Chapter 7

Techniques of regression

What we'll cover...

- Visualising **bivariate data**
- Assessing the **association** between two continuous variables:
  - Informally
  - More objectively – the sample **correlation coefficient**
- Assessing the **signficance** of the correlation coefficient
- Modelling the relationship between two continuous variables: **Simple linear regression**
- Extensions to the **multiple linear regression** model, and possibly beyond...

7.1 Introduction

In this chapter we will investigate relationships between **continuous variables**.

Initially, we will assume our data consists of n pairs of observations on two variables, say X and Y:

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

- Could be a random sample of n individuals from a population, on which two measurements/observations (x_i, y_i), $i = 1, \ldots, n$ were made
- Could be from an experiment in which one variable (X) is held fixed or controlled at certain chosen levels and independent measurements of the **response** variable (Y), are taken at each of these levels

7.1 Introduction

The first step is *always* to plot the data on a **scatter diagram**.

We considered scatter diagrams, often called **scatter plots**, in Chapter 4 (Section 4.5.3, page 92).

From this diagram we can get an initial impression about the relationship between X and Y and form some subjective assessments.

7.1 Introduction

The main aim of this Chapter is to supplement such descriptive analyses with more formal techniquesb — in terms of

- **quantifying** the association between X and Y, and
- **modelling** any relationship between X and Y

7.1 Introduction

Towards the end of the course, we will also consider extending these techniques to investigate the relationship between the response variable Y and more than one **predictor** variable – perhaps to include several predictor variables $X_1, X_2, \ldots$.

7.2 Example: The *Saint Clair Estate Winery*

The *Saint Clair Estate Winery* is a vineyard in the Marlborough region of the South Island of New Zealand.

- **Multi–million** dollar industry
- *Saint Clair* one of the country's leading producers/exporters, of wine

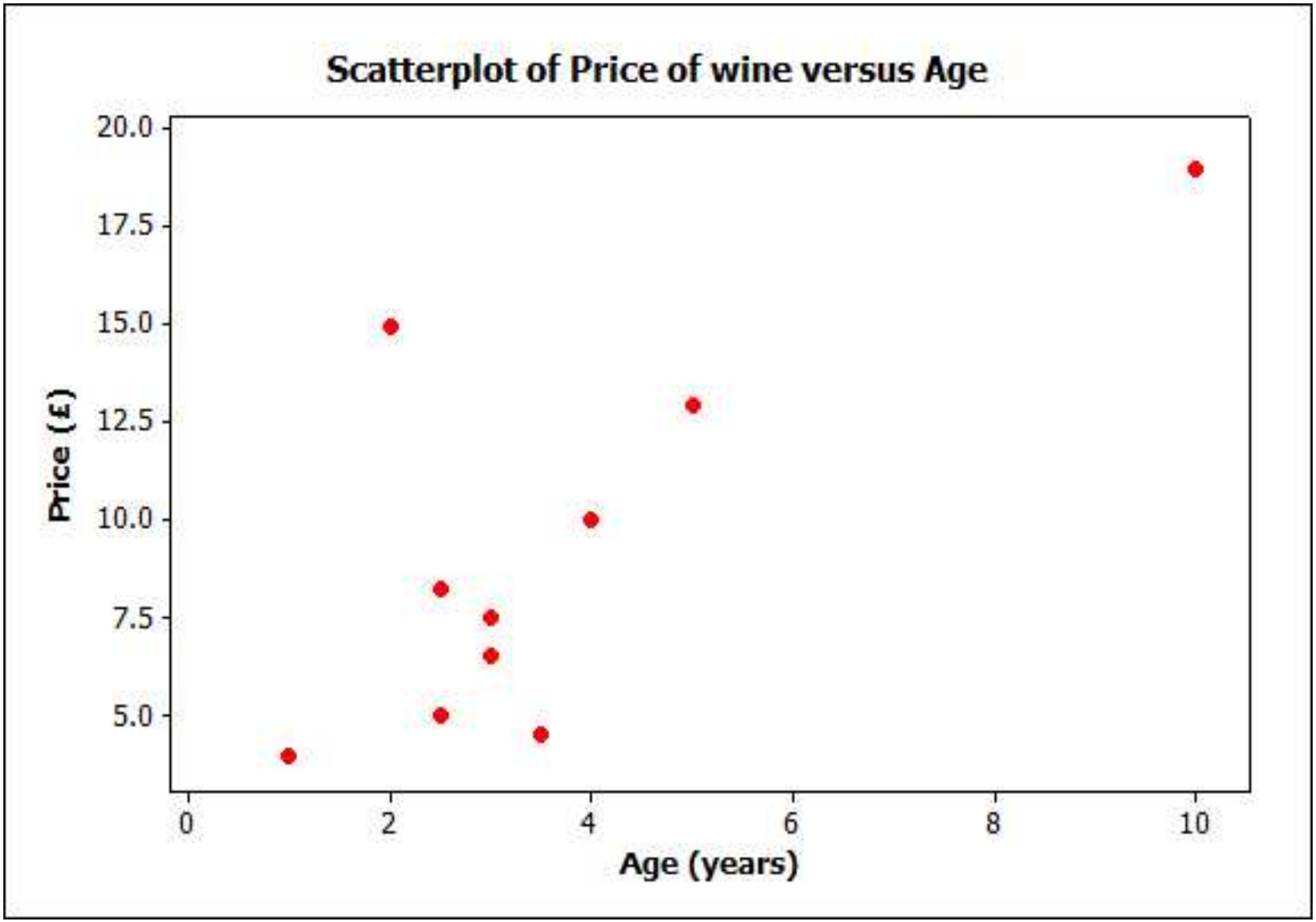
7.2 Example: The *Saint Clair Estate Winery*

The price of a bottle of wine is thought to depend on many factors, such as its age, the quality of the grapes used to produce it, the amount of rainfall during the growing season, where the wine was produced, etc.

The table below shows the price of 10 randomly selected bottles of wine from www.tanners-wines.co.uk, an online wine merchant. Also shown is the age of each wine selected.

Bottle	1	2	3	4	5	6	7	8	9	10
Age (X)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Price (Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00

7.2 Example: The *Saint Clair Estate Winery*



7.2 Example: The *Saint Clair Estate Winery*

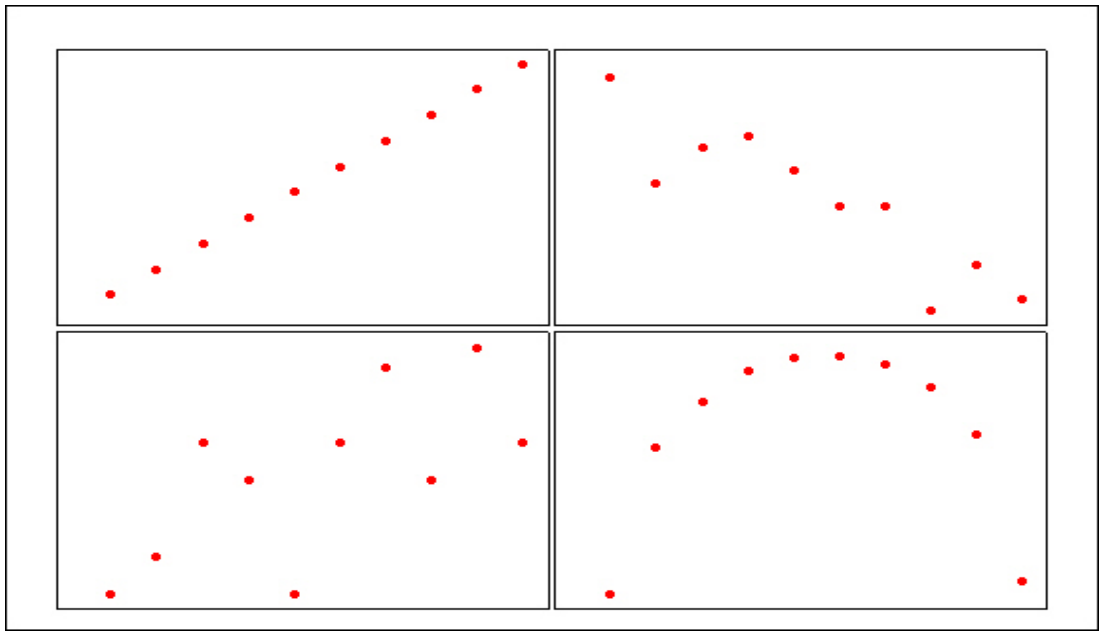
Looking at the scatter plot (and maybe just the raw data themselves!), what can you say about the relationship between age of wine and price?

- *Generally*, as the age of wine increases, the price also increases
- There is a **linear** relationship
- There is a **strong** linear relationship? Or maybe **moderate**?
- There is a **positive correlation**

7.3 Quantifying the relationship: Correlation

There is clearly a relationship between the age and price of wine; the relationship is **strong**, **positive** and **linear**.

How would you describe, in words, the relationship between X and Y in the following scatter plots?



7.3 Quantifying the relationship: Correlation

Scatterplots such as the one in the bottom left–hand corner can be difficult to interpret using words alone, since different people might say different things.

Some might think there is a moderate/fairly strong relationship between X and Y here, whilst others might conclude that there is a relatively weak relationship between these two variables.

Interpreting such relationships with words alone can be subjective; **quantifying** such relationships **numerically** can circumvent this problem of subjectivity.

7.3 Quantifying the relationship: Correlation

One way of doing this is to calculate the **product moment correlation coefficient**, often denoted by the letter r.

The formula for r is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}},$$

where

$$\begin{aligned} S_{xy} &= \left(\sum xy\right) - n\bar{x}\bar{y}, \\ S_{xx} &= \left(\sum x^2\right) - n\bar{x}^2 \qquad \text{and} \\ S_{yy} &= \left(\sum y^2\right) - n\bar{y}^2, \end{aligned}$$

n is the number of pairs and $\bar{x}$ and $\bar{y}$ correspond to the mean of X and the mean of Y (respectively).

7.3 Quantifying the relationship: Correlation

The correlation coefficient r always lies between -1 and $+1$.

- If r is close to $+1$, there is a strong *positive* linear relationship
- If r is close to -1 there is a strong *negative* relationship
- If r is close to zero, there is *no* linear relationship between the variables.

Note that $r \approx 0$ does not imply no relationship at all, simply no *linear* relationship.

Can you estimate the value of r for the wine age/price data? And for the four datasets shown on page 169?

7.3 Quantifying the relationship: Correlation

Using the wine price/age data, we can calculate the value of r. The easiest way to do this is to draw up a table:

x	y	x^2	y^2	xy
3.5	4.50	12.25	20.25	15.75
5	12.95	25	167.7025	64.75
3	6.50	9	42.25	19.5
2.5	4.99	6.25	24.9001	12.475
3	7.50	9	56.25	22.5
2	14.95	4	223.5025	29.9
2.5	8.25	6.25	68.0625	20.625
1	3.95	1	15.6025	3.95
10	18.99	100	360.6201	189.900
4	10.00	16	100	40
36.5	**92.58**	**188.75**	**1079.14**	**419.35**

Then we have:

$$\begin{aligned} \bar{x} &= \frac{36.5}{10} \\ &= 3.65 \qquad \text{and} \\ \bar{y} &= \frac{92.58}{10} \\ &= 9.258. \end{aligned}$$

We can now calculate S_{xy}, S_{xx} and S_{yy}:

$$\begin{aligned} S_{xy} &= \left(\sum xy\right) - n\bar{x}\bar{y} \\ &= 419.35 - 10 \times 3.65 \times 9.258 \\ &= 81.433, \end{aligned}$$

$$\begin{aligned} S_{xx} &= \left(\sum x^2\right) - n\bar{x}^2 \\ &= 188.75 - 10 \times 3.65 \times 3.65 \\ &= 55.525 \qquad \text{and} \end{aligned}$$

7.3 Quantifying the relationship: Correlation

$$\begin{aligned} S_{yy} &= \left(\sum y^2\right) - n\bar{y}^2 \\ &= 1079.14 - 10 \times 9.258 \times 9.258 \\ &= 222.0344. \end{aligned}$$

7.3 Quantifying the relationship: Correlation

Thus,

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} \\ &= \frac{81.433}{\sqrt{55.525 \times 222.0344}} \\ &= \frac{81.433}{111.0336} \\ &= 0.7334. \end{aligned}$$

7.3 Quantifying the relationship: Correlation

Since this is fairly close to +1, we have a moderate/strong positive linear association between the age and price of wine.

Remember that this correlation coefficient can only be used to detect **linear** associations.

For information, the value of r for the plots on page 169, from top–left and moving clockwise, is:$r = 1$, -0.899, 0.699 and 0.064.

Note there is clearly a relationship between X and Y in the bottom–right plot, but here $r = 0.064$ which is very close to zero: this is because the relationship here is plainly non–linear.

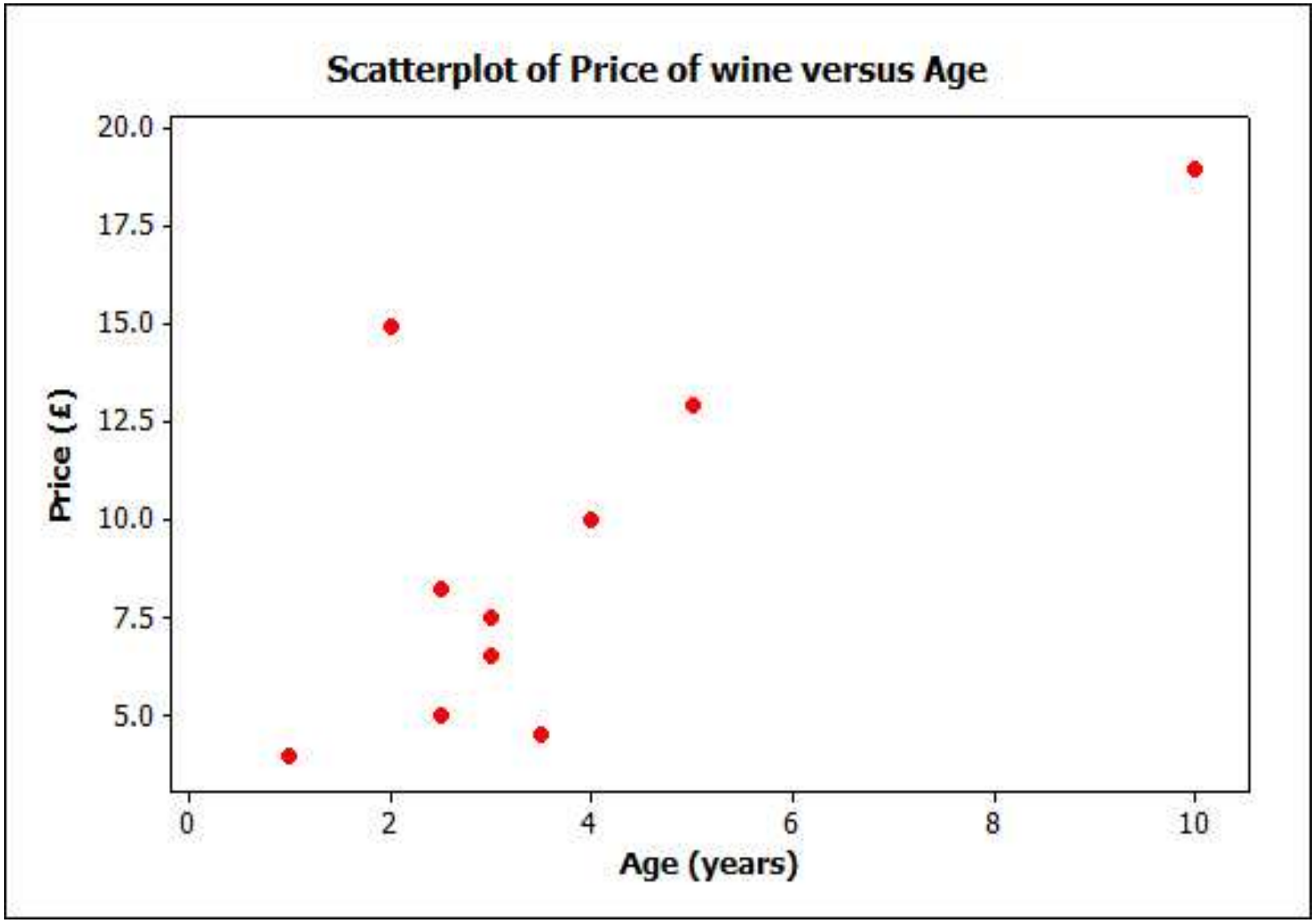
7.4 Modelling the relationship: linear regression

A correlation analysis may establish a linear relationship but it 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.

Look at the scatter plot of the price of wine against the corresponding age of each bottle.

7.4 Modelling the relationship: linear regression



7.4 Modelling the relationship: linear regression

A “line of best fit” can be drawn through the data, and from this line we can make predictions of price based on age.

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 completely objective way.

We do this through the **regression equation**.

7.4 Modelling the relationship: linear regression

Recall from Chapter 1 that the equation of a straight line takes the general form

$$y = mx + c,$$

where m is the **gradient** and c is the **intercept**.

Statisticians tend to use different notation for their **regression equation**:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where Y is the **response** variable and X the **predictor** variable, and ϵ ("epsilon") is a "random error" term.

7.4 Modelling the relationship: linear regression

The unknown parameters β_0 ("beta nought") and β_1 ("beta one") represent the intercept and slope of the population regression line $\beta_0 + \beta_1 X$.

Obviously, we need to find β_0 and β_1; the best values will minimise the vertical 'gaps' between the regression line and the data. These 'gaps' are known as the **residuals** – see diagram.

7.4 Modelling the relationship: linear regression

Each of the points i, $i = 1, \ldots, n$, in our scatter plot has a y co–ordinate y_i.

Recall from above that the corresponding y co–ordinates of points on the line, say $\hat{y}_i$, are given by

$$\hat{y}_i = \beta_0 + \beta_1 x_i.$$

Thus, the vertical "gaps" between the points and the line are given by

$$\begin{aligned} y_1 - \hat{y}_1 &= y_1 - \beta_0 - \beta_1 x_1 \\ y_2 - \hat{y}_2 &= y_2 - \beta_0 - \beta_1 x_2 \\ &\vdots \\ y_n - \hat{y}_n &= y_n - \beta_0 - \beta_1 x_n \end{aligned}$$

7.4 Modelling the relationship: linear regression

Some of these "gaps" will be negative, as defined above, as some points will lie below the line.

To get rid of any "negative gaps", we square all of these quantities:

$$
\begin{aligned}
&(y_1 - \beta_0 - \beta_1 x_1)^2 \\
&(y_2 - \beta_0 - \beta_1 x_2)^2 \\
&\qquad \vdots \\
&(y_n - \beta_0 - \beta_1 x_n)^2
\end{aligned}
$$

7.4 Modelling the relationship: linear regression

The very best line of best fit – that is, the line which minimises the sum of these "squared gaps", is what we call the **regression line**.

So we want the regression line to minimise the quantity

$$\Delta(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

And how do we minimise a function? We use calculus! So we solve

$$\begin{aligned} \frac{d}{d\beta_0} \Delta(\beta_0, \beta_1) &= 0 \\ \frac{d}{d\beta_1} \Delta(\beta_0, \beta_1) &= 0 \end{aligned}$$

for β_0 and β_1.

7.4 Modelling the relationship: linear regression

Doing so gives some very nice formulae:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where the "hat" notation $\widehat{\ }$ implies that we obtain **estimates** of the gradient and intercept from our sample data, and not the **true** values of these parameters.

Estimate the regression equation for the wine data, and superimpose this on the original scatter plot.

7.4 Modelling the relationship: linear regression

For the wine data, we have

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
&= \frac{81.433}{55.525} \\
&= 1.467 \qquad \text{and} \\
\\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= 9.258 - 1.467 \times 3.65 \\
&= 3.903.
\end{aligned}$$

7.4 Modelling the relationship: linear regression

Thus, the regression equation is

$$Y = 3.903 + 1.467X + \epsilon.$$

To plot this line (page 172), we just need to co–ordinates of two points. We know one of them: $(0, 3.9)$. For another, choose any x–value – say $x = 8$:

$$Y = 3.903 + 1.467 \times 8 = 15.639 \longrightarrow (8, 15.6).$$

Modelling the relationship: simple linear regression

Modelling the relationship: simple linear regression

We can use the estimated regression equation to make predictions of wine price given a certain age.

For example, suppose we produce a bottle of wine that has been ageing for $4\frac{1}{2}$ years. How much should we sell it for?

Based on our regression equation, we could estimate a selling price per bottle as:

$$\begin{aligned} Y &= 3.903 + 1.467 \times 4.5 \\ &= 10.505, \end{aligned}$$

i.e. about £10.50.

A cautionary note (I)

As far as possible, avoid **extrapolation**.

- Only use the regression equation to make predictions using X–values that lie within the range of the data observed
- So we should not use our regression equation to estimate the selling price of a bottle of wine that has been ageing for 12 years
- We cannot be sure that this linear association will continue beyond the range of our data!

A cautionary note (II)

Also care should be taken not to read too much into the regression equation.

For example, consider sales of ice cream and sales of sun tan lotion.

- In hot weather sales of ice cream increase and sales of sun tan lotion also increase
- So ice cream sales may be a useful predictor of sun tan lotion sales
- However, the act of buying an ice cream does not *cause* someone to by some sun tan lotion...
- ...What is happening is that both ice cream sales and sun tan lotion sales are directly influenced by a third factor: in this case, the weather.

7.5 Testing the strength of a correlation

In Section 7.3 we thought about how we can quantify the strength of (linear) association between a pair of variables X and Y.

We then moved on, in Section 7.4, to think about how we can **model** this relationship through the simple linear regression model.

Surely, though, there is no point in estimating the regression equation if there is little, or no, linear association between X and Y?

That is, if the correlation coefficient is close zero, we have shown that the strength of linear association is negligible and so why would we then be interested in modelling this negligible/non–existent relationship?

Example 7.1

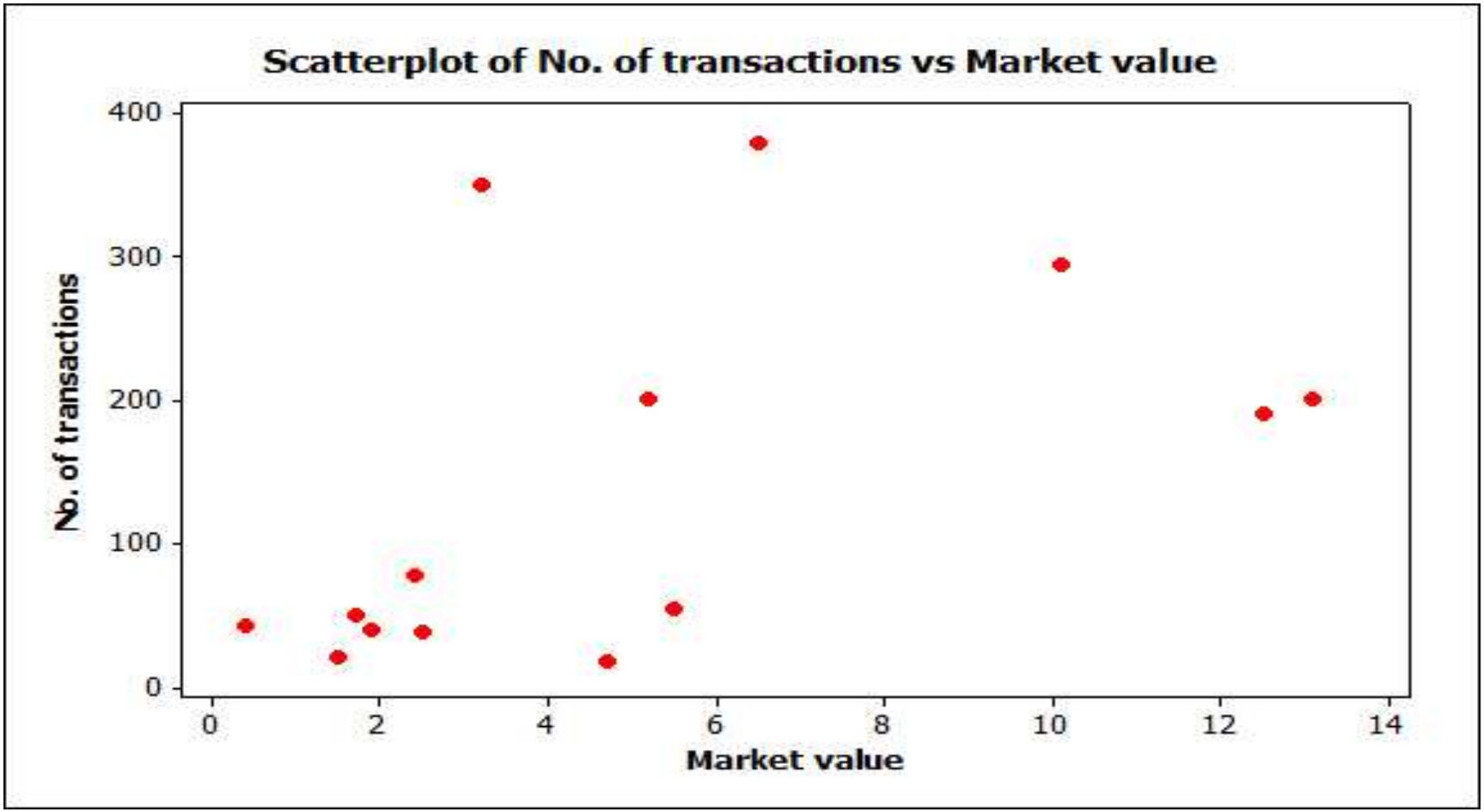
The following table shows the total market value of 14 companies (in £million) and the number of stock exchange transactions in that company's shares occurring on a particular day. Underneath, you are given some numerical summaries; the graph below shows a scatter plot for these data, as produced by `Minitab`.

Market value	6.5	5.2	0.4	1.7	1.9	2.4	3.2	4.7	10.1	12.5	13.1	5.5	2.5	1.
Transactions	380	200	42	50	40	78	350	18	295	190	200	55	38	2

$$\sum_{i=1}^{14} x_i = 71.2 \qquad \sum_{i=1}^{14} y_i = 1956$$

$$\sum_{i=1}^{n} x_i^2 = 582.66 \qquad \sum_{i=1}^{14} y_i^2 = 487166 \qquad \sum_{i=1}^{14} x_i y_i = 13481.6$$

Example 7.1


Scatterplot of No. of transactions vs Market value
No. of transactions
400
300
200
100
0
0
2
4
6
8
10
12
14
Market value

Example 7.1

(a) Find the sample correlation coefficient r, and comment.

(b) Formally test the strength of correlation as suggested by your answer to part (a).

Example 7.1(a): Solution

We have

$$S_{xx} = 582.66 - 14 \times \left(\frac{71.2}{14}\right)^2 = 220.55714$$

$$S_{yy} = 487166 - 14 \times \left(\frac{1956}{14}\right)^2 = 213884.8571$$

$$S_{xy} = 13481.6 - 14 \times \left(\frac{71.2}{14}\right) \times \left(\frac{1956}{14}\right) = 3533.942857$$

So

$$r = \frac{3533.942857}{\sqrt{220.55714 \times 213884.8571}} = 0.515.$$

Comment: Moderate, positive linear correlation.

Example 7.1(b): Solution

We have

$$H_0 \quad : \quad \rho = 0 \qquad \text{(i.e. no correlation in the population)}$$
$$H_1 \quad : \quad \rho \neq 0 \qquad \text{(i.e. there } \textit{is} \text{ a correlation!)}$$

The test statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.515\sqrt{\frac{12}{1-0.515^2}} = 2.08.$$

To get the p–value, we use t tables (see **Page 156**, Chapter 6) with $\nu = n - 2 = 14 - 2 = 12$, giving:

p–value	10%	5%	1%
Critical value	1.782	2.179	3.055

So here, p lies between 5% and 10%.

Example 7.1(b): Solution

Thus

- There is **slight** evidence against H_0
- We **retain** H_0
- The correlation observed in the sample is not strong enough to suggest a true correlation between X and Y in the population

7.6 Multiple linear regression

In this section we will show how the linear regression model can be **extended** to include any number of predictor variables.

The model we have considered so far, namely

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

has been, and is often, referred to as the **simple** linear regression model, because it only involves a single predictor variable.

However, frequently two or more predictor variables may be useful together to predict Y.

7.6 Multiple linear regression

For instance, the sales of a product may depend on

- the product's unit price, **and**
- the amount of advertising expenditure **and**
- the price of a competing product

The simple linear regression model can be extended to include any number of predictor X variables, in which case it is called the **multiple linear regression model**.

7.6.1 Back to the simple linear regression model

Recall the example relating age of wine (X years) to price of wine (Y pounds).

By hand, we found the equation of our regression line (the very best line of best fit!) to be

$$Y = 3.903 + 1.467X + \epsilon.$$

Is **Age** an important predictor of **price**?

7.6.1 Back to the simple linear regression model

We can attempt to answer this by testing

$$\begin{aligned} H_0 &: \beta_1 = 0 \qquad \text{versus} \\ H_1 &: \beta_1 \neq 0. \end{aligned}$$

We will use `Minitab` here!

7.6.1 Back to the simple linear regression model

Notice that $p = 1.6\%$, and so

- We have **moderate** evidence against H_0
- We should **reject** H_0 in favour of H_1
- β_1 is significantly different from zero, and so Age **is** an important predictor of price!

7.6.2 Extending the simple linear regression model

Bottle	1	2	3	4	5	6	7	8	9	10
Price (£Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00
Age (X_1 years)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rainfall (X_2 mm)	126	121	125	106	107	112	124	105	116	108
Temp. (X_3°C)	16	20	17	18	18	22	19	15	21	20

7.6.2 Extending the simple linear regression model

A multiple linear regression model that may be suitable simply extends on the simple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon;$$

as before, ϵ is our "random error" term, and β_0, β_1, β_2 and β_3 are parameters in the model that we need to estimate.

7.6.2 Extending the simple linear regression model

So how do we find $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ – the estimated parameters of the model?

We *can* compute these by hand, as we did in Section 7.4 for the simple linear regression model, but this requires knowledge of matrix algebra which many of you won't have (even if you did A Level maths!).

Anyway, `Minitab` can perform the calculations for us!

7.6.2 Extending the simple linear regression model

Thus, the full (multiple) regression model is:

$$Y = -22.5 + 0.807X_1 - 0.0004X_2 + 1.55X_3 + \epsilon,$$

where X_1, X_2 and X_3 represent the age of a bottle of wine, the total rainfall during the growing season and the corresponding average afternoon temperature (respectively).

7.6.2 Extending the simple linear regression model

The estimated coefficients of the model indicate the **direction** of the relationship between the price of a bottle of wine and each of the corresponding predictors.

For example:

- Since $\hat{\beta}_1 = 0.807$ is positive – positive relationship between age and price (i.e. generally, older wines are more expensive);
- Since $\hat{\beta}_2 = -0.0004$ is negative – negative relationship between rainfall and price (i.e. generally, wines from regions with higher rainfall are cheaper);
- Since $\hat{\beta}_3 = 1.55$ is positive – positive relationship between temperature and price (i.e. generally, wines from regions with higher temperatures are more expensive).

7.6.3 Testing the importance of our predictor variables

Recall Section 7.6.1, where we used `Minitab` to test the significance of the parameter β_1 in our model.

The null hypothesis here was $H_0 : \beta_1 = 0$; retention of this hypothesis would imply that the predictor variable attached to this parameter (in Section 7.6.1 this was "Age") is not an important predictor of the response variable (Price).

The output from `Minitab` for our multiple linear regression, which also uses rainfall and temperature as predictors, is shown at the top of page 182 and can be used in a similar way.

7.6.3 Testing the importance of Age

For example, let us once again consider the importance of **Age** in our model:

$$\begin{aligned} H_0 &: \beta_1 = 0 \qquad \text{versus} \\ H_1 &: \beta_1 \neq 0. \end{aligned}$$

The *p*–value for this is 3%. Thus, *age appears to be important in our model*.

7.6.3 Testing the importance of Rainfall

Rainfall is variable X_2, which has coefficient β_2. Our hypotheses are:

$$\begin{aligned} H_0 &: \beta_2 = 0 \qquad \text{versus} \\ H_1 &: \beta_2 \neq 0. \end{aligned}$$

The p–value for this is 99.6%. Thus, *rainfall is NOT important in our model*.

7.6.3 Testing the importance of Temperature

Temperature is variable X_3, which has coefficient β_3. Our hypotheses are:

$$\begin{aligned} H_0 &: \beta_3 = 0 \qquad \text{versus} \\ H_1 &: \beta_3 \neq 0. \end{aligned}$$

The p–value for this is 0.2%. Thus, *temperature appears to be important in our model*.

7.6.3 Testing the importance of our predictor variables

Since rainfall is not an important linear predictor in our model, we should now remove it and re–fit the model using only age and temperature.

In `Minitab`, we perform the regression again, but this time include only age and temperature as predictor variables.

7.6.3 Testing the importance of our predictor variables

Notice that the regression equation has now changed, and now only includes age and temperature.

We now have:

$$Y = -22.6 + 0.806X_1 + 1.55X_3 + \epsilon,$$

where X_1 represents the age of a bottle of wine and X_3 represents the average temperature during the growing season.

Notice that the p–values for both age and temperature are still less than 0.05, so performing a hypothesis test for both would conclude that both are important in the model.

7.6.3 Testing the importance of our predictor variables

The regression equation above represents our "**final**" model, in that we have excluded all variables that are not important predictors of price, and the model now includes only those predictors that *are* important.

We could now use this model to make predictions.

7.6.3 Testing the importance of our predictor variables

For example, suppose you run a vineyard and have just produced a 7 year–old vintage wine.

During the growing season, the average afternoon temperature was 18.5^{o}C and the total amount of rainfall was 117mm. How much, per bottle, might this wine sell for?

7.6.3 Testing the importance of our predictor variables

We have $X_1 = 7$, $X_2 = 117$ and $X_3 = 18.5$. Thus

$$Y = -22.6 + 0.806 \times 7 + 1.55 \times 18.5 = 11.717,$$

i.e. about £11.72 per bottle.

7.6.4 The R^2 statistic

In the `Minitab` output shown in these lecture notes so far, you may have noticed something called `R-Sq`.

Each time you perform a regression analysis in `Minitab`, the output includes the value of the R^2 statistic, and this is sometimes used as an overall assessment of the quality of our model.

Technically, the R^2 statistic tells us how much of the variation in our Y data is explained by the predictors in the model.