



A • P • U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

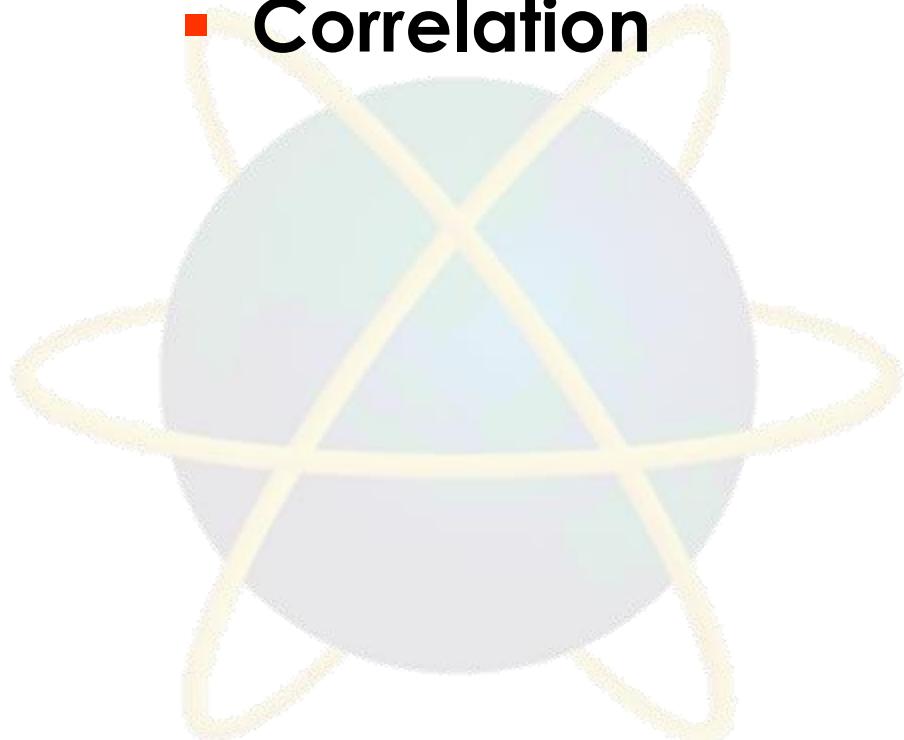
Probability & Statistical Modelling

AQ077-3-2-PSMOD and Version VD1

Correlation & Regression Analysis

Topic & Structure of The Lesson

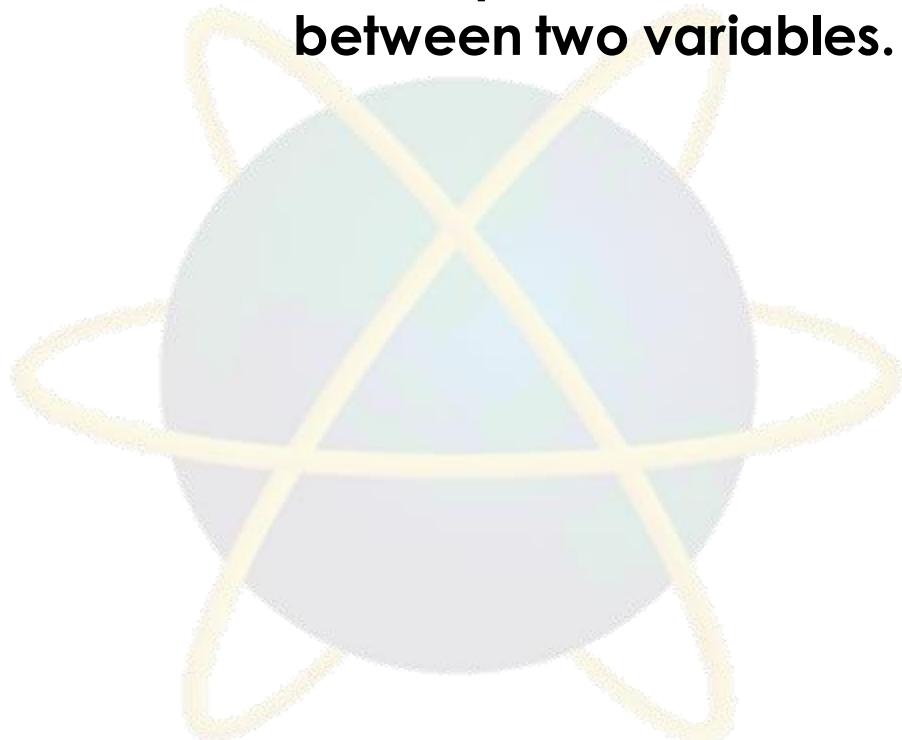
- **Introduction**
- **Regression**
- **Correlation**



Learning Outcomes

➤ At the end of this section, You should be able to:

- Analyse bi-variate data using regression & correlation techniques used to measure the linear relationship between two variables.



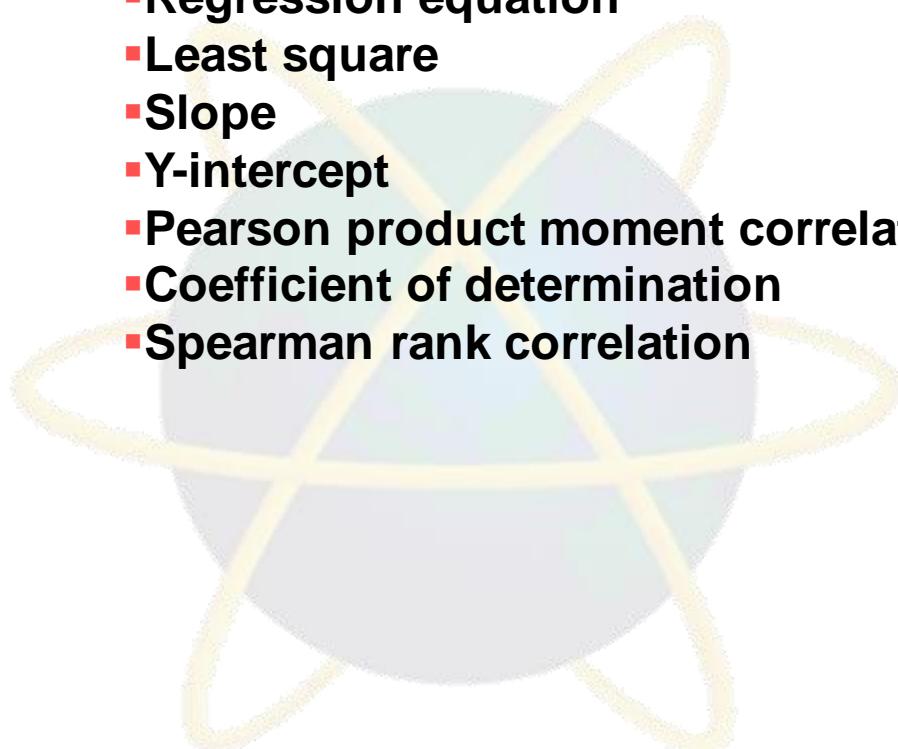
Key Terms You Must Be Able To Use



If you have mastered this topic, you should be able to use the following terms correctly in your assignments and exams:

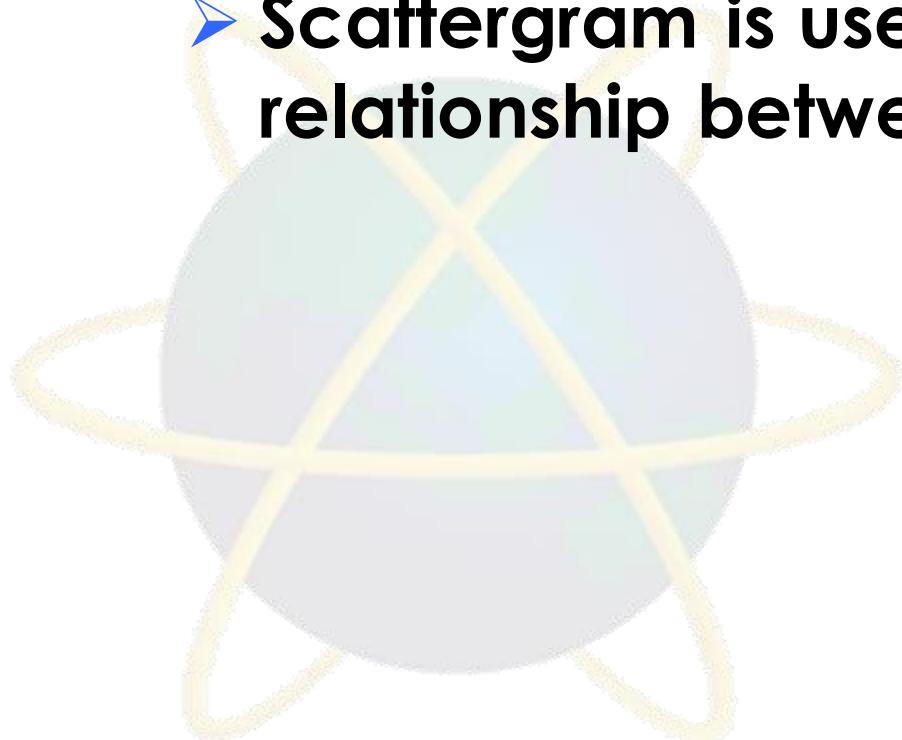
(Prepare your own list)

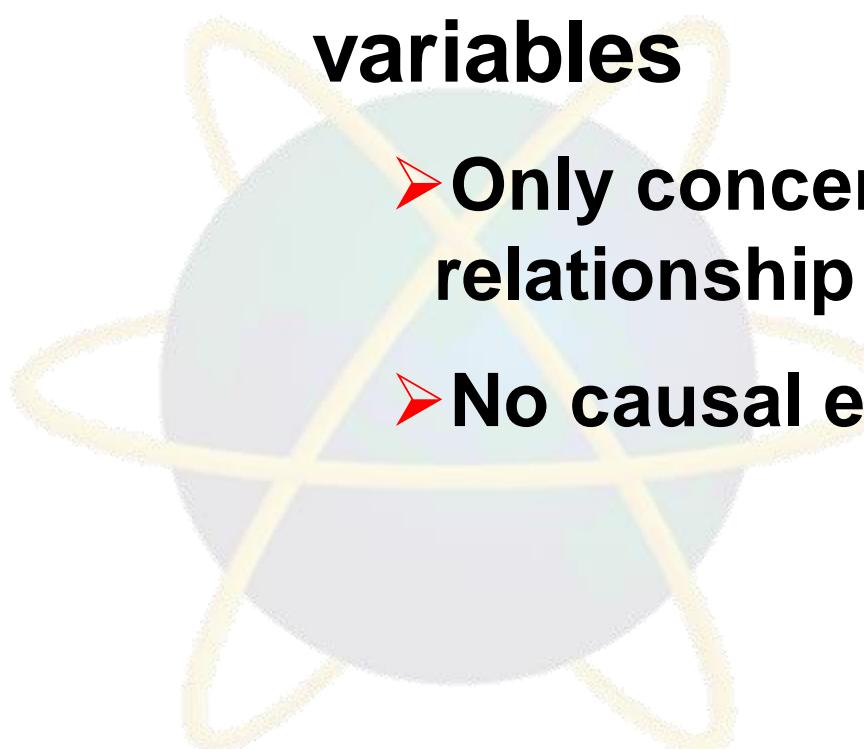
- Regression equation
- Least square
- Slope
- Y-intercept
- Pearson product moment correlation
- Coefficient of determination
- Spearman rank correlation



Introduction

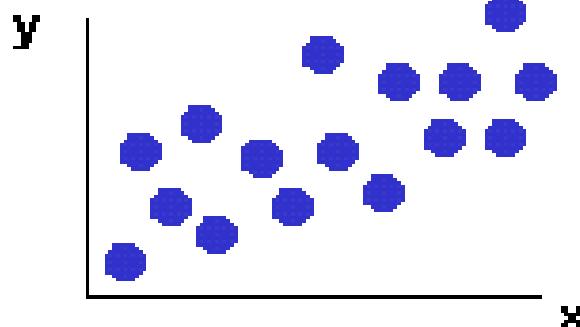
- Correlation & Regression are concerned with measuring the linear relationship between two variables.
- Scattergram is used to illustrate any relationship between two variables.



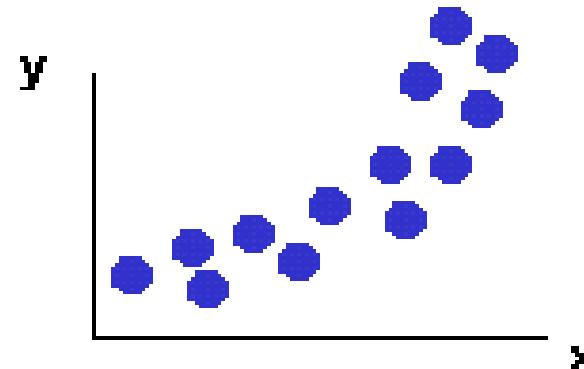
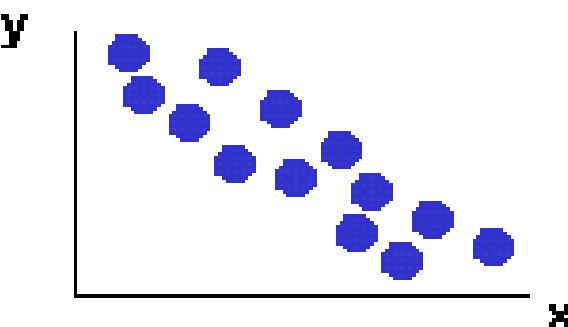
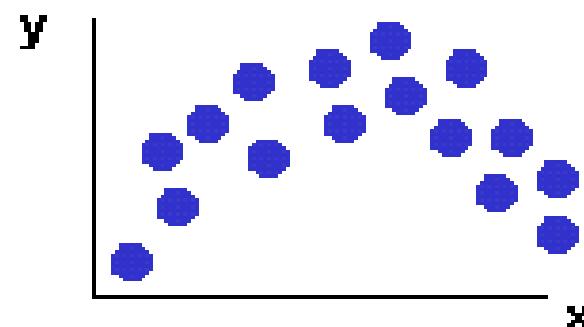
- 
- **Correlation analysis is used to measure strength of the association (linear relationship) between two variables**
 - Only concerned with strength of the relationship
 - No causal effect is implied

➤ Examples of scatter plots

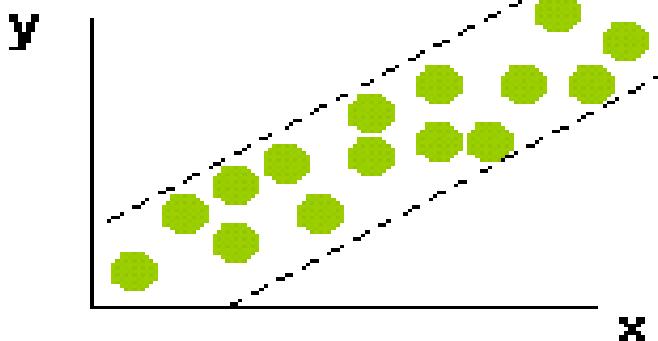
Linear relationships



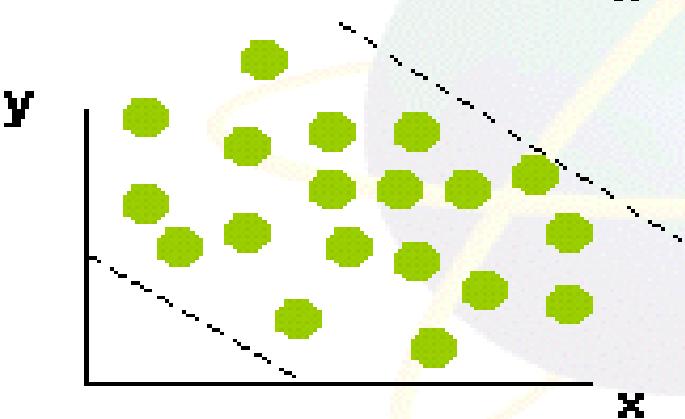
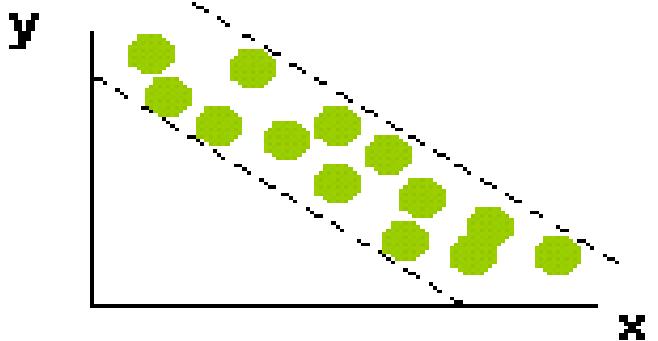
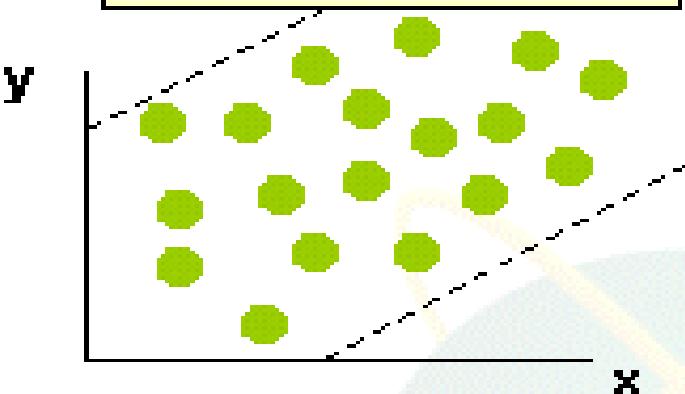
Curvilinear relationships



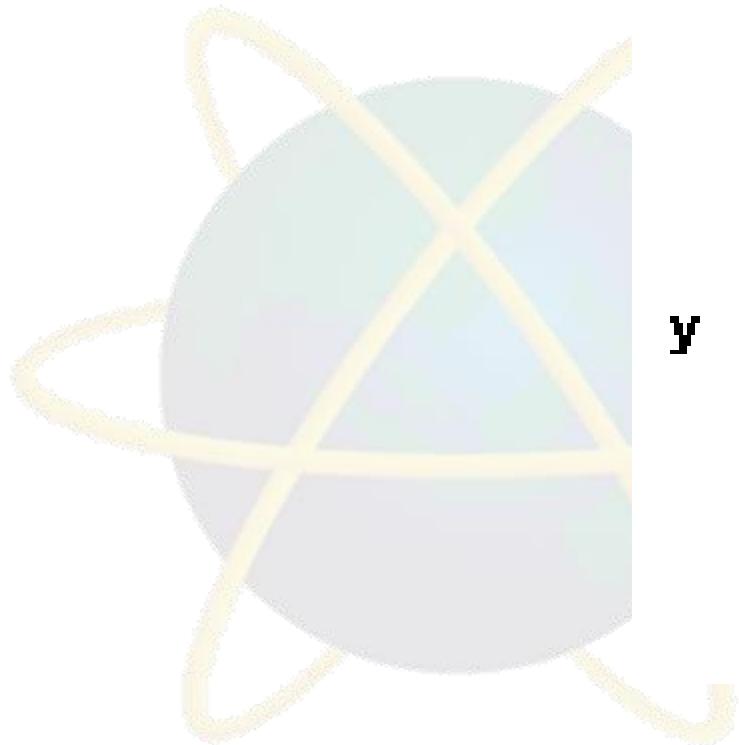
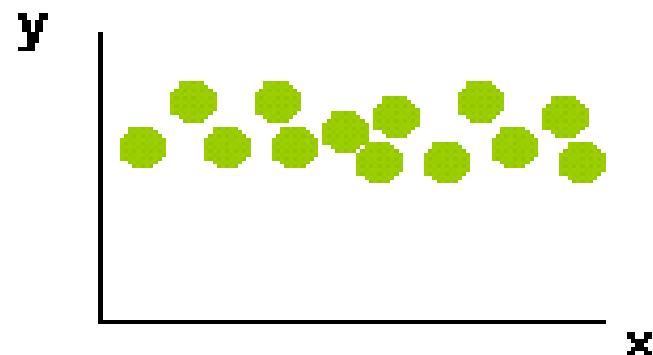
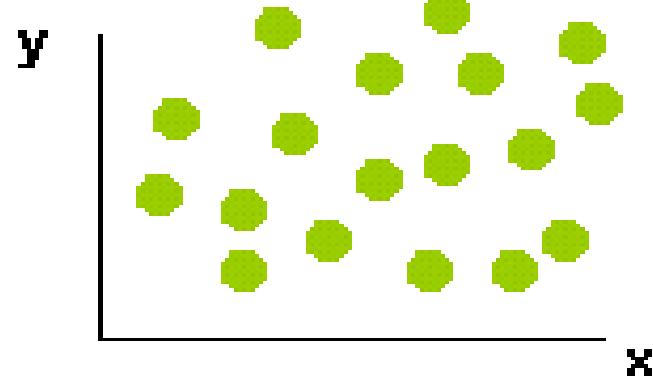
Strong relationships



Weak relationships



No relationship

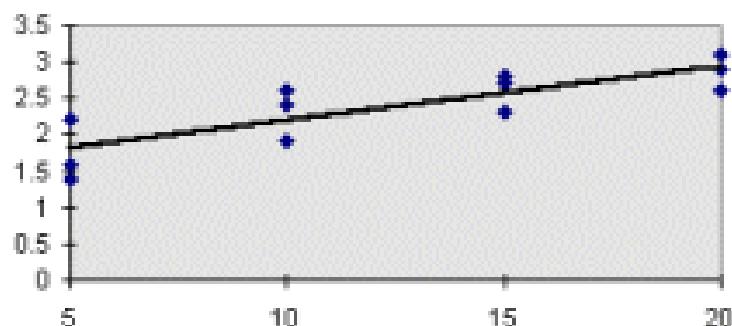


■ Regression

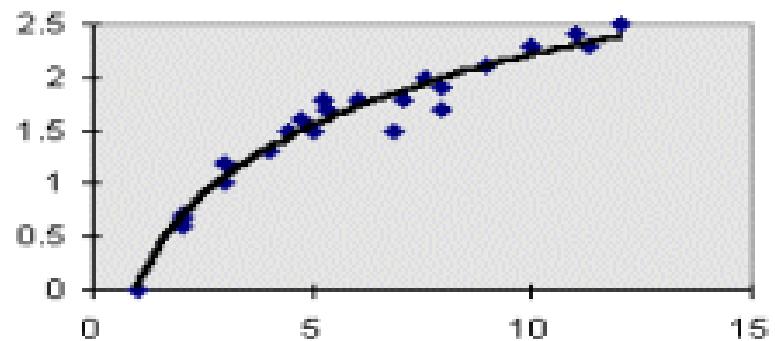
- Regression is concerned with obtaining a mathematical equation which describes the relationship between two variables.
 - The independent variable is the one that is chosen freely or occurs naturally.
 - The dependent variable occurs as a consequence of the value of the independent variable.
- It is normally used for estimation purposes.

➤ Types of Regression Models

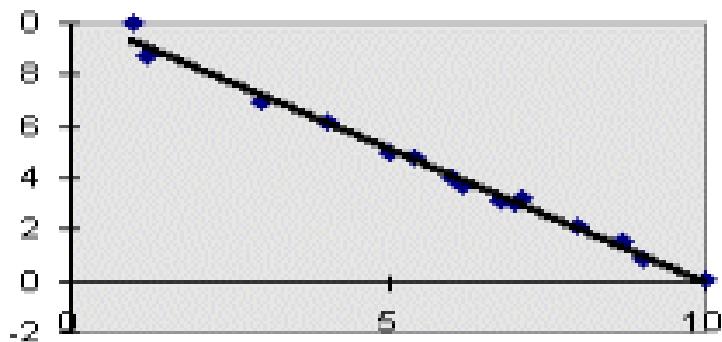
Positive Linear Relationship



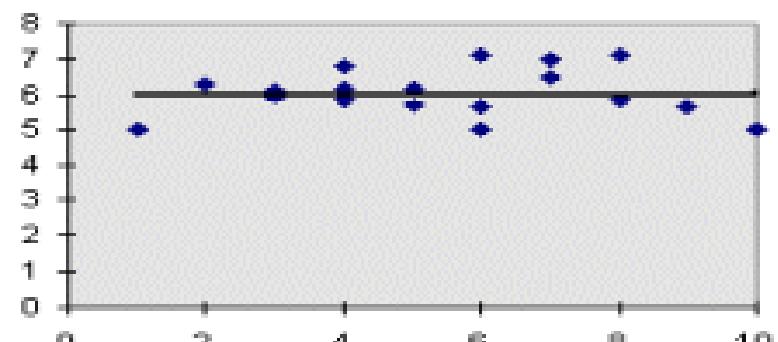
Relationship NOT Linear



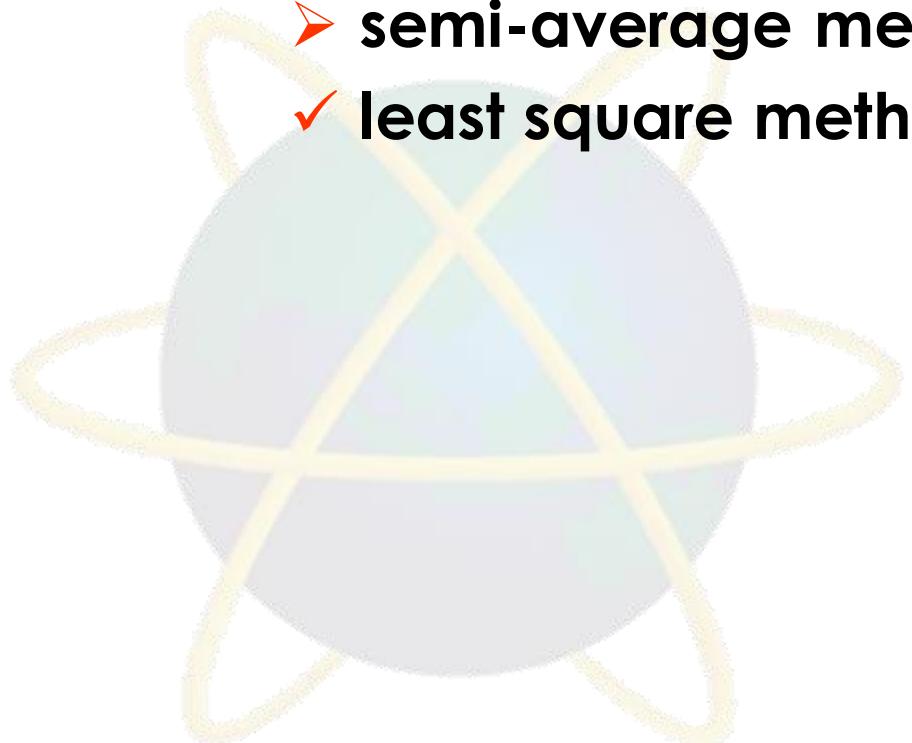
Negative Linear Relationship



No Relationship



- **3 common methods used to determine a regression line**
 - inspection method
 - semi-average method
 - ✓ least square method



■ Least square method

- the standard method of obtaining a regression line.
- For any set of bivariate, there are two regression line which can be obtained
 - x on y regression line
 - used for estimating x given a value of y
 - ✓ y on x regression line
 - used for estimating y given a value of x.
- Note that for this syllabus, only the y on x regression line is dealt with.

- If the least square equation is given by
 $y = a + bx$,
then,

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n}$$

Example 1

- The following table shows the amount spent on advertising and the corresponding sales of the product from 6 companies.

Company	Advertising Cost (\$000)	Sales (\$000)
A	8	25
B	12	35
C	11	29
D	5	24
E	14	38
F	3	12

- (a) Plot a scattergram showing the relationship between advertising cost and sales of the product.
- (b) Calculate the equation of the regression line of sales on advertising costs. Draw the regression line on the scattergram.
- (c) Use the regression line to forecast sales if advertising costs were
 - (i) \$10000
 - (ii) \$1000
- (d) Justify your answer in part (c)(ii).

■ Correlation

- It is a technique used to measure the strength of relationship between two variables by measuring the degree of 'scatter' of the data values.
- The less scatter the data values are, the stronger the correlation.
- Two types of correlation
 - Positive (direct)
 - Negative (inverse)

➤ **Measures of correlation**

- **Product moment correlation coefficient**
- **Coefficient of determination**
- **Spearman rank correlation coefficient**



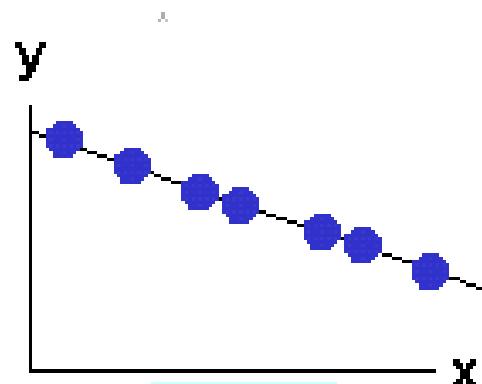
■ Product moment correlation coefficient, r

- It measures the extent to which two variables move in sympathy with or in opposition to one another.

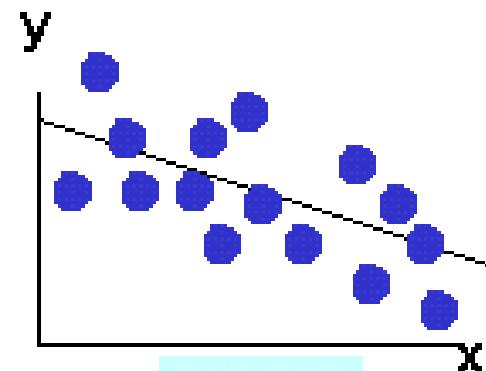
$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

- The correlation coefficient, r lies between 0 and ± 1 .
- When $r = 0$, it signifies there is no correlation present
- When $r = 1$, it signifies perfect positive correlation
- When $r = -1$, it signifies perfect negative correlation
- The further away r is from 0, the stronger is the correlation.

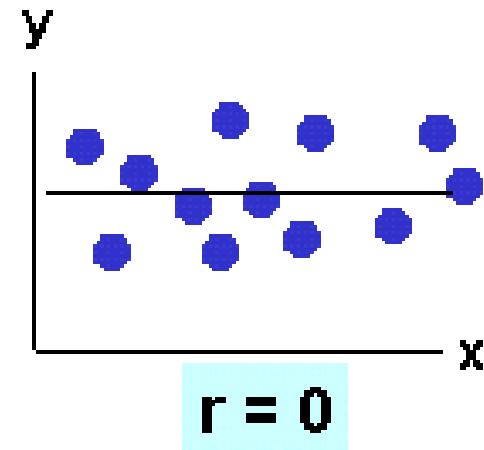
➤ Examples of appropriate r values



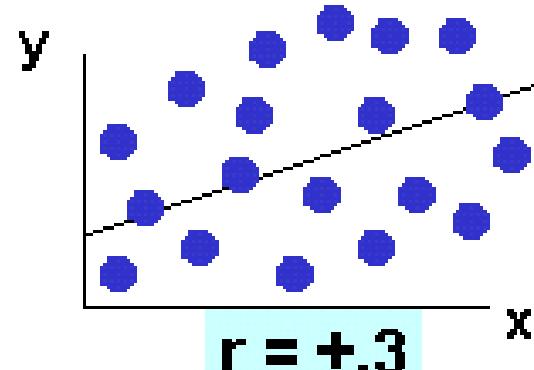
$r = -1$



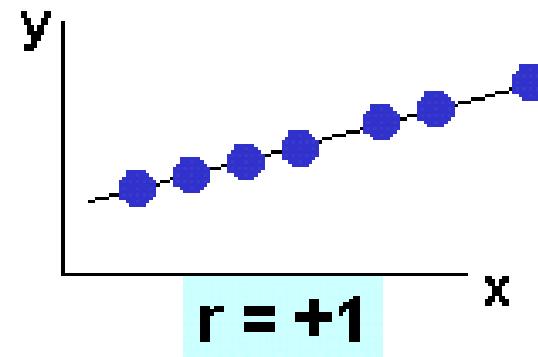
$r = -.6$



$r = 0$



$r = +.3$



$r = +1$

■ Coefficient of determination, r^2

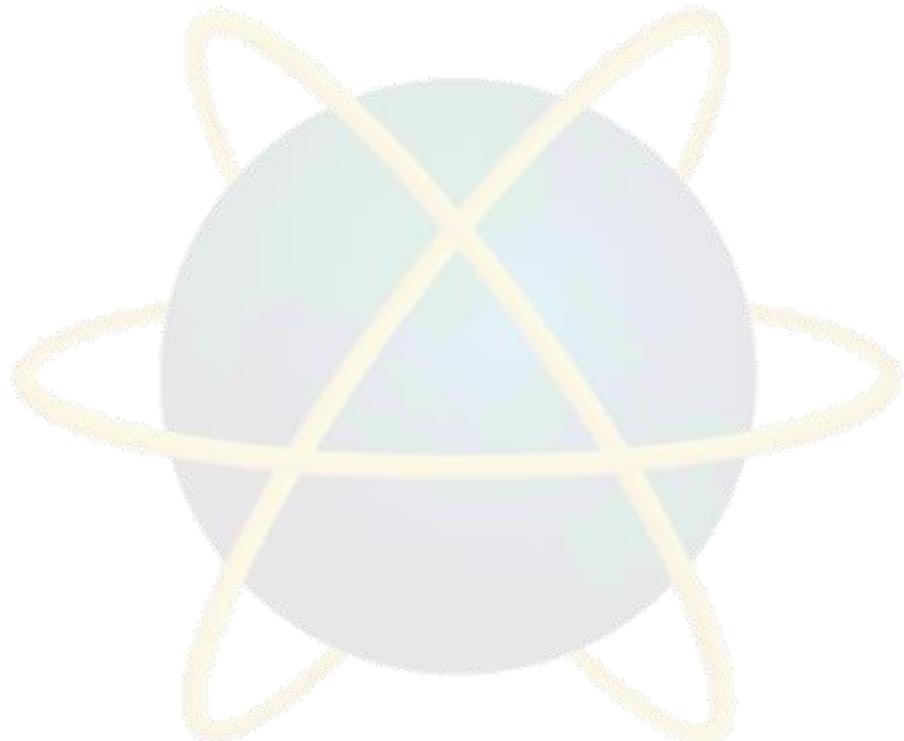
- It indicates the proportion of variance in the dependent variable that is explained statistically by knowledge of the independent variable and vice versa.
- Notice that, since $-1 \leq r \leq +1$, it follows that $0 \leq r^2 \leq +1$

Example 2

Use the data of *Example 1*, calculate the

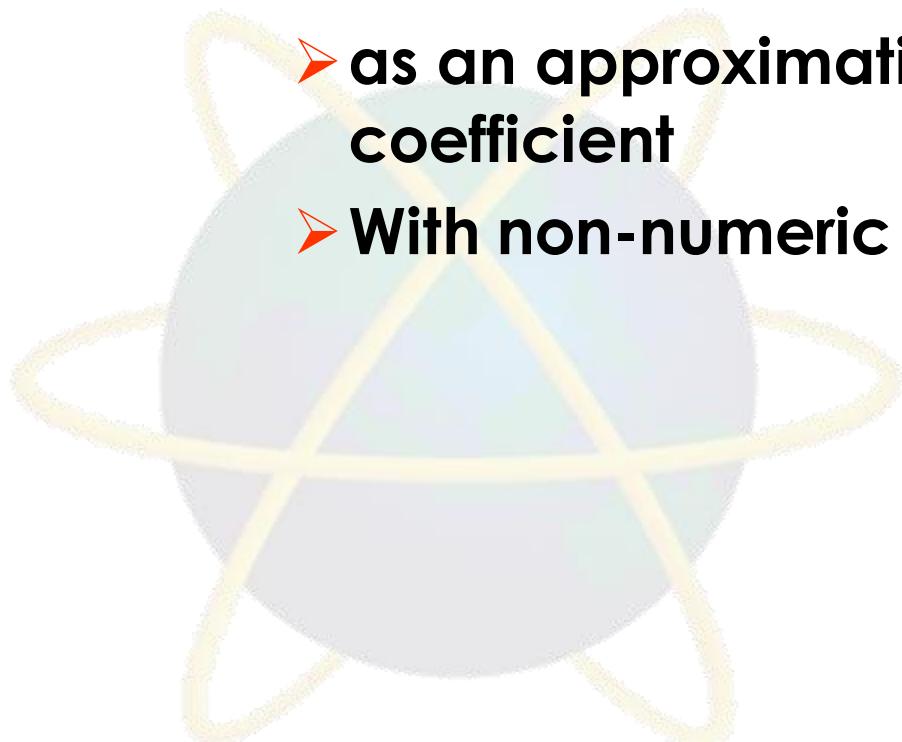
- (i) Product moment correlation coefficient
- (ii) Coefficient of determination

and interpret the result.



■ Spearman rank correlation coefficient, r_s

- It can be used:
 - as an approximation to the product moment coefficient
 - With non-numeric data that can be ranked



➤ Procedure for obtaining r_s

- Rank the x values, r_x
- Rank the y values, r_y
- For each pair of ranks, calculate $d^2 = (r_x - r_y)^2$
- Calculate $\sum d^2$
- The value of the rank correlation coefficient can then be calculated as below:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Example 3

In a survey of TV viewers in Sabah and KL, the following programme were ranked in order of preference. Calculate the Spearman's rank correlation coefficient for the data. Comment on the result.

TV programme	Sabah	KL
Biggest Loser	1	2
Amazing Race	2	5
TV3 news	4	4
Hero	3	6
24	5	1
CSI	7	7
Ultraman	6	3

Example 4

- The following table shows the marks of eight pupils in biology and chemistry. Find the value of Spearman's coefficient of rank correlation.

Biology (x)	65	65	70	75	75	80	85	85
Chemistry (y)	50	55	58	55	65	58	61	65

- **Comparison of rank and product moment correlation** (with (+) and (-) signifying whether the feature can be thought of as an advantage or disadvantage respectively)

- **Product moment coefficient**

- The standard measure of correlation (+)
 - Data must be numeric (-)
 - The calculations can be awkward. (-)

➤ Rank coefficient

- Only an approximation to the product moment coefficient. (-)
- Easier to use with less involved calculations. (+)
- Can be used with non-numeric data. (+)
- Can be insensitive to small changes in actual values. (-)

■ Practical difficulties in drawing conclusions from correlation coefficient

- A high correlation coefficient does not necessarily imply that the variables are related to one another- spurious correlation
- A low correlation coefficient between two variables does not necessarily mean that there is little relationship between them but there are also some additional factors exerting an influence.

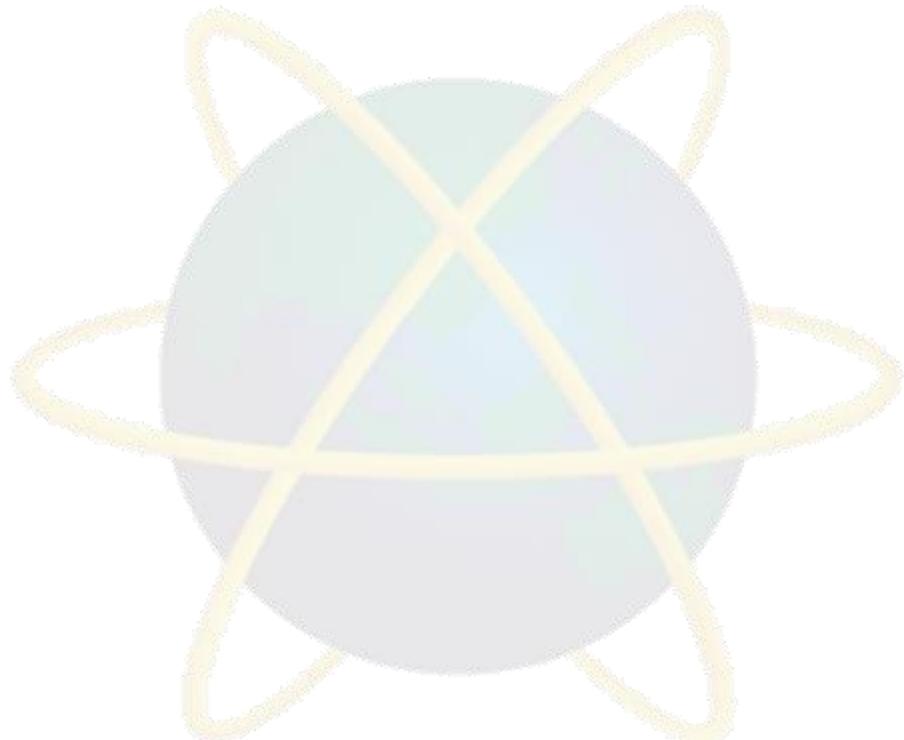
EXCEL



=SLOPE()

=INTERCEPT()

=PEARSON()

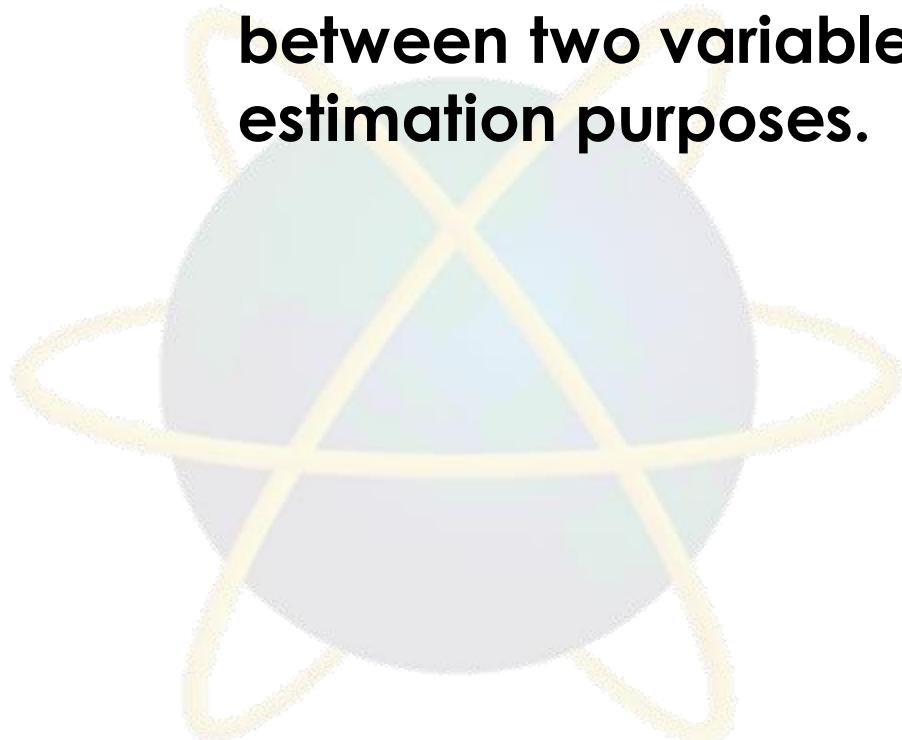


Summary of Main Teaching Points



■ Regression

- It is concerned with producing a mathematical function which describes the relationship between two variables. It is normally used for estimation purposes.



- There are three common methods used to determine a regression line for a set of bivariate data.
 - Inspection
 - Method of semi-averages
 - Method of least squares.
 - This is the standard technique for obtaining a regression line.

- In most examinations questions the bivariate variables involved will be labelled (usually x and y) and the regression line of y on x will be asked for. Where this is not the case, it is usual to label the independent variable as x and the dependent variable as y and thus the y on x regression line will be appropriate.
- Please take note that
 - Other forms of regression are sometimes appropriate and calculated, e.g. curvilinear regression.

- An independent variable can be affected by more than 1 dependent variable.
- Interpolation involves estimating a value of the dependent variable given a value of the independent variable within the range of the data used to calculate the regression line and can be carried out with some confidence. Estimation outside this range is known as extrapolation and the results should be treated with caution.

■ Correlation

- Concerned with describing how well two variables are associated by measuring the degree of 'scatter' of the data values.
- Two types
 - Positive (direct)
 - Increases in one variable are associated with increases in the other.
 - Negative (inverse)
 - Increases in one variable are associated with decreases in the other.

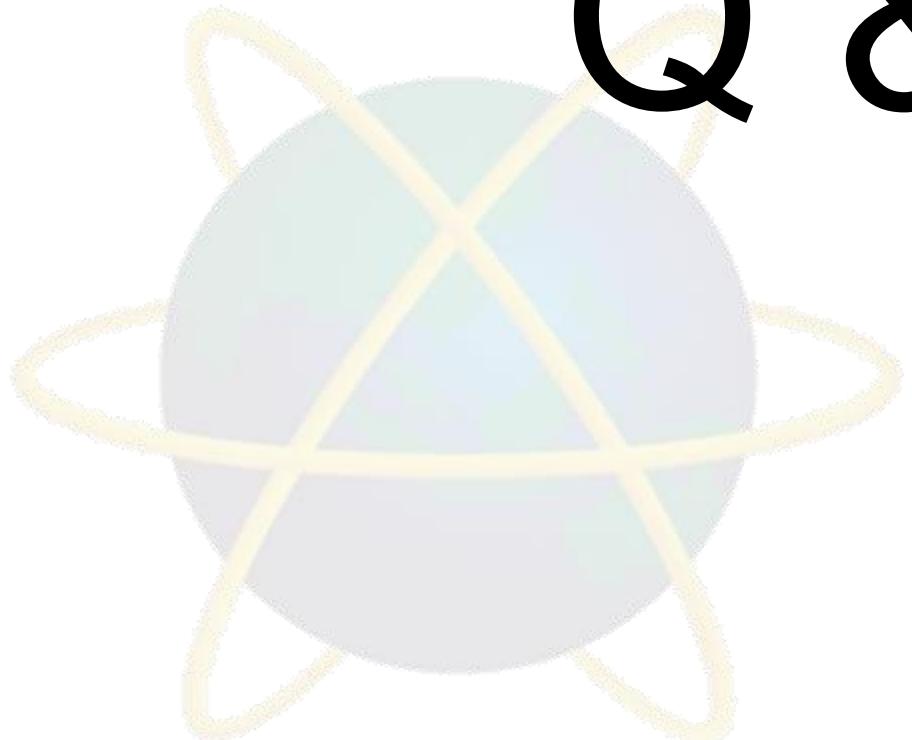
- A quantitative measure of correlation is given by the (product moment) correlation coefficient, r .
 - $-1 \leq r \leq 1$
 - $r = -1$ signifies perfect negative correlation
 - $r = 0$ signifies no correlation
 - $r = +1$ signifies perfect positive correlation
 - The product moment correlation coefficient is the standard measure of correlation for numeric data. It cannot be calculated for non-numeric data.

- The coefficient of determination, r^2 , is used to indicate the proportion of the total variation in the dependent variable (y) that is due to variations in the independent variable (x).
- Spearman's rank correlation coefficient can be used:
 - As an approximation to the product moment coefficient
 - With non- numeric data that can be ranked.
- Correlation does not necessarily imply causality.

Question and Answer Session



Q & A



What we will cover next

- **Probability Distribution**

