Chapter 7 Regression and Correlation

Regression

⇨ a study to identify the relationship between two or more variables using a mathematical equation

⇨ is normally used for estimation purposes

Independent Variable / Explanatory Variable

⇨ the variable that is used to explain the variation in the dependent variable

Dependent Variable

⇨ the variable to be predicted or explained

Example 1:

A study on relationship between the sales of ice cream and the temperatures

- temperature is an independent variable since it can be used to explain the sales of ice cream
- sales is a dependent variable since the sales depends on temperature

Scatter Diagram

⇨ a plot of paired observations

⇨ illustrates whether

- ♦ any relationship between the dependent and independent variables
- ♦ the relationship is positive or negative
- ♦ the relationship is linear or non-linear

Simple Linear Regression

⇨ the simplest form of linear relationship between two variables

⇨ $Y = a + bX$

where Y ≡ dependent variable
a ≡ interception of the line at the y-axis
b ≡ regression coefficient
≡ slope or gradient of the line
X ≡ independent variable

Note:
1. b indicates the changes in Y when a unit change in X
2. b is positive $\Rightarrow$ positive linear relationship between X and Y
3. b is negative $\Rightarrow$ negative linear relationship between X and Y

Multiple Linear Regression Equation

⇨ $Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$

where Y ≡ dependent variable

$b_1, b_2, \cdots, b_k$ ≡ regression coefficients

$X_1, X_2, \cdots, X_k$ ≡ *k* independent variables

Univariate distribution

⇨ data of single characteristics is grouped together

⇨ Example: 1. height of student
2. price of items

Bivariate distribution

⇨ data of two characteristics are grouped together

⇨ Example: 1. monthly sales of ice cream and monthly temperatures
2. the sales of product and the advertisement expenses

Least squares method

⇨ the standard technique for obtaining a linear regression line such that the error sum of square is the minimum

Least squares regression line

⇨ the regression line obtained by Least Square method

Fit a set of bivariate data (X, Y) with a straight line $Y = a + bX$.
By Least Squares method,

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

$$a = \overline{Y} - b\overline{X} \quad \text{or} \quad a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

where n ≡ total observations in a set of bivariate data (X, Y)

Notes:
For any set of bivariate data, the least squares regression line of Y on X

1. is used to estimate a value of Y given a value of X
2. passes through the mean point $(\overline{X}, \overline{Y})$ of the data

Example 2:
The following table shows the output at a factory and costs of production over the past 5 months. Find the equation of the least squares regression line.

Month	Output (000's units) / X	Costs (RM'000) / Y
1	20	82
2	16	70
3	24	90
4	22	85
5	18	73

Solution:

X	Y	XY	X^2
20	82		
16	70		
24	90		
22	85		
18	73		
$\Sigma X = 100$	$\Sigma Y = 400$	$\Sigma XY =$	$\Sigma X^2 =$

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

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} =$$

⇨ the regression line is $Y = a + bX =$
where Y = total costs in RM'000; X = output in thousand units

Regression Analysis as a forecasting tool

⇨ Two types of estimation using the regression equation

1. Extrapolation estimate
   - ⇨ Extrapolation ≡ find the value of Y outside the range of X
   - ⇨ most commonly used for forecasting using a time series
   - ⇨ may less accurate and unreliable to a certain extent
2. Interpolation estimate
   - ⇨ Interpolation ≡ find the value of Y within the range of X
   - ⇨ forecasting using interpolation is more accurate and more reliable than using extrapolation

Example 3:

The data in the following table relate the weekly maintenance cost (RM) to the age (in months) of ten machines of similar type in a manufacturing company. Find the least squares regression line of maintenance cost on age and use this to predict the maintenance cost for a machine of this type which is 40 months old.

Machine	1	2	3	4	5	6	7	8	9	10
Age (X)	5	10	15	20	30	30	30	50	50	60
Cost (Y)	190	240	250	300	310	335	300	300	350	395

Solution:

X	Y	XY	X^2
5	190		
10	240		
15	250		
20	300		
30	310		
30	335		
30	300		
50	300		
50	350		
60	395		
$\Sigma X = 300$	$\Sigma Y = 2970$	$\Sigma XY =$	$\Sigma X^2 =$

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

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} =$$

⇨ Least squares regression line of cost on age is $Y =$

When $X = 40$,

$$Y = 212.90 + 2.8033X =$$

⇨ The estimated maintenance cost for a similar machine which is 40 months old is RM325

Regression line (trend line) for a time series data

★ Time-series data is data collected on the same element for the same variable at different points in time or for different periods of time

★ Use least squared method to obtain the regression line (trend line)

★ This is particular useful for purpose of forecasting

Example 4:

The sales of a product in Kuala Lumpur are shown for the years 1996 to 2000

Year	1996	1997	1998	1999	2000
Sales (RM'000)	103	167	214	262	309

Use the method of least squares to find the trend line. Use the line to estimate the sales of 2001

Solution:

Let X=0 for 1996, X=1 for 1997, and so on.

Year	X	X^2	Y	XY
1996				
1997				
1998				
1999				
2000				
	$\Sigma X =$	$\Sigma X^2 =$	$\Sigma Y =$	$\Sigma XY =$

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

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} =$$

⇨ $Y = a + bX =$

To estimate the sales for 2001, substitute $X =$

$Y = 109.6 + 50.7(X) = 363.1$

The estimate of sales in 2001 is RM363100

Correlation

⇨ measure the strength of the relationship between two variables

⇨ involves a bivariate data / distribution

Two types of correlation

1. linear correlation
   ⇨ correlation is said to be linear if $\frac{dy}{dx} =$ constant
2. non-linear correlation (or curvilinear correlation)
   ⇨ correlation is said to be non-linear if $\frac{dy}{dx} \neq$ constant

Positive linear correlation

⇨ An increase in the independent variable (X) will result an increase in the dependent variable (Y)

Negative linear correlation

⇨ An increase in the independent variable (X) will result a decrease in the dependent variable (Y)

Correlation Coefficient (r)

⇨ measure the strength of linear relationship between two variables

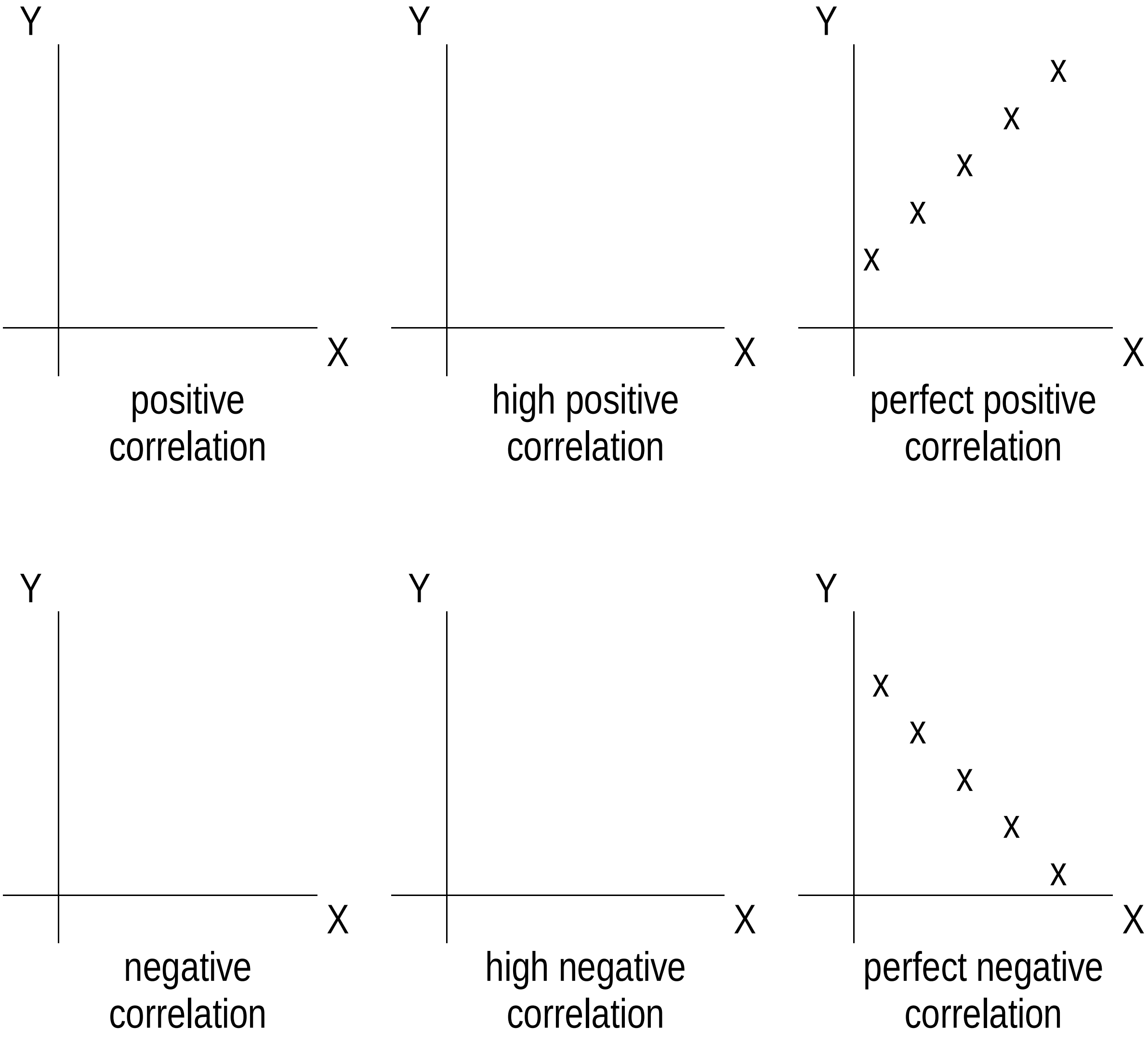
⇨ has a range of values from –1 to +1 i.e. $-1 \le r \le +1$

⇨ if $r = 0$, then there is no linear relationship between the two variables

Degree of correlation	Positive correlation	Negative correlation
Perfect	+1	-1
Very High	$0.8 \leq CC < 1.0$	$-1.0 < CC \leq -0.8$
High	$0.6 \leq CC < 0.8$	$-0.8 < CC \leq -0.6$
Some	$0.4 \leq CC < 0.6$	$-0.6 < CC \leq -0.4$
Low	$0.2 \leq CC < 0.4$	$-0.4 < CC \leq -0.2$
Very Low	$0.0 < CC < 0.2$	$-0.2 < CC < 0.0$
Absent	0	0

$CC \equiv$ Correlation coefficient

Scatter diagrams and correlation



positive correlation

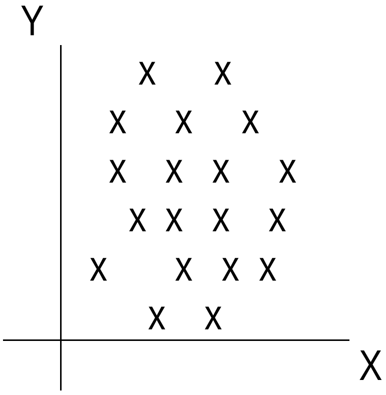
high positive correlation

perfect positive correlation

negative correlation

high negative correlation

perfect negative correlation



No linear relationship

Product Moment Correlation Coefficient for a set of bivariate (X,Y) data

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

where n is the number of pair bivariate (X,Y) values.

Example 5:
Find the product moment correlation coefficient for the following data.

X	2	3	1	4	3	5
Y	9	11	7	13	11	15

Solution:

X	Y	XY	X^2	Y^2
2	9			
3	11			
1	7			
4	13			
3	11			
5	15			
$\Sigma X=18$	$\Sigma Y=66$	$\Sigma XY=$	$\Sigma X^2=$	$\Sigma Y^2=$

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

Example 6:

Calculate the coefficient of correlation for the following data. What does the value of the coefficient indicate?

X	5	6	7	9	8
Y	8	9	9	11	13

Solution:

X	Y	XY	X^2	Y^2
5	8			
6	9			
7	9			
9	11			
8	13			
$\Sigma X = 35$	$\Sigma Y = 50$	$\Sigma XY =$	$\Sigma X^2 =$	$\Sigma Y^2 =$

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

$r = 0.7906$ shows a fairly high positive relationship between X and Y. As X increase, Y would also increase and the agreement between X and Y are fairly well.

Coefficient of Determination (r^2) for Simple Linear Regression

- ⇨ is the square of the coefficient of correlation
- ⇨ indicates the percentage of the total variance in dependent variable (Y) that is explained by the given independent variable (X)
- ⇨ also known as the index of determination
- ⇨ since $-1 \le r \le 1$, it follows that $0 \le r^2 \le 1$

Example 7:

If $r = 0.8$ then $r^2 = 0.8^2 = 0.64$

- ⇨ 64% of the variation in the dependent variable (Y) is explained by the given independent variable (X)
- ⇨ the rest are obviously unexplained and may be due to other factors

Correlation of a time series data

- ⇨ correlation exists if there is a trend line

Example 8:

Sales of a product between 1996 and 2000 were as follows:

Year	1996	1997	1998	1999	2000
Unit sold ('000)	20	18	15	14	11

Is there a trend in sales? In other words, is there any correlation between the year and the number of units sold?

Solution:

Let X=0 for 1996, X=1 for 1997, and so on.

Year	X	Y	XY	X^2	Y^2
1996		20			
1997		18			
1998		15			
1999		14			
2000		11			
	$\Sigma X =$	$\Sigma Y = 78$	$\Sigma XY =$	$\Sigma X^2 =$	$\Sigma Y^2 =$

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

There is a trend in sales. The coefficient of correlation indicates that there is very high negative correlation between year and unit sold. As year increase, the number of units sold decrease.

Rank Correlation

⇨ non-parametric method to measure the correlation between two variables

⇨ measure the correlation based on the ranks of two sets of data (X and Y)

⇨ data are ranked from 1 to n

⇨ Example: 1. Discipline and exam marks
2. Job performance and qualification

Spearman's Rank Correlation Coefficient

⇨ a measure of rank correlation

⇨ can be used even though the variables to be correlated are not representable in numeric form

⇨ $r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$ with $-1 \le r_s \le +1$

where r_s ≡ rank coefficient of correlation
d ≡ difference between two corresponding ranks ($d = r_X - r_Y$)
r_X ≡ rank of X
r_Y ≡ rank of Y
n ≡ number of pairs of observations

Example 9: (data had been ranked)
X and Y were judges at a beauty contest in which there were 10 competitors. Their ranking are shown below.

Competitor	A	B	C	D	E	F	G	H	I	J
X	4	9	2	5	3	10	6	7	8	1
Y	6	10	2	8	1	9	7	4	5	3

Calculate a coefficient of rank correlation between these two ranks and comment briefly on your result.

Solution:

r_X	4	9	2	5	3	10	6	7	8	1	
r_Y	6	10	2	8	1	9	7	4	5	3	
$d = r_X - r_Y$											
d^2											$\Sigma d^2 =$

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

Spearman’s coefficient of rank correlation for the data is 0.7455, indicating that there is some degree of agreement between X and Y.

Example 10: (data had not been ranked)
The following data show the average rent and rates (RM per square feet) for a selection of areas.

Rate (X)	1.68	1.46	1.57	13.37	3.18	1.95	1.07	1.71	1.22	6.46
Rent (Y)	3.81	4.19	4.87	22.85	6.47	6.48	2.66	6.49	5.33	15.23

Calculate Spearman’s rank correlation coefficient to access whether there is any correlation between rate and rent.

Solution:

Rate (X)	Rank of X (r_X)	Rent (Y)	Rank of Y (r_Y)	$d = r_X - r_Y$	d^2
1.68		3.81			
1.46		4.19			
1.57		4.87			
13.37		22.85			
3.18		6.47			
1.95		6.48			
1.07		2.66			
1.71		6.49			
1.22		5.33			
6.46		15.23			
					$\Sigma d^2 =$

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

The result demonstrates relatively high positive correlation. Thus, high rates tend to be paired with high rents and vice versa.

Note:
Ranks are usually allocated in ascending order; rank 1 to the smallest item; rank 2 to the next larger and so on, although it is perfectly feasible to allocate in descending order. However, which method is selected must be used on both variables.

Note:

1. In some cases, it may have a ‘tied’ in the rankings. It may be necessary to rank two or more individuals or entries as equal. In such cases, each individual is given an equal rank.
2. For example, the four numbers 14, 26, 26 and 28 would be allocated ranks 1, 2.5, 2.5 and 4 respectively (since two items have same value of 26, they each must be allocated the average of ranks 2 and 3, i.e. 2.5)

Example 11: (data had tied rank)
The following data relate to the number of vehicles owned per 100 population (X) and road deaths per 100,000 population for 12 countries. Calculate the Spearman's rank correlation coefficient and comment on the result.

X	30	31	32	30	46	30	19	35	40	46	57	30
Y	30	14	30	23	32	26	20	21	23	30	35	26

Solution:

X	r_X	Y	r_Y	$d = r_X - r_Y$	d^2
30		30			
31		14			
32		30			
30		23			
46		32			
30		26			
19		20			
35		21			
40		23			
46		30			
57		35			
30		26			
					$\Sigma d^2 =$

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

The result shows that there is some positive correlation between vehicles owned and number of road deaths, but the relationship is not strong.

Example 12: (data had tied rank)
Calculate the rank correlation coefficient for the following data.

X	92	89	87	86	86	77	71	63	53	50
Y	86	83	91	77	68	85	52	82	37	57

Solution:

X	r_X	Y	r_Y	$d = r_X - r_Y$	d^2
92		86			
89		83			
87		91			
86		77			
86		68			
77		85			
71		52			
63		82			
53		37			
50		57			
					$\Sigma d^2 =$

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

<u>Comparison of rank and product moment correlation</u>

Product moment coefficient

1. The standard measure of correlation
2. Data must be numeric

Rank coefficient

1. Only an approximation to the product moment coefficient
2. Easier to use with less calculations
3. Can be used with non-numeric data
4. Can be insensitive to small changes in actual values