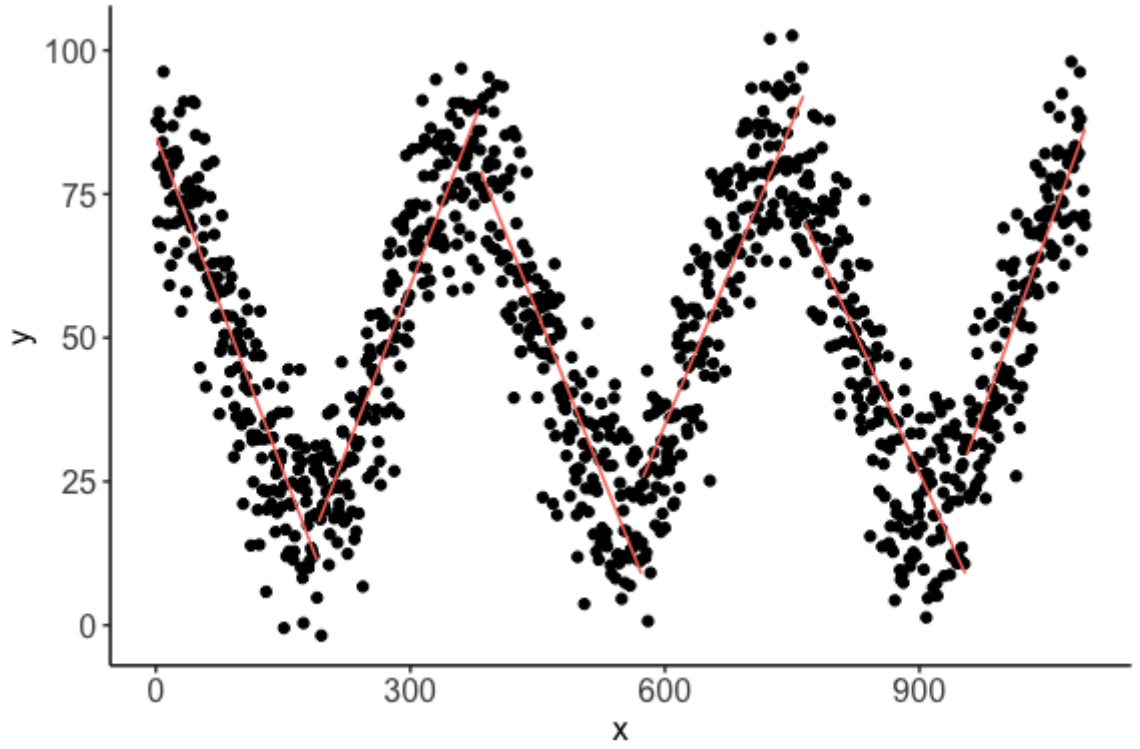


Negative Correlation

No-Correlation



Non-Linear Correlation



Relation between X and Y

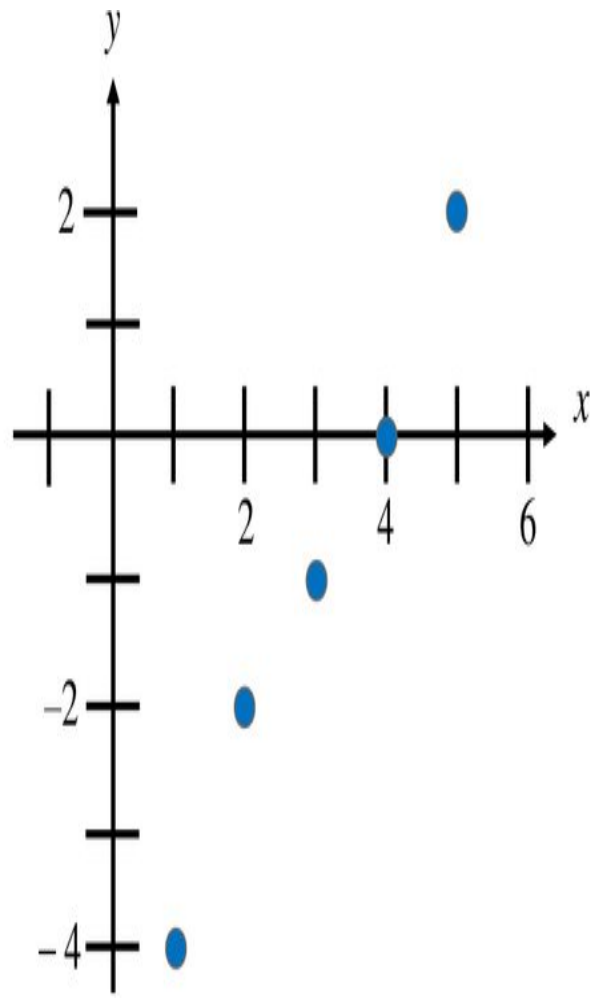
- If X and Y variables move in the same direction (i.e., either both of them increase or both decrease) the relationship between them is said to be **positive correlation**.
- if X and Y variables move in the opposite directions (i.e., if variable X increases and variable Y decreases or vice-versa) the relationship between them is said to be **negative correlation** .
- If Y is unaffected by any change in X variable, then the relationship between them is said to be **un-correlated**
- If the amount of variations in variable X bears a constant ratio to the corresponding amount of variations in Y, then the relationship between them is said to be **linear-correlation**
- otherwise it is **non-linear or curvilinear correlation**
- Since measuring non-linear correlation for data analysis is far more complicated, we therefore, generally make an assumption that the association between two variables is of the linear type

Simple Correlation

- If the relationship is confined to two variables only, it is called simple correlation

Example:

x	1	2	3	4	5
y	−4	−2	−1	0	2



Simple Correlation-Example

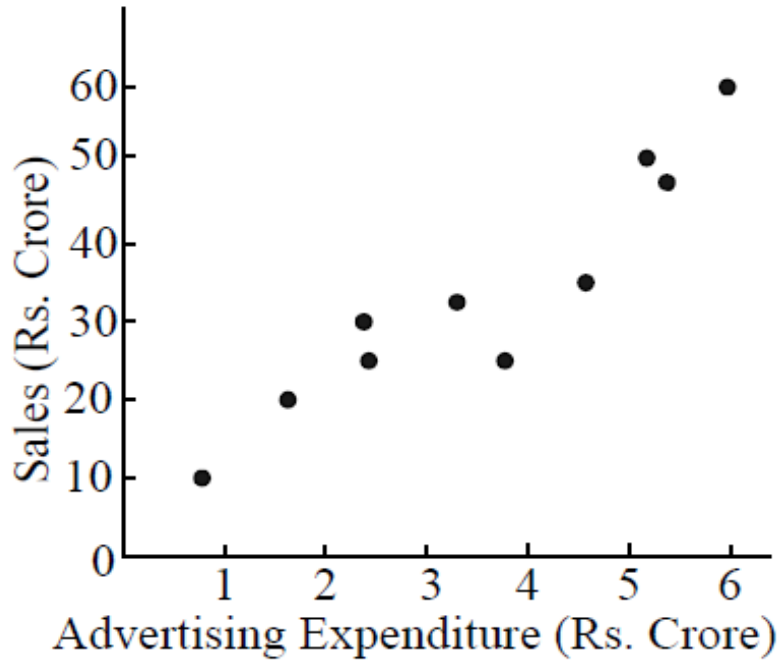
A Company's Advertising Expenses and Sales Data (Rs. in crore)

Years :	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Advertise-ment expenses (X)	6	5	5	4	3	2	2	1.5	1.0	0.5
Sales (Y)	60	55	50	40	35	30	20	15	11	10

- The company's sales manager claims the sales variability occurs because the marketing department constantly changes its advertisement expenditure.
- He/she is quite certain that there is a relationship between sales and advertising
- but does not know what the relationship is

correlation coefficient

Scatter Diagram of Sales and Advertising Expenditure for a Company



- Scatter Chart indicates that advertising expenditure and sales seem to be linearly (positively) related.
- However, the strength of this relationship is not known
- that is, how close do the points come to fall on a straight line is yet to be determined.
- The quantitative measure of strength of the linear relationship between two variables (here sales and advertising expenditure) is called the **correlation coefficient**

Practise Exercise

1) Suggest eight pairs of variables, four in each, which you expect to be positively correlated and negatively correlated

2) How does a scatter diagram approach help in studying the correlation between two variables?

Coefficient of correlation

- Coefficient of correlation helps in measuring the degree of relationship between two variables, X and Y.
- The methods which are used to measure the degree of relationship
  - **Karl Pearson's Correlation Coefficient**
  - **Spearman's Rank Correlation**

Karl Pearson's Correlation Coefficient

- Karl Pearson's coefficient of correlation (r) is one of the mathematical methods of measuring the degree of correlation between any two variables X and Y is given as:

- Population correlation coefficient

$$r = \frac{\sum(X - \bar{X})\ (Y - \bar{Y})/n}{\sigma_X\ \sigma_Y}$$

Covariance Formula

cuemath
THE MATH EXPERT

Covariance Formula For Population

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Covariance Formula For Sample

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Karl Pearson's Correlation Coefficient

The simplified formulae (which are algebraic equivalent to the above formula) are:

1) $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}, \text{ where } x = X - \overline{X}, \quad y = Y - \overline{Y}$$

Note: This formula is used when $\overline{X}$ and $\overline{Y}$ are integers.

2) $$r = \frac{\Sigma XY - \frac{\Sigma X . \Sigma Y}{n}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}}$$

Important Points –Correlation Coefficient

i) r is a dimensionless number whose numerical value lies between +1 to -1.

- The value +1 represents a perfect positive correlation
- The value -1 represents a perfect negative correlation
- The value 0 (zero) represents lack of correlation

ii) The coefficient of correlation is a pure number and is independent of the units of measurement of the variables

iii) The correlation coefficient is independent of any change in the origin and scale of X and Y values

Interpreting correlation result

Avoid spurious or nonsense correlation

Remark: Care should be taken when interpreting the correlation results.

Although a change in advertising may, in fact, cause sales to change, the fact that the two variables are correlated does not guarantee a cause and effect relationship.

Two seemingly unconnected variables may often be highly correlated.

For example, we may observe a high degree of correlation:

(i)between the height and the income of individuals

(ii) between the size of the shoes and the marks secured by a group of persons

even though it is not possible to conceive them to be casually related. When correlation exists between such two seemingly unrelated variables, it is called spurious or nonsense correlation. Therefore we must avoid basing conclusions on spurious correlation.

Spurious Correlation Examples

- Interesting correlations are easy to find, but many will turn out to be spurious.
- Three examples are the skirt length theory, the super bowl indicator, and a suggested correlation between race and college completion rates.
- **Skirt Length Theory**: Originating in the 1920s, the skirt length theory holds that skirt lengths and stock market direction are correlated. If skirt lengths are long, the correlation is that the stock market is bearish. If shirt lengths are short, the market is bullish.[1]
- **Super Bowl Indicator**: In late January, there is often chatter about the so-called Super Bowl indicator, which suggests that a win by the American Football Conference team likely means that the stock market will go down in the coming year, whereas a victory by the National Football Conference team portends a rise in the market. Since the beginning of the Super Bowl era, the indicator has been accurate around 74% of the time, or 40 out of the 54 years, according to OpenMarkets.[2] It is a fun conversation piece but probably not something a serious financial advisor would recommend as an investment strategy for clients.
- **Educational Attainment and Race**: Social scientists have focused on identifying which variables impact educational attainment. According to government research, 56% of White 25- to 29-year-olds had completed a college degree in 2019, compared to just 36% of black individuals of the same age.[3] The implication being that race has a causal effect on college completion rates.

Ref: https://www.investopedia.com/terms/s/spurious_correlation.asp#:~:text=Spurious%20correlation%2C%20or%20spuriousness%2C%20occurs,a%20third%20%22confounding%22%20factor.

Calculation of Correlation Coefficient

- Data of advertisement expenditure (X) and sales (Y) of a company for 10 years shown in the table

Years :	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
Advertise-ment expenses (X)	6	5	5	4	3	2	2	1.5	1.0	0.5
Sales (Y)	60	55	50	40	35	30	20	15	11	10

Calculation of Correlation Coefficient

$$r = \frac{\sum XY - \frac{\sum(X)\sum(Y)}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}}$$

Advertisement expenditure Rs. (X)	Sales Rs. (Y)	XY	X^2	Y^2
6	60	360.0	36	3600
5	55	275.0	25	3025
5	50	250.0	25	2500
4	40	160.0	16	1600
3	35	105.0	9	1225
2	30	60.0	4	900
2	20	40.0	4	400
1.5	15	22.5	2.25	225
1.0	11	11.0	1	121
0.5	10	5.0	0.25	100
$\Sigma X = 30$	$\Sigma Y = 326$	$\Sigma XY = 1288.5$	$\Sigma X^2 = 122.50$	$\Sigma Y^2 = 13696$

Calculation of Correlation Coefficient

$$r = \frac{\sum XY - \frac{\sum(X)\sum(Y)}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}}$$

$$= \frac{1288.5 - \frac{(30)(326)}{10}}{\sqrt{122.5 - \frac{(30)^2}{10}}\sqrt{13696 - \frac{(326)^2}{10}}} = \frac{310.5}{315.7}$$

$$= 0.9835$$

Note: The calculated coefficient of correlation r = 0.9835 shows that there is a high degree of association between the sales and advertisement expenditure. For this particular problem, it indicates that an increase in advertisement expenditure is likely to yield higher sales.

If the results of the calculation show a strong correlation for the data, either negative or positive, then the line of best fit to that data will be useful for forecasting

Spearman Correlation

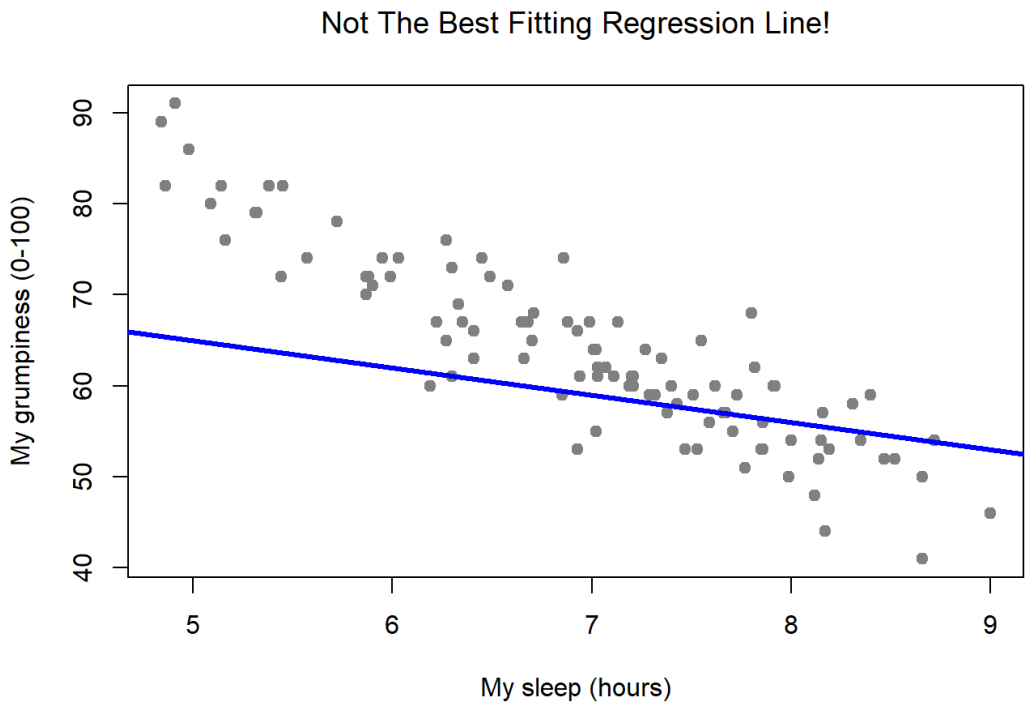
SIMPLE LINEAR REGRESSION

- When we identify the fact that the correlation exists between two variables, we shall develop an estimating equation, known as regression equation or estimating line.
- **Regression equation or Estimating line:** a methodological formula, which helps us to estimate or predict the unknown value of one variable from known value of another variable.
- Regression analysis attempts to establish the nature of the relationship between variables, that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting.
- For example, if we confirmed that advertisement expenditure (independent variable), and sales (dependent variable) are correlated, we can predict the required amount of advertising expenses for a given amount of sales or vice-versa.
- The statistical method which is used for prediction is called regression analysis.
- When the relationship between the variables is linear, the technique is called **simple linear regression**.

Regression and Correlation coefficient

- Hence, the technique of regression goes one step further from correlation
- It is about relationships that have been true in the past as a guide to what may happen in the future.
- To do this, need the regression equation and the correlation coefficient.
- The latter is used to determine that the variables are really moving together.

Best Fitting Regression Line



Basic ideas in regression are closely tied to correlation

Example

- we were trying to find out why Dan is so very grumpy all the time, and our working hypothesis was that I'm not getting enough sleep. We drew some scatterplots to help us examine the relationship between the amount of sleep I get, and my grumpiness the following day.
- correlation of r=−.90,
- but what we find ourselves secretly imagining is something that looks closer to Figure 2 in previous slide. That is, we mentally draw a straight line through the middle of the data. In statistics, this line that we're drawing is called a ***regression line***.
- Ref:https://learningstatisticswithr.com/book/regression.html

Regression Line Vs Straight Line

- The formula for a straight line

  $y=mx+c$

  - The two *variables* are x and y
  - two *coefficients*, m and c.
  - The coefficient m represents the *slope* of the line
  - The coefficient c represents the y-*intercept* of the line

Intercept is interpreted as “the value of y that you get when x=0”

Slope of m means that if you increase the x-value by 1 unit, then the y-value goes up by m units;

A negative slope means that the y-value would go down rather than up

Regression Line Vs Straight Line

- If Y is the outcome variable (the DV) and X is the predictor variable (the IV), then the formula that describes our regression is written like this:

$$\acute{Y} = b1Xi + b0$$

- Xi is the value of predictor variable for the ith observation (i.e., the number of hours of sleep that I got on day i of my little study),
- Yi is the corresponding value of the outcome variable (i.e., my grumpiness on that day).
- Assuming is that this formula works for all observations in the data set (i.e., for all i).
- $\acute{Y}$ and not Yi- This is because we want to make the distinction between the *actual data* Yi, and the *estimate* $\acute{Y}$ (i.e., the prediction that our regression line is making).
- letters used to describe the coefficients from m and c to b1 and b0. That's just the way that statisticians like to refer to the coefficients in a regression model.

The Regression Line

- The data don't fall perfectly on the line.
- Or, to say it another way, the data $\acute{Y}$ are not identical to the predictions of the regression model .
- The difference between the model prediction and that actual data point as a *residual*, ϵi
- Written using mathematics, the residuals are defined as:

$$\epsilon i = Yi - \acute{Y}$$

complete linear regression model as:

$$Yi = b1Xi + b0 + \epsilon i$$

simple linear regression

The objective of simple linear regression is to represent the relationship between two variables with a model of the form shown below:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

wherein

Y_i = value of the dependent variable,

β_0 = Y-intercept,

β_1 = slope of the regression line,

X_i = value of the independent variable,

e_i = error term (i.e., the difference between the actual Y value and the value of Y predicted by the model.

ESTIMATING THE LINEAR REGRESSION

A good regression model is with small residuals.
The “best fitting” regression line is the one that has the smallest residuals

ESTIMATING THE LINEAR REGRESSION

- If we consider the two variables (X variable and Y variable), we shall have two regression lines. They are:

  i) Regression of Y on X - estimates value of Y for given value of X
  ii) Regression of X on Y- estimates the value of X for given value of Y

- These two regression lines will coincide if correlation between the variable is either perfect positive or perfect negative.

Regression Equation of Y on X

$$\hat{Y} = a + bx$$

where, $\hat{Y}$ is the computed values of Y (dependent variable) from the relationship for a given X, 'a' and 'b' are constants (fixed values), 'a' determines the level of the fitted line at Y-axis (Y-intercept), 'b' determines the slope of the regression line, X represents a given value of independent variable.

The alternative simplified expression for the above equation is:

$$\hat{Y} - \overline{Y} = byx\ (X - \overline{X})$$

$$byx = r\frac{\sigma_y}{\sigma_x} = \frac{(\Sigma XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

Regression Equation of Y on X

$$\hat{X} = a + by$$

Alternative simplified expression is:

$$\hat{X} - \overline{X} = bxy\,(Y - \overline{Y})$$

$$bxy = r\frac{\sigma x}{\sigma y} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}$$

Reduce error - **Least Squares Method**

- we may get an infinite number of possible regression lines for a set of data points. We must, therefore, establish a criterion for selecting the best line.
- The criterion used is the **Least Squares Method**
- According to the least squares criterionBest regression line is the one that minimizes the sum of squared vertical distances between the observed (X, Y) points and the regression line. $\sum (Y - \hat{Y})^2$

i.e:

It is the least value and the sum of the positive and negative deviations is zero

$$\sum (Y - \hat{Y}) = 0$$

It is important to note that the distance between (X, Y) points and the regression line is called the error.

Exercise

From the following 12 months sample data of a company, estimate the regression lines and also estimate the value of sales when the company decided to spend Rs. 2,50,000 on advertising during the next quarter.

(Rs. in lakh)

Advertisement Expenditure:	0.8	1.0	1.6	2.0	2.2	2.6	3.0	3.0	4.0	4.0	4.0	4.6
Sales:	22	28	22	26	34	18	30	38	30	40	50	46