Topic 2

Correlation Theory

Learning objectives

- Distinguish between covariance and correlation
- Calculate and interpret the covariance coefficient
- Calculate and interpret the linear correlation coefficient
- Calculate and interpret rank correlation coefficient
- Examine the limitations of the theory of linear correlation

Introduction

Two measures of the relationship between variables are:

- Covariance
  - a measure of the direction of a linear relationship between two variables
- Correlation Coefficient
  - a measure of both the direction and the strength of a linear relationship between two variables

Covariance

- The covariance measures the direction in which two variables are linearly associated
- The population covariance:

$$\mathrm{Cov}(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- The sample covariance:

$$\mathrm{Cov}(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1}$$

  - Only how changes in one variable are associated with changes in a second variable
  - No causal effect is implied

Copyright © 2013 Pearson Education, Inc. Publishing as Prentice Hall

Interpreting Covariance

- **Covariance** between two variables:

Cov(x,y) > 0 → x and y tend to move in the same direction

Cov(x,y) < 0 → x and y tend to move in opposite directions

Cov(x,y) = 0 → x and y are independent

Copyright © 2013 Pearson Education, Inc. Publishing as Prentice Hall

Correlation

- Correlation represents the degree of relationship or strength of association between two or more variables.
  - Simple correlation: degree of association between two variables.
  - Multiple correlation: degree of association between three or more variables.
- One way to check whether two variables are related is to graph them using a scatter plot.
  - Correlation is linear when all points (X,Y) on a scatter diagram cluster near a straight line.

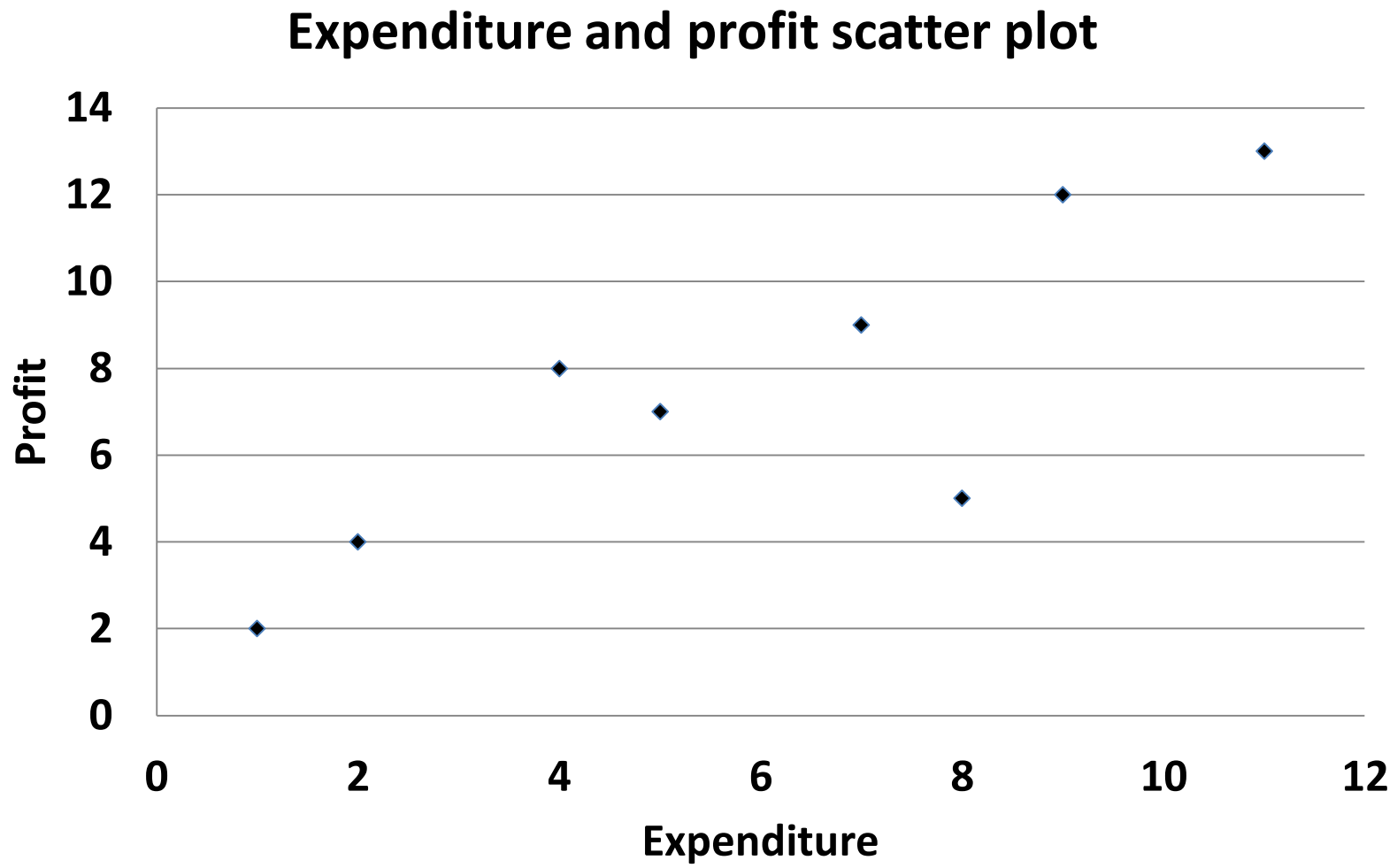
The data below relates to expenditure and profits for 8 firms.

Expenditure ('000 Kshs)	9	5	1	4	11	7	2	8
Profit ('000 Kshs)	12	7	2	8	13	9	4	5

- Draw a scatter diagram (scatter plot)
- What relationship do you observe?

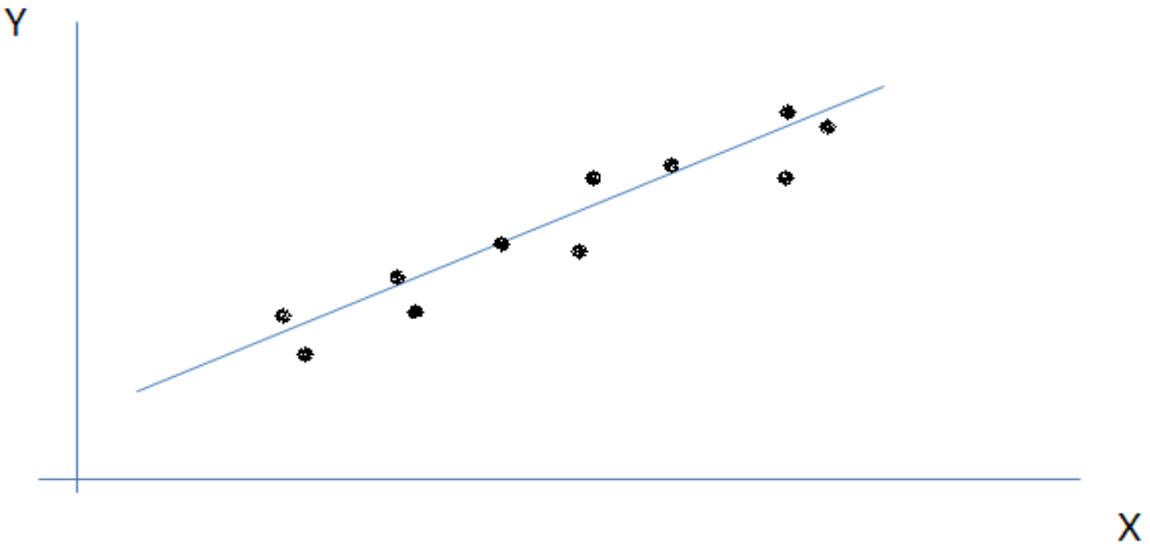
The data below relates to expenditure and profits for 8 firms.

Expenditure ('000 Kshs)	9	5	1	4	11	7	2	8
Profit ('000 Kshs)	12	7	2	8	13	9	4	5



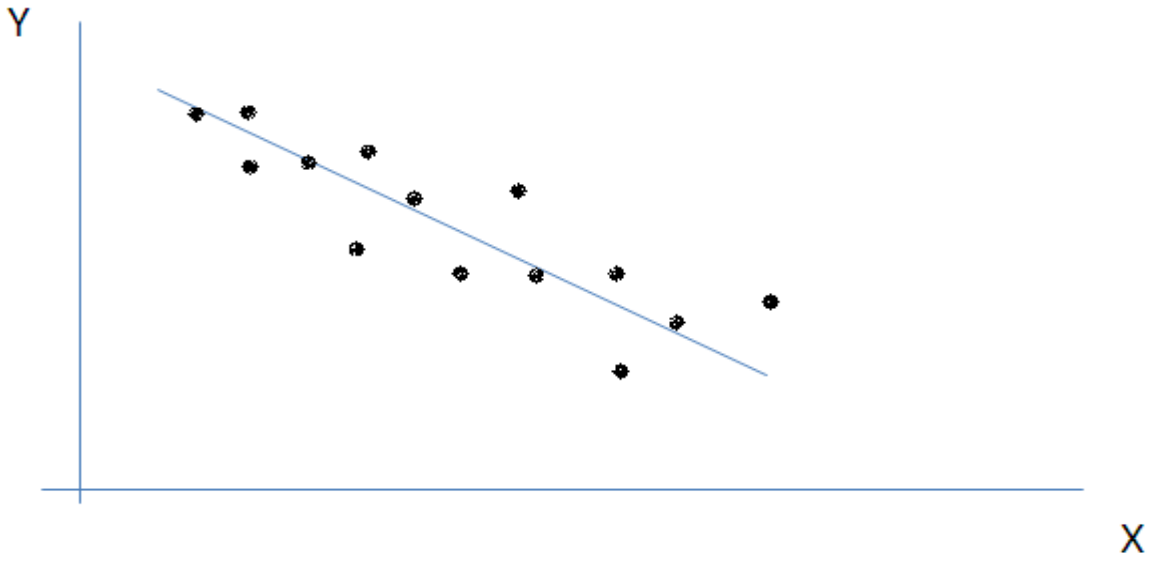
Linear Correlation Coefficient

- Two variables are positively correlated if they tend to change in the same direction i.e. increase or decrease together e.g. Qs and P.
- Points cluster around a line with a positive slope.
  - If all points lie on the line correlation is said to be perfect positive i.e. r =+1
  - The diagram below illustrates positive linear correlation



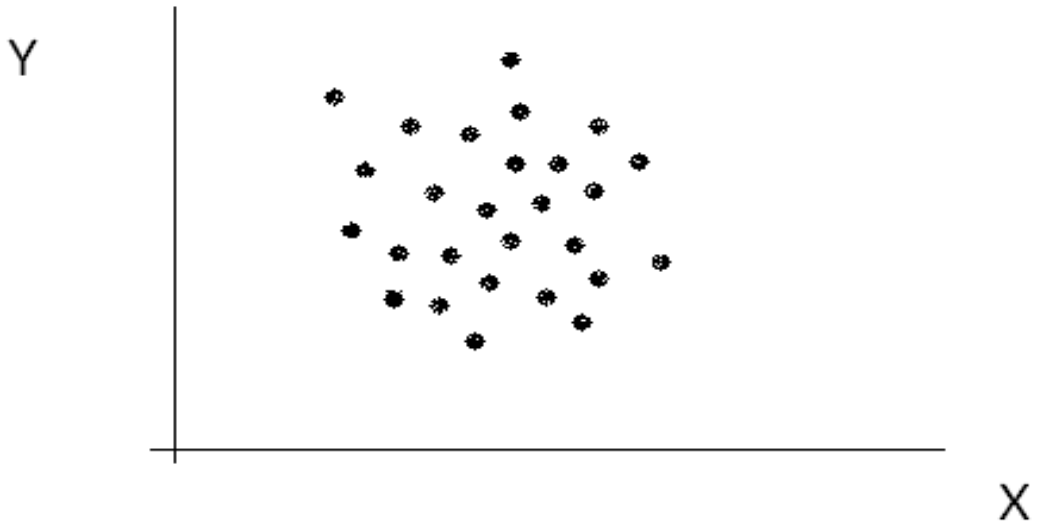
Linear Correlation Coefficient

- Two variables are negatively correlated if they tend to change in the opposite direction e.g. Qd and P.
- Points cluster around a line with a negative slope.
  - If all points lie on the line, correlation is said to be perfect negative i.e. r =-1
  - The diagram below illustrates negative linear correlation



Linear Correlation Coefficient

- Two variables are uncorrelated when they tend to change with no relation to each other. Points are dispersed all over the surface of the XY plane i.e. r = 0



Linear Correlation Coefficient

- In the underlying population form which the sample of points (x_i,y_i) is selected, the population correlation between the variables X and Y. (Greek letter: ρ; read rho)
- The quantifies the strength of the linear relationship between the outcomes *x* and *y*.
- The estimator of ρ is known as Pearson's coefficient of correlation or correlation coefficient (*r*).

Linear Correlation Coefficient

- The sample correlation coefficient is denoted by *r*.

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

or

$$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

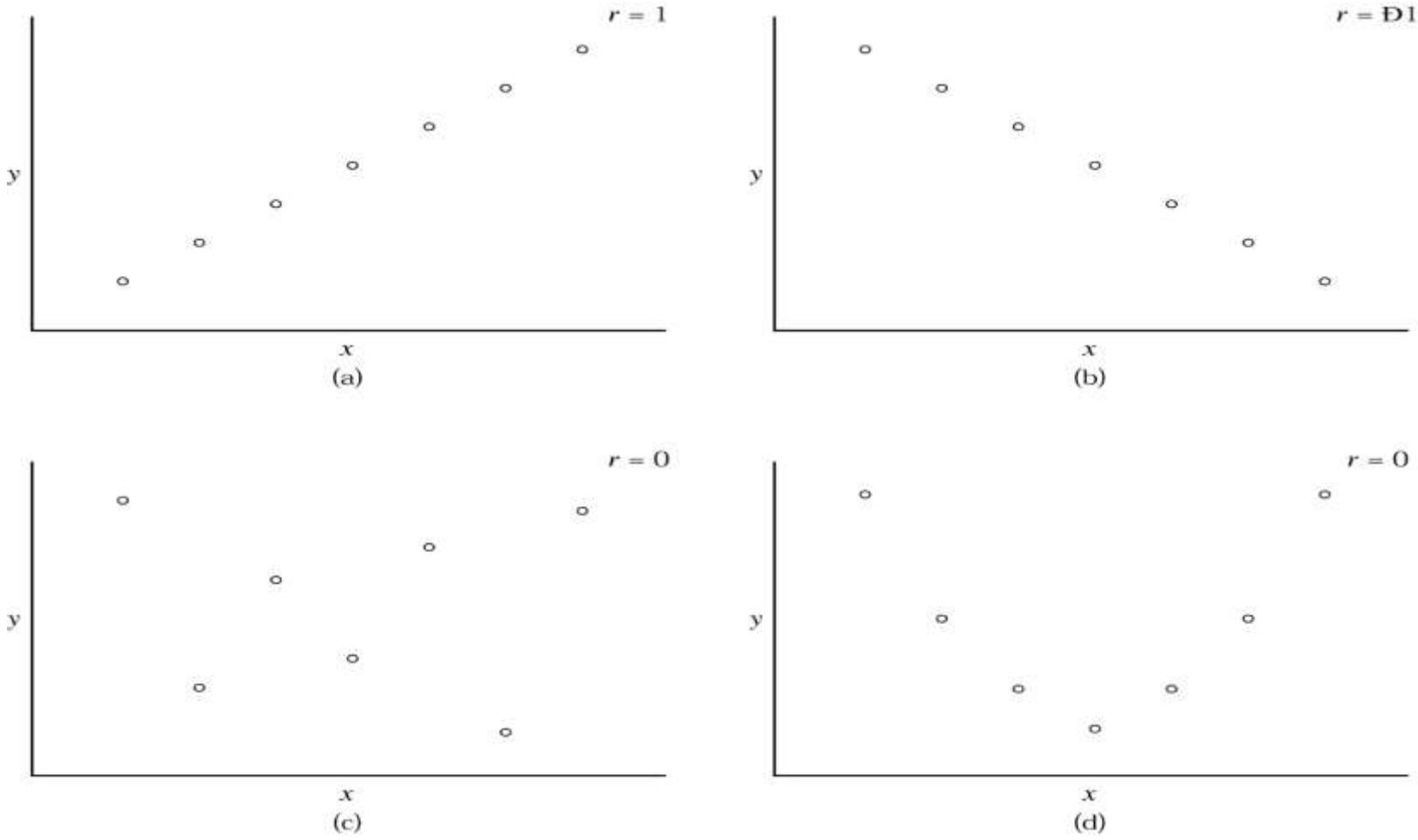
Linear Correlation Coefficient

- The correlation coefficient is dimensionless number; it has no units of measurement.
  - $-1 \leq r \leq 1$
  - The value $r=1$ and $r=-1$ occur when there is an exact linear relationship between x and y. (Figure 17.2 (a)(b))
  - If y tends to increase in magnitude as x increases, r is greater than 0; x any y are said to be positively correlated. ($r>0$)
  - If y decreases as x increases, r is less than 0 and the two variables are *negatively correlated*. ($r<0$)
  - If $r=0$, there is no linear relationship between x and y and the variables are uncorrelated. ($r=0$) (Figure 17.2 (c)(d))

Linear Correlation Coefficient

FIGURE 17.2

Scatter plots showing possible relationships between X and Y



Limitations of the coefficient of correlation r :

1. It quantifies only the **strength of the linear relationship** between two variables.
2. Care must be taken when the data contain any **outliers**, or pairs of observations that lie considerably outside the range of the other data points.
3. The estimated correlation should never be **extrapolated** beyond the observed ranges of the variables; the relationship between X and Y may change outside of this region.
4. A high correlation between two variables **does not imply a cause-and-effect relationship**.

Example 1

The data below relates to expenditure and profits for 8 firms.

Expenditure ('000 Kshs)	9	5	1	4	11	7	2	8
Profit ('000 Kshs)	12	7	2	8	13	9	4	5

Do the following:

- Plot a scatter diagram (include a title and label the axes)
- Calculate the covariance and interpret your result.
- Calculate and interpret the linear correlation coefficient.

Spearman's Rank Correlation Coefficient

- Pearson's correlation coefficient is very sensitive to outlying values. We may be interested in calculating a measure of association that is more robust.
- Pearson's correlation coefficient is also not appropriate when points on a scatter graph seem to follow a curve (i.e. a curvilinear relationship).
- One approach is to rank the two sets of outcomes x and y separately and known as Spearman's rank correlation coefficient.(non-parametric method: data does not have to follow a normal distribution).

Spearman's Rank Correlation Coefficient

- Spearman's rank correlation coefficient is denoted by r_s :

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

  - n: the number of data points in the sample
  - d_i is the different between the rank of x_i and the rank of y_i
- $-1 \le r_s \le 1$
- High degree of correlation between x any y: r_s =-1 or 1
- A lack of linear association between two variables: r_s= 0
- When type of data is ordinal or the conditions do not hold, we should used r_s .

Spearman's rank correlation coefficient r_s :

1. It is much less sensitive to outlying values than Pearson's correlation coefficient.
2. It is used when points on a scatter graph follow a curve i.e. a non-linear (curvilinear) relationship
3. It can be used when one or both of the relevant variables are ordinal.
4. It relies on ranks rather than on actual observations.

Coefficient of Determination, r^2 or R^2 :

- The $\boldsymbol{r^2}$ is the ratio of the explained variation to the total variation.
- $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between *x* and *y*
- It represents the percent of the data that is the closest to the line of best fit.
- For example, if $r = 0.84$, then $r^2 = 0.70$, which means that 70% of the total variation in *y* can be explained by the linear relationship between *x* and *y* (as described by the regression equation). The other 30% of the total variation in *y* remains unexplained.
- $\boldsymbol{R^2}$ is a measure of how well the regression line represents the data

Spearman's rank correlation coefficient Example 2

- Two students are considering applying to the same six universities (A, B, C, D, E, F) to study Economics. Their orders of preference are as follows:

Student 1:	B	E	A	F	D	C
Student 2:	F	C	A	B	D	E

- Calculate Spearman's Rank Correlation Coefficient and interpret your answer.

Solution

We start by presenting the data differently:

University	A	B	C	D	E	F
Student 1, r_x	3	1	6	5	2	4
Student 2, r_y	3	4	2	5	6	1
d						
d^2						

Example 3

- Note 1: If the orders of preferences were:

Student 1:	B	E	A	F	D	C
Student 2:	B	E	A	F	D	C

Example 4

- Note 2: If the orders of preference were:

Student 1:	B	E	A	F	D	C
Student 2:	C	D	F	A	E	B