CHAPTER 6 LINEAR REGRESSION AND CORRELATION

Objectives: In business and economic applications, frequently interest is in relationships between two or more random variables, and the association between variables is often approximated by postulating a linear functional form for their relationship.

After working through this chapter, you should be able to:

(i) understand the basic concepts of regression and correlation analyses;

(ii) determine both the nature and the strength of the linear relationship between two variables.

6.1 Introduction

This chapter presents some statistical techniques to analyze the association between two variables and develop the relationship for prediction.

6.2 Curve Fitting

Very often in practice a relation is found to exist between two (or more) variables.

It is frequently desirable to express this relationship in mathematical form by determining an equation connecting the variables.

To aid in determining an equation connecting variables, a first step is the collection of data showing corresponding values of the variables under consideration.

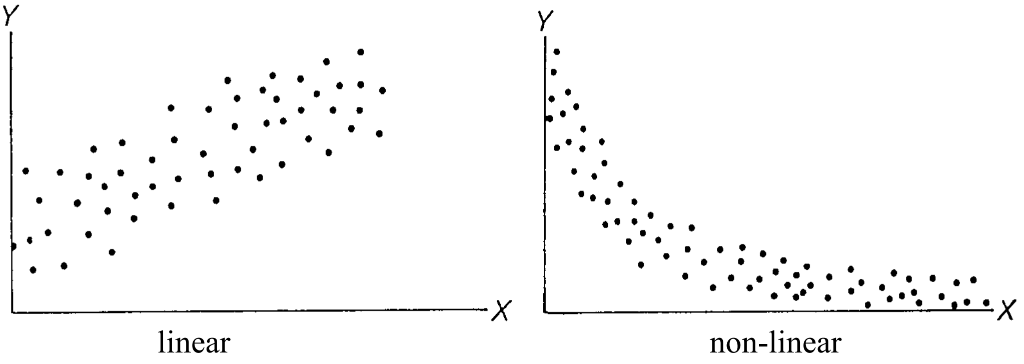

Figure 1. Scatter diagram

6.3 Fitting a Simple Linear Regression Line

To determine from a set of data, a line of best fit to infer the relationship between two variables.

6.3.1 The Method of Least Squares

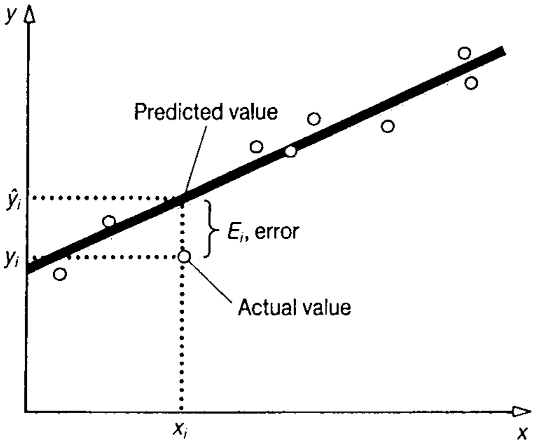

Figure 2. Sample observations and the sample regression line

Determining the line of "best fit":

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

by minimizing $\sum E_i^2$.

To minimize $\sum E_i^2$, we apply calculus and find the following "normal equations":

$$\sum y = na + b\sum x \quad (1)$$
$$\sum yx = a\sum x + b\sum x^2 \quad (2)$$

Solve (1) and (2) simultaneously, we have:

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

Notes:

1. The formula for calculating the slope b is commonly written as

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

which the numerator and denominator then reduce to formulas

$$\begin{aligned}\sum(x-\bar{x})(y-\bar{y}) &= \sum(xy-\bar{x}y-x\bar{y}+\bar{x}\,\bar{y})\\ &= \sum xy-\bar{x}\sum y-\bar{y}\sum x+\sum\bar{x}\,\bar{y}\\ &= \sum xy-n\bar{x}\,\bar{y}-n\bar{x}\,\bar{y}+n\bar{x}\,\bar{y}\\ &= \sum xy-n\bar{x}\,\bar{y}\end{aligned}$$

and

$$\begin{aligned}\sum(x-\bar{x})^2 &= \sum\left(x^2-2x\bar{x}+\bar{x}^2\right)\\ &= \sum x^2-2\bar{x}\sum x+\sum\bar{x}^2\\ &= \sum x^2-2n\bar{x}^2+n\bar{x}^2\\ &= \sum x^2-n\bar{x}^2\end{aligned}$$

respectively; and $a=\bar{y}-b\bar{x}$ is the y-intercept of the regression line.

2. When the equation $\hat{y}=a+bx$ is calculated from a sample of observations rather than from a population, it is referred as a sample regression line.

Example 1

Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue and obtains the following results

Month	Advertising Expenditure	Sales Revenue
	(in $1,000)	(in $10,000)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Find the sample regression line and predict the sales revenue if the appliance store spends 4.5 thousand dollars for advertising in a month.

From the data, we find that

$$n=5,\ \sum x=15,\ \sum y=10,\ \sum xy=37,\ \sum x^2=55.$$

Hence $\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$ and $\bar{y} = \frac{\sum y}{n} = \frac{10}{5} = 2$.

Then the slope of the sample regression line is

$$b = \frac{\sum xy - n\bar{x}\,\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$=$

$=$

and the y-intercept is

$$a = \bar{y} - b\bar{x}$$

$=$

$=$

The sample regression line is thus

$\hat{y}=$

So if the appliance store spends 4.5 thousand dollars for advertising in a month, it can expect to obtain $\hat{y}=$ $=$ ten-thousand dollars as sales revenue during that month.

Example 2

Obtain the least squares prediction line for the data below:

	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
	101	1.2	1.44	121.2	10201
	92	0.8	0.64	73.6	8464
	110	1.0	1.00	110.0	12100
	120	1.3	1.69	156.0	14400
	90	0.7	0.49	63.0	8100
	82	0.8	0.64	65.6	6724
	93	1.0	1.00	93.0	8649
	75	0.6	0.36	45.0	5625
	91	0.9	0.81	81.9	8281
	105	1.1	1.21	115.5	11025
Sum	959	9.4	9.28	924.8	93569

$$b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{10(924.8) - (9.4)(959)}{10(9.28) - (9.4)^2} = \frac{233.4}{4.44} = 52.568$$

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{959}{10} - 52.568\left(\frac{9.4}{10}\right) = 46.486$$

Therefore, $\hat{y} = 46.486 + 52.568x$

Example 3

Find a regression curve in the form $y = a + b\ln x$ for the following data:

x_i	1	2	3	4	5	6	7	8
y_i	9	13	14	17	18	19	19	20

$\ln x_i$	0	0.693	1.099	1.386	1.609	1.792	1.946	2.079
y_i	9	13	14	17	18	19	19	20

$\sum \ln x_i = 10.604$ $\qquad$ $\sum\left(\ln x_i\right)^2 = 17.518$

$\sum y_i = 129$ $\qquad$ $\sum\left(\ln x_i\right)y_i = 189.521$

$$b = \frac{n\sum(\ln x)y - \left(\sum \ln x\right)\left(\sum y\right)}{n\sum(\ln x)^2 - \left(\sum \ln x\right)^2} = \frac{8(189.521) - (10.604)(129)}{8(17.518) - (10.604)^2} = 5.35$$

$$a = \frac{\sum y}{n} - b\frac{\sum \ln x}{n} = \frac{129}{8} - 5.35\left(\frac{10.604}{8}\right) = 9.03$$

Therefore, $\hat{y} = 9.03 + 5.35\ln x$

6.4 Linear Correlation Analysis

Correlation analysis is the statistical tool that we can use to determine the degree to which variables are related.

6.4.1 Coefficient of Determination, r^2

Problem: how well a least squares regression line fits a given set of paired data?

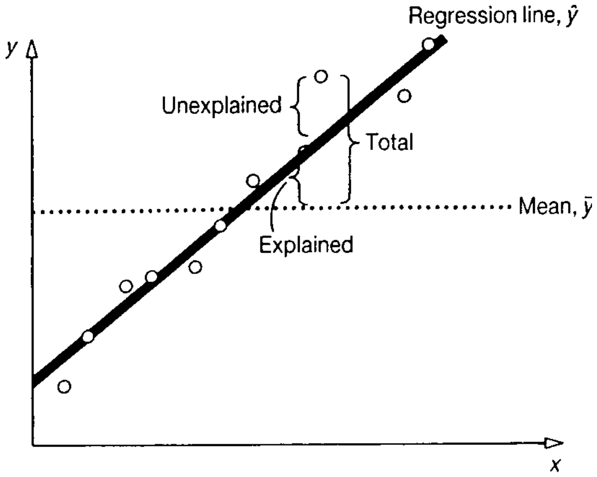

Figure 3. Relationships between total, explained and unexplained variations

Variation of the y values around their own mean = $\sum(y - \bar{y})^2$

Variation of the y values around the regression line = $\sum(y - \hat{y})^2$

Regression sum of squares = $\sum(\hat{y} - \bar{y})^2$

We have:

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2$$

$$\Rightarrow \qquad 1 = \frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2} + \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$$

$$\Rightarrow \qquad \frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2} = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}.$$

Denoting $\frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2}$ by r^2, then

$$r^2 = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}.$$

r^2, the coefficient of determination, is the proportion of variation in *y* explained by a sample regression line.

For example, r^2 =0.9797; that is, 97.97% of the variation in *y* is due to their linear relationship with *x*.

6.4.2 Correlation Coefficient

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

and $-1 \le r \le 1$.

Notes:

The formulas for calculating r^2 (sample coefficient of determination) and r (sample coefficient of correlation) can be simplified in a more common version as follows:

$$r^2 = \frac{\left(\sum (x-\bar{x})(y-\bar{y})\right)^2}{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2} = \frac{\left(\sum xy - n\bar{x}\,\bar{y}\right)^2}{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}$$

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

Since the numerator used in calculating *r* and *b* are the same and both denominators are always positive, *r* and *b* will always be of the same sign. Moreover, if r=0 then b=0; and vice versa.

Example 4

Calculate the sample coefficient of determination and the sample coefficient of correlation for example 1. Interpret the results.

From the data we get

$$n=5,\ \sum x=15,\ \sum y=10,\ \sum xy=37,\ \sum x^2=55,\ \sum y^2=26.$$

Then, the coefficient of determination is given by

$$r^2=\frac{\left(\sum xy-n\bar{x}\,\bar{y}\right)^2}{\left(\sum x^2-n\bar{x}^2\right)\left(\sum y^2-n\bar{y}^2\right)}$$

$$=$$

$$=$$

$$=$$

and

$$r= \qquad\qquad =$$

$r^2=$ implies that of the sample variability in sales revenue is explained by its linear dependence on the advertising expenditure. r = indicates a very strong positive linear relationship between sales revenue and advertising expenditure.

Example 5

Interest rates (x) provide an excellent leading indicator for predicting housing starts (y). As interest rates decline, housing starts increase, and vice versa. Suppose the data given in the accompanying table represent the prevailing interest rates on first mortgages and the recorded building permits in a certain region over a 12-year span.

	Year					
	1985	1986	1987	1988	1989	1990
Interest rates (%)	6.5	6.0	6.5	7.5	8.5	9.5
Building permits	2165	2984	2780	1940	1750	1535
	Year					
	1991	1992	1993	1994	1995	1996
Interest rates (%)	10.0	9.0	7.5	9.0	11.5	15.0
Building permits	962	1310	2050	1695	856	510

(a) Find the least squares line to allow for the estimation of building permits from interest rates.

(b) Calculate the correlation coefficient r for these data.

(c) By what percentage is the sum of squares of deviations of building permits reduced by using interest rates as a predictor rather than using the average annual building permits $\bar{y}$ as a predictor of y for these data?

6.5 Spearman's Rank Correlation

Occasionally we may need to determine the correlation between two variables where suitable measures of one or both variables do not exist.

However, variables can be ranked and the association between the two variables can be measured by r_s:

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)}$$, where d is the difference of rank between x and y.

$-1 \le r_s \le 1$

if r_s closes to 1: strong positive association
if r_s closes to -1: strong negative association
if r_s closes to 0: no association

Notes:

1. The two variables must be ranked in the same order, giving rank 1 either to the largest (or smallest) value, rank 2 to the second largest (or smallest) value and so forth.

2. If there are ties, we assign to each of the tied observations the mean of the ranks which they jointly occupy; thus, if the third and fourth ordered values are identical we assign each the rank of $\frac{3+4}{2} = 3.5$, and if the fifth, sixth and seventh ordered values are identical we assign each the rank of $\frac{5+6+7}{3} = 6$.
3. The ordinary sample correlation coefficient r can also be used to calculate the rank correlation coefficient where x and y represent ranks of the observations instead of their actual numerical values.

Example 6

Calculate the rank correlation coefficient r_s for example 1.

Month (1)	Value x (2)	rank (x) (3)	Value y (4)	rank (y) (5)	d (6)=(3)-(5)	d^2 (7)
1	1	1	1	1.5	-0.5	0.25
2	2	2	1	1.5	0.5	0.25
3	3	3	2	3.5	-0.5	0.25
4	4	4	2	3.5	0.5	0.25
5	5	5	4	5	0	0

By formula

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

$$=$$

$$=$$

$r_s =$ indicates a correlation between the rankings of advertising expenditure and sales revenue. Note that if we apply the ordinary formula of correlation coefficient r to calculate the correlation coefficient of the rankings of the variables in example 6, the result would be slightly different. Since

n=5, $\sum \text{rank}(x)$=15, $\sum \text{rank}(y)$=15, $\sum(\text{rank}(x))(\text{rank}(y))$=54,
$\sum(\text{rank}(x))^2$=55, $\sum(\text{rank}(y))^2$=54,

then r =

which is very close to the result of r_s.

Example 7

Calculate the Spearman's rank correlation, r_s, between x and y for the following data:

y_i	$\text{rank}(y_i)$	x_i	$\text{rank}(x_i)$	$(\text{rank}(y_i) - \text{rank}(x_i))^2$
52		10		
54		14		
47		6		
42		8		
49		6		
38		4		
50		8		
49		8		

Example 8

The data in the table represent the monthly sales and the promotional expenses for a store that specializes in sportswear for young women.

Month	Sales (in $1,000)	Promotional expenses (in $1,000)
1	62.4	3.9
2	68.5	4.8
3	70.2	5.5
4	79.6	6.0
5	80.1	6.8
6	88.7	7.7
7	98.6	7.9
8	104.3	9.0
9	106.5	9.2
10	107.3	9.7
11	115.8	10.9
12	120.1	11.0

(a) Calculate the coefficient of correlation between monthly sales and promotional expenses.

(b) Calculate the Spearman's rank correlation between monthly sales and promotional expenses.

(c) Compare your results from part a and part b. What do these results suggest about the linearity and association between the two variables?

EXERCISE: LINEAR REGRESSION AND CORRELATION

1. The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Find the equation of the regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report but was ill at the time of the final examination.

2. (a) From the following information draw a scatter diagram and by the method of least squares draw the regression line of best fit.

Volume of sales (thousand units)	5	6	7	8	9	10
Total expenses (thousand $)	74	77	82	86	92	95

(b) What will be the total expenses when the volume of sales is 7,500 units?

(c) If the selling price per unit is $11, at what volume of sales will the total income from sales equal the total expenses?

3. The following data show the unit cost of producing certain electronic components and the number of units produced:

Lot size, x	50	100	250	500	1000
Unit cost, y	$108	$53	$24	$9	$5

It is believed that the regression equation is of the form

$$y = ax^b .$$

By simple linear regression technique or otherwise estimate the unit cost for a lot of 400 components.

4. Two variables x and y are related by the law:

$$y = \alpha x + \beta x^2 .$$

State how α and β can be estimated by the simple linear regression technique.

5. Compute and interpret the correlation coefficient for the following grades of 6 students selected at random.

Mathematics grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

6. The following table below shows a traffic-flow index and the related site costs in respect of eight service stations of ABC Garages Ltd.

Site No.	Traffic-flow index	Site cost (in 1000)
1	100	100
2	110	115
3	119	120
4	123	140
5	123	135
6	127	175
7	130	210
8	132	200

(a) Calculate the coefficient of correlation for this data.

(b) Calculate the coefficient of rank correlation.

7. As a result of standardized interviews, an assessment was made of the IQ and the attitude to the employing company of a group of six workers. The IQ's were expressed as whole numbers within the range 50-150 and the attitudes are assigned to five grades labeled 1, 2, 3, 4 and 5 in order of decreasing approval. The results obtained are summarized below:

Employee	A	B	C	D	E	F
IQ	127	85	94	138	104	70
Attitude score	2	4	3	1	2	5

Is there evidence of an association between the two attributes?