Chapter 6

Correlation and linear regression

6.1 Introduction

This chapter concerns itself with relationships between *continuous variables*, such as height and weight, market value and number of transactions, or temperature and sales of ice cream. How would you expect the three pairs of variables listed here to relate to one another? You have already seen (in semester 1) how to present paired data through a **scatter diagram**; this descriptive analysis is now supplemented with more formal techniques. The data consist of n pairs of observations on two variables X and Y:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

These data could have arisen from a random sample of n individuals from a population, or from an experiment in which one variable, usually the X variable, is held fixed or controlled at certain chosen levels and independent measurements of the **response** variable, conventionally Y, are taken at each of these levels. The first step is *always* to plot the data on a scatter diagram.

Example: ice cream sales
Consider the following data (shown overleaf) for ice cream sales at Luigi Minchella's ice cream parlour. Is there any relationship between average temperature and ice cream sales? How would you *describe* this relationship?

Figure 6.1 shows a scatter plot of these data, drawn in `Minitab`. Looking at this diagram, we can immediately see that the two variables are related; in fact, it appears that, as the average temperature increases, sales also increase. We say such a relationship is a **positive** relationship. If there was a "downwards" or "downhill" slope to the scatter diagram, we would say the relationship was **negative**. It looks like we could also draw a straight line through the middle of the data without the points straying too much from this line. Thus, we can also say that the two variables have a **linear** relationship. The more the points stray from this line, the weaker the (linear) relationship between the two variables. So average temperatures and ice cream sales have a *strong*, *positive*, *linear* relationship.

Month	Average Temp (°C)	Sales (£ 000's)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

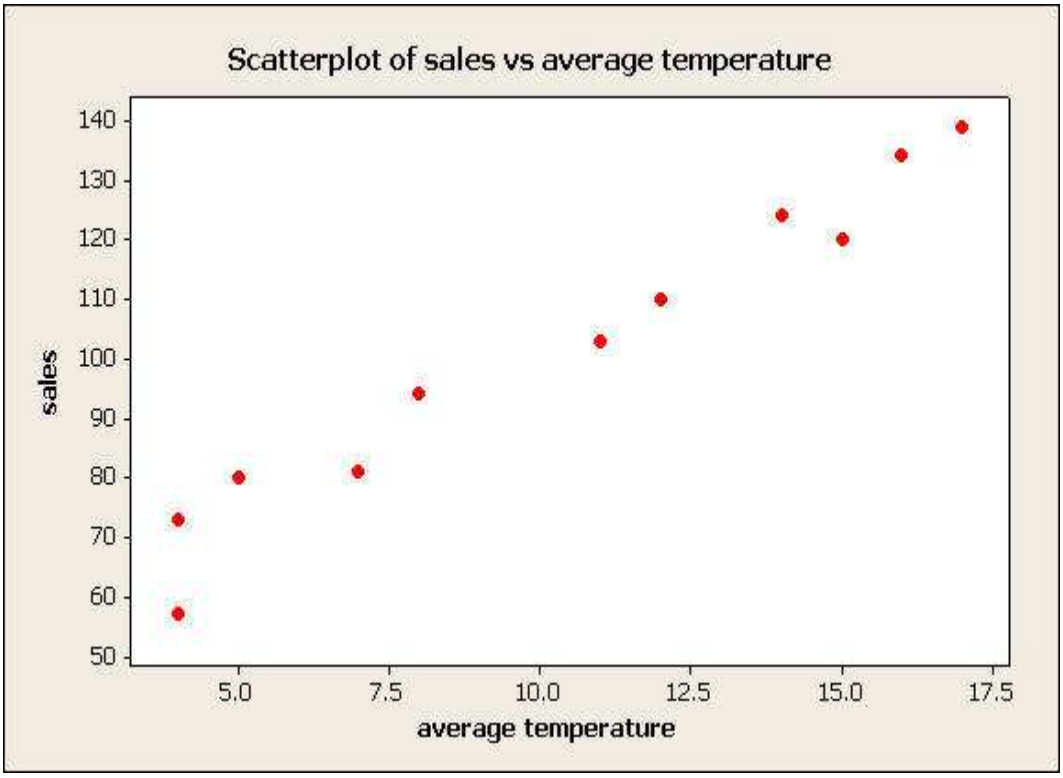

Figure 6.1: Scatter plot showing the relationship between average temperature and ice cream sales

6.2 Correlation

The only measure of association we will be covering here is the **Pearson product moment correlation coefficient**, r. Sometimes r is more briefly referred to as the **sample linear correlation coefficient**. The formula for r is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y}, \\ S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2, \\ S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2. \end{aligned}$$

The sample linear correlation coefficient r always lies between -1 and $+1$. Also, if r is near $+1$, there is a strong *positive* linear relationship between the two variables; if r is near -1 there is a strong *negative* relationship. If r is near zero, there is *no* linear relationship between the variables. Note that this does not imply no relationship at all, simply no *linear* relationship.

Example: ice cream sales
The easiest way to calculate the correlation coefficient between two variables (other than using a computer!) is to draw up a table:

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16	3249	228
7	81	49	6561	567
8	94	64	8836	752
12	110	144	12100	1320
15	124	225	15376	1860
16	134	256	17956	2144
17	139	289	19321	2363
14	124	196	15376	1736
11	103	121	10609	1133
7	81	49	6561	567
5	80	25	6400	400
120	1200	1450	127674	13362

We have a sample size of 12 (not 24!), and so $n = 12$. Thus, using the sums of the first two columns we can find the sample means of temperature (X) and ice cream sales (Y):

$$\begin{aligned} \bar{x} &= \frac{120}{12} \\ &= 10 \qquad \text{and} \end{aligned}$$

$$\begin{aligned} \bar{y} &= \frac{1200}{12} \\ &= 100. \end{aligned}$$

Similarly,

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y} \\ &= 13362 - 12 \times 10 \times 100 \\ \text{‘} &= 13362 - 12000 \\ &= 1362, \end{aligned}$$

$$\begin{aligned} S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2 \\ &= 1450 - 12 \times 10 \times 10 \\ &= 1450 - 1200 \\ &= 250 \qquad \text{and} \end{aligned}$$

$$\begin{aligned} S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2 \\ &= 127674 - 12 \times 100 \times 100 \\ &= 127674 - 120000 \\ &= 7674. \end{aligned}$$

Thus,

$$\begin{aligned} r &= \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} \\ &= \frac{1362}{\sqrt{250 \times 7674}} \\ &= \frac{1362}{1385.099274} \\ &= 0.983 \text{ (to 3 decimal places).} \end{aligned}$$

We have a correlation coefficient of 0.983. Remember, this implies that there is a strong, positive (linear) relationship between average temperature and ice cream sales since the correlation coefficient is very close to +1. This is what we might expect, given the pattern shown in figure 6.1.

Remember, if your calculated correlation coefficient does not lie between -1 and $+1$ then you've done something wrong! You should also check to see if your calculated coefficient agrees with what you can see in the scatter diagram – an "uphill" slope ties in with a positive correlation coefficient, and a "downhill" slope with a negative correlation coefficient. If there appears to be just a random scatter of points, you might expect to get a correlation coefficient which is close to zero. The closer to a straight line the points will lie, then the closer to either $+1$ or -1 the correlation coefficient should be.

Data sets which show non–linear association can be analysed by the technique described above after transforming the data to linearity, or by using **rank correlation** methods which are outside the scope of this course.

6.3 Simple linear regression

A correlation analysis may establish a linear relationship but does not allow us to *use* it to say, predict the value of one variable given the value of another. **Regression analysis** allows us to do this and more. It is also applicable when one of the variables (X) is controlled.

We will assume that the scatter plot of Y versus X shows roughly a straight line and, in addition, that the **spread** in the Y–direction is roughly constant with X.

Look at the scatter plot of ice cream sales against average temperatures. A "line of best fit" can be drawn through the data, and from this line we can make predictions of ice cream sales based on temperature for temperatures which we have no data for. The problem is, everyone's line of best fit is bound to be slightly different! And so everyone's predictions will be slightly different! The aim of regression analysis is to find the very best line which goes through the data in a less subjective way. We do this through the **regression equation**. Before you can understand a regression equation, you need to know what a straight line equation actually is.

6.3.1 Equation of a straight line

Consider the equation

$$Y = 2X.$$

We can draw the line which has this equation quite easily by drawing up a table:

X	0	1	2	3	4	5
$Y = 2X$	0	2	4	6	8	10

In the space below, plot the line which has the equation $Y = 2X$.

You probably noticed that we only needed two points to draw the line; this is true for *all* straight line equations. In the space below, plot the line which has the equation $Y = 3 + 4X$.

There might be many equations which produce lines which pass through the ice cream data; however, in regression analysis, we want to find the equation which gives the "best" line, i.e. the equation which gives a straight line lying closer to the data than any other equation.

6.3.2 The regression equation

We assume a simple linear regression model for the data (and for the population from which the data have been drawn) in which

$$Y = \alpha + \beta X + \epsilon$$

where Y is the **response** variable, X is the **explanatory** variable, and ϵ ("epsilon") is a random error with zero mean and constant variance. The unknown parameters α ("alpha") and β ("beta") represent the intercept and slope of the population regression line $\alpha + \beta X$. Obviously, we need to find α and β; the best values will minimise the gaps between the regression line and the data. These "gaps" are known as the **residuals**. The values of α and β which give rise to the "best" regression line, i.e. the line which minimises the residuals, are

$$\begin{aligned} \hat{\beta} &= \frac{S_{XY}}{S_{XX}} \qquad \text{and} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}, \end{aligned}$$

where S_{XY} and S_{XX} are as before. The "hats" on α and β are there to remind ourselves that we have *estimated* α and β using our sample data; these estimates will change from sample to sample. Since the error term (ϵ) is assumed to have zero mean, in practice we don't estimate this and just ignore it in any further analysis.

Example: ice cream sales
We now use simple linear regression to fit a regression line through the ice cream sales data. Remember, the equation of the regression line is

$$Y = \alpha + \beta X + \epsilon,$$

where we can estimate α and β using

$$\begin{aligned} \hat{\beta} &= \frac{S_{XY}}{S_{XX}} \qquad \text{and} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}. \end{aligned}$$

Notice that $\hat{\beta}$ must be found first, since we need $\hat{\beta}$ in order to calculate $\hat{\alpha}$. Thus,

$$\begin{aligned} \hat{\beta} &= \frac{1362}{250} \\ &= 5.448 \qquad \text{and} \\ \hat{\alpha} &= 100 - 5.448 \times 10 \\ &= 100 - 54.48 \\ &= 45.52. \end{aligned}$$

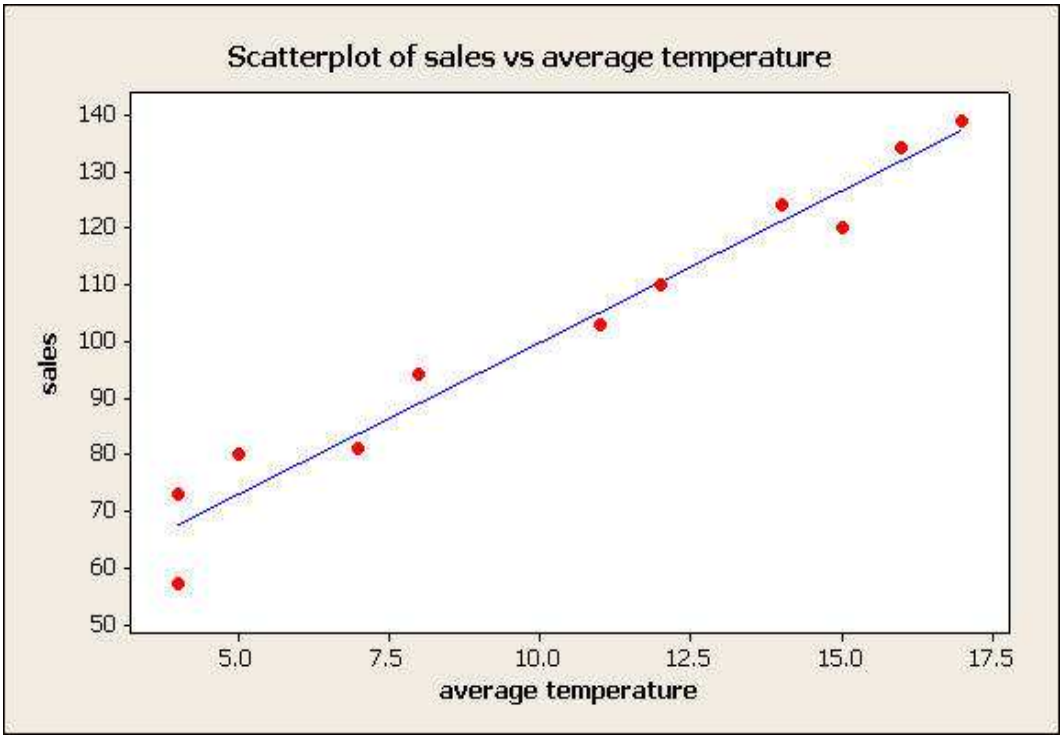

Figure 6.2: Scatter plot showing the relationship between average temperature and ice cream sales, with superimposed regression line

Thus, the regression equation is

$$Y = 45.52 + 5.448X + \epsilon.$$

The scatter plot in figure 6.2 has this regression line superimposed. We can use this regression equation to predict ice cream sales for a given temperature. For example, if we were interested in knowing what the sales would be if the monthly average temperature was 10°C, we can either

(1) take a reading from the graph, or

(2) substitute 10 into our regression equation and solve for Y.

The second approach is probably better, since the accuracy of a reading from your graph depends on how accurately you have drawn your graph in the first place. Thus, using our regression equation, the number of sales we can expect if the average temperature is 10°C is

$$\begin{aligned} Y &= 45.52 + 5.448 \times 10 \\ &= 45.52 + 54.48 \\ &= 100, \end{aligned}$$

i.e. £100,000.

You should only use your regression line to make predictions within the range of the observed data; we cannot be certain that an association between the two variables will continue in the future, and even if it does, it might not be linear.

6.4 Extensions

Regression is a very flexible tool and there are a range of extensions that can be investigated; for example, non–linear regression for situations where a linear relationship between the two variables might not be justified. Multiple regression can be used when there is more than one predictor variable. There are also techniques for using regression on ordinal data, which are particularly useful when considering survey results. The calculations for these more complex regressions are beyond the scope of this course, and are invariably performed by a computer.

6.5 Exercises

1. Consider the following data for a company's monthly advertising expenditure and their sales.

   (a) Produce a scatter plot for these data, and comment on the relationship between advertising and sales.

   (b) Calculate the sample correlation coefficient. Does this agree with what you can see in your plot in part (a)?

   (c) Perform a linear regression analysis on these data, and obtain the linear regression equation.

   (d) Plot the regression line on your scatter diagram in part (a).

   (e) If the company were to spend £112,000 on advertising in a month, what could we expect their sales to be?

Month	Advertising (£000's)	Sales (£ Millions)
January	100	11.2
February	90	12.1
March	110	13.2
April	120	15.1
May	115	14.2
June	95	10.2
July	105	12.5
August	130	16.6
September	118	14.8
October	100	10.8
November	115	11.2
December	128	15.9

2*. (a) In this chapter we have looked at the **Pearson Product Moment Correlation Coefficient**; however, this is just one of many correlation coefficients we could use to quantify the strength of linear dependence between two variables. Write down the name of one other correlation coefficient, give its formula, and explain how it differs to the Pearson Product Moment Correlation Coefficient.

   (b) What regression models could be used for data which do not display a linear association?

* Prize question