Chapter 7

Techniques of regression

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:

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

These data could have arisen from a random sample of n individuals from a population, on which two measurements/observations (x_i, y_i), $i = 1, \ldots, n$ were made; 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*.

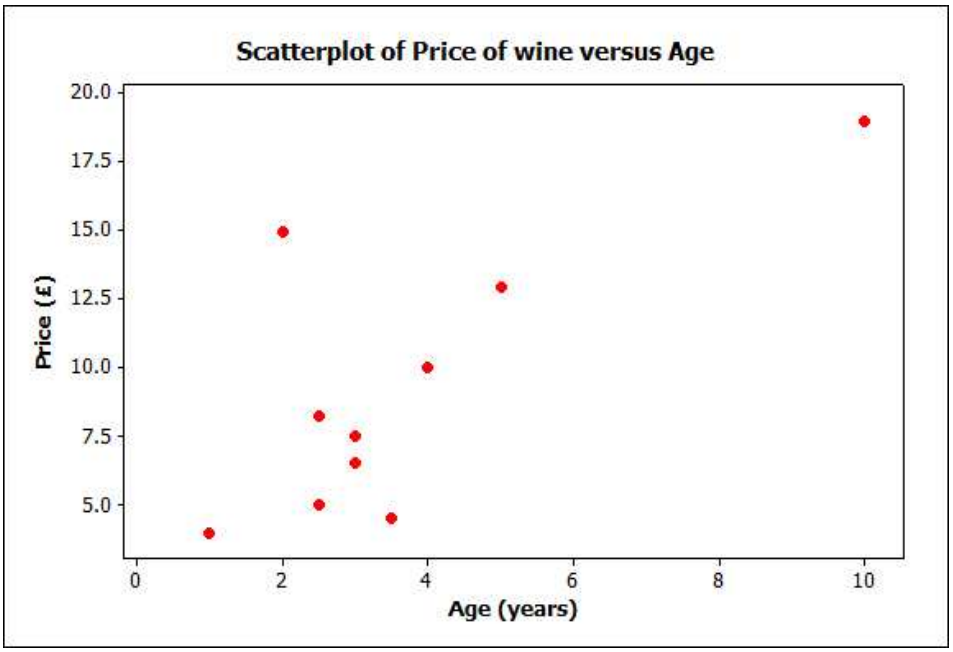
We considered scatter diagrams, often called *scatter plots*, in Chapter 3. From this diagram we can get an initial impression about the relationship between X and Y and form some subjective assessments. The main aim of this chapter is to supplement such descriptive analyses with more formal techniques – both in terms of quantifying the association between X and Y, but also being able to *model* any relationship between X and Y. 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. In New Zealand, wine production is a multi–million dollar industry, and *Saint Clair* is one of the country's leading producers, and exporters, of Sauvignon Blanc wine. *Tanner's Wines*, a fine wine stockist in the U.K., imports wine from all over the world, including 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`. Also shown in this table is the age of each wine selected. The graph below shows a scatter plot of Price (Y) against Age (X); we produce this by plotting points with x and y co–ordinates given by the observed values for X and Y, i.e. $(3.5, 4.50), (5, 12.95), \ldots, (4, 10.00)$. This plot was produced in `Minitab`; you will be reminded of the `Minitab` commands necessary for producing scatter plots in the next practical session.

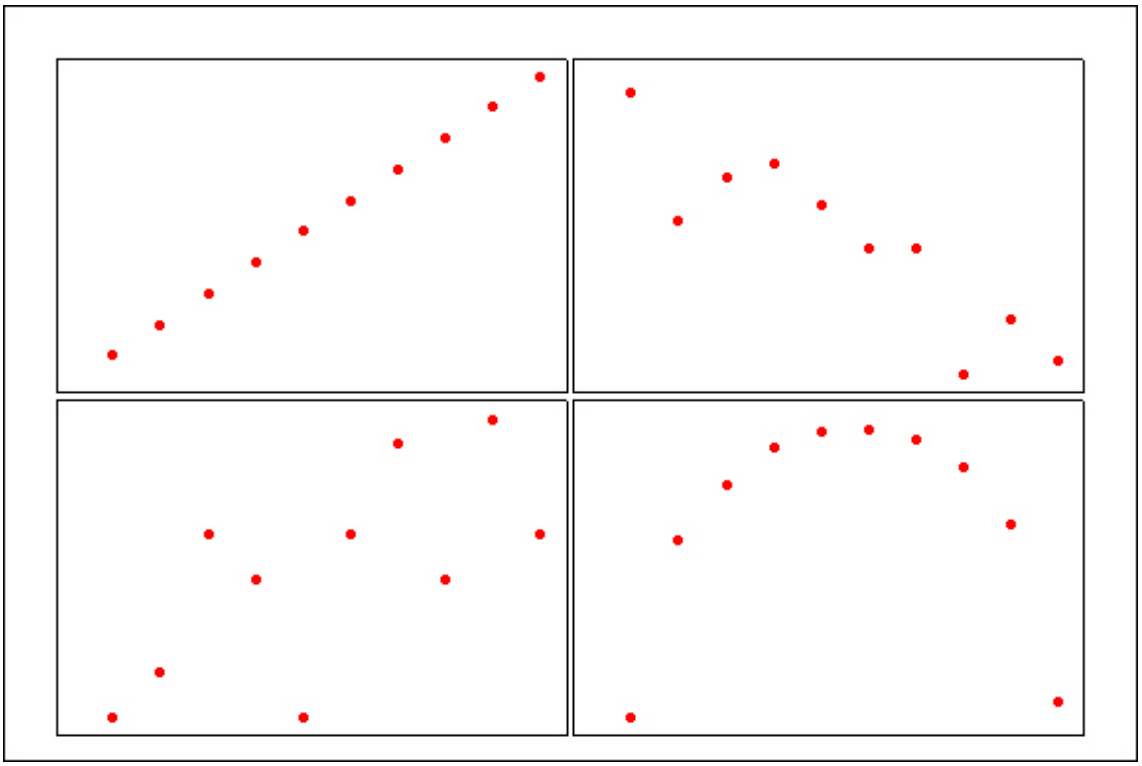
Bottle	1	2	3	4	5	6	7	8	9	10
Age (X years)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Price (£Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00



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?

7.3 Quantifying the relationship: Correlation

There is clearly a relationship between the age and price of wine; the relationship is *positive*, or *direct*, and most people would agree that this positive relationship appears to be *quite strong* and *linear*. How would you describe, in words, the relationship between X and Y in the following scatter plots?



Scatterplots such as the one in the bottom left–hand corner above 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 very subjective; quantifying such relationships numerically can circumvent this problem of subjectivity. 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).

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

Based on this information, can you estimate the value of r for the wine age/price data? And for the four datasets shown in the plots above?

Since we have the data for the wine example, we can calculate the value of r here. The easiest way to calculate the correlation coefficient (other than using a computer/stats mode of a calculator!) 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:

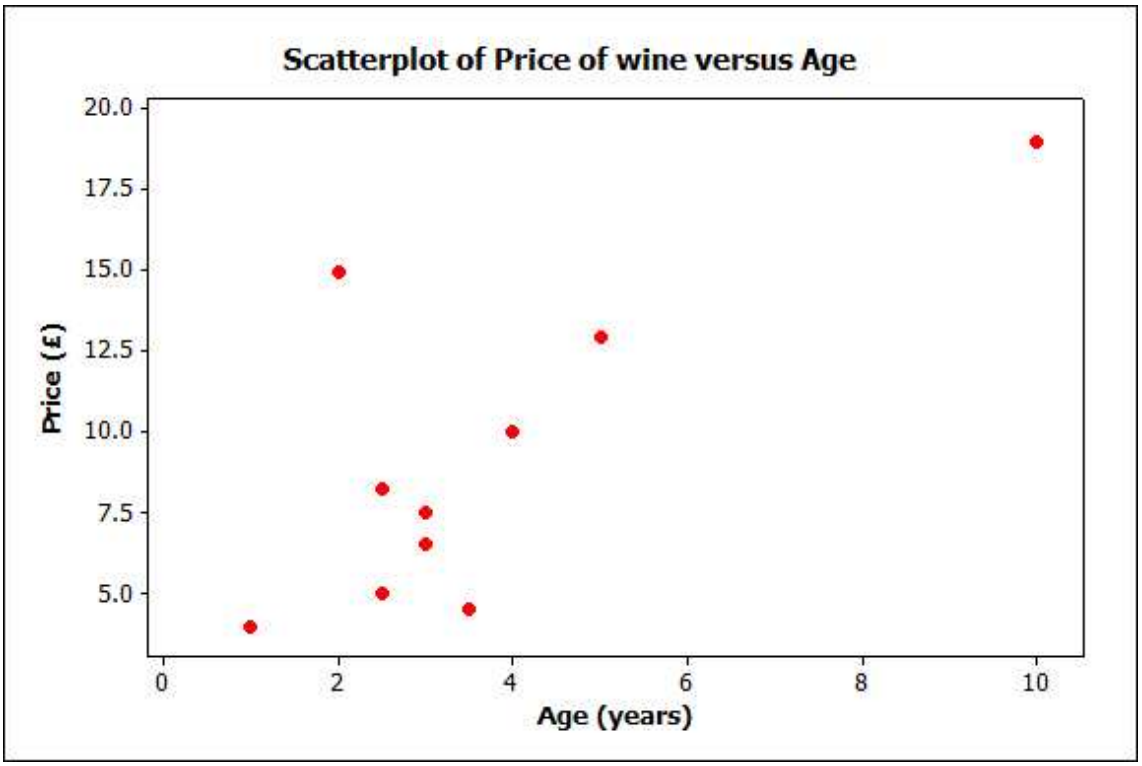
Since this is fairly close to +1, we have a moderate/strong positive linear association between the age and price of wine. Remember that this correlation coefficient can only be used to detect *linear* associations.

For information, the value of r for the plots at the start of this section, from top–left and moving clockwise, is $r = 1, -0.899, 0.699, 0.064$. 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. Recall the scatter plot of the price of wine against the corresponding age of each bottle (shown again below). A "line of best fit" can be drawn through the data, and from this line we can make predictions of price based on age, for ages for which we have no data.

Try this yourself, and use your line of best fit to predict the price of a bottle if wine which is six years old.



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 relatively objective way. We do this through the *regression equation*.

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

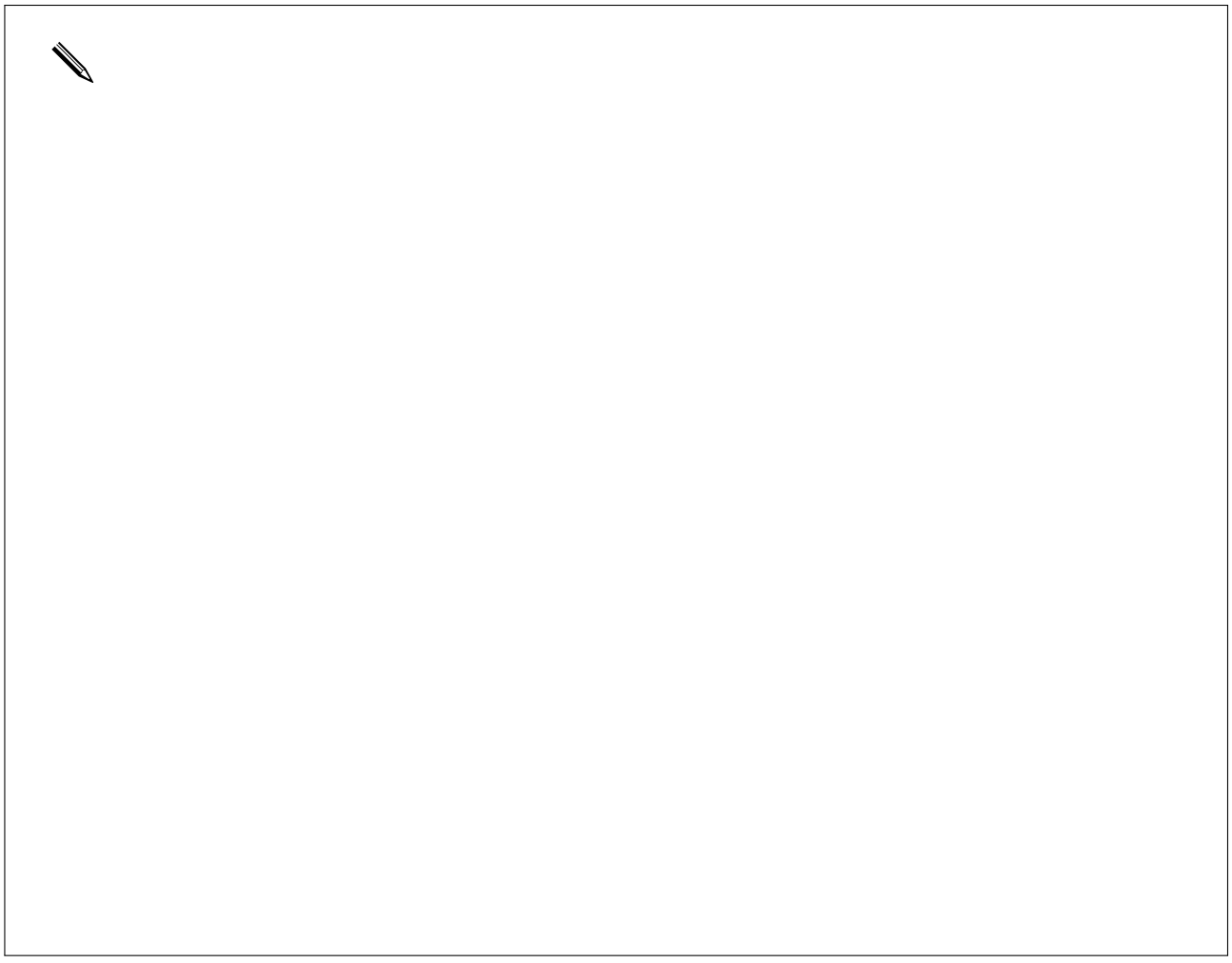
$$y = mx + c,$$

where m and c are the gradient, and intercept, respectively. Statisticians tend to use different notation for their regression equation; the convention is to use

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

where Y is the response variable and X the predictor variable. The unknown parameters β_0 ("beta nought") and β_1 ("beta one") represent the intercept and slope of the population regression line $\beta_0 + \beta_1 X$. Notice we also have an unusual addition to this equation of a straight line: ϵ ("epsilon"). This quantity represent "random error", and is added to the equation to allow for deviations from the straight line (in most real–life applications, we never see a 'perfect' relationship – there is usually some "scatter" about the line!). If we get time, we will think about the role of ϵ in more detail later on.

So we need to find β_0 and β_1; the best values will minimise the vertical distances between the regression line and the data. These distances are known as the *residuals*; this is best seen through a diagram:



Now 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 distances between the points and the line are given by

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

Now some of these distances will be negative, as defined above, as some points will lie below the line; thus, to get rid of any "negative distances", we square all of these quantities:

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

The very best line of best fit – that is, the line which minimises the sum of these "squared distances", 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, without showing the details here (see the last prize question for this!), 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. Doing so gives some very nice formulae:

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \qquad \text{and} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned}$$

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.

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:

$$Y = 3.903 + 1.467 \times 4.5 = 10.505$$

i.e. about £10.50. It is clear from our regression equation that, for every one year increase in age, the selling price of a bottle of wine increases by about £1.47.

A cautionary note

Care should be taken when using the regression equation to make predictions of the response variable. In particular, we should only use our regression equation to make predictions using X–values that lie within the range of the data observed. So, for example, we should not use this regression equation to estimate the selling price of a bottle of wine that has been ageing for 12 years.

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?

The answer is – "we wouldn't"! If our correlation coefficient $r \approx 0$, there is no (linear) relationship between X and Y and so the story ends. However, what if your value for r is about ± 0.5 or ± 0.6? Thus suggests there is *some* linear association, but is this linear association strong enough to warrant further analysis (i.e. regression)?

We can at least attempt to answer this question by performing a hypothesis test for the correlation coefficient.

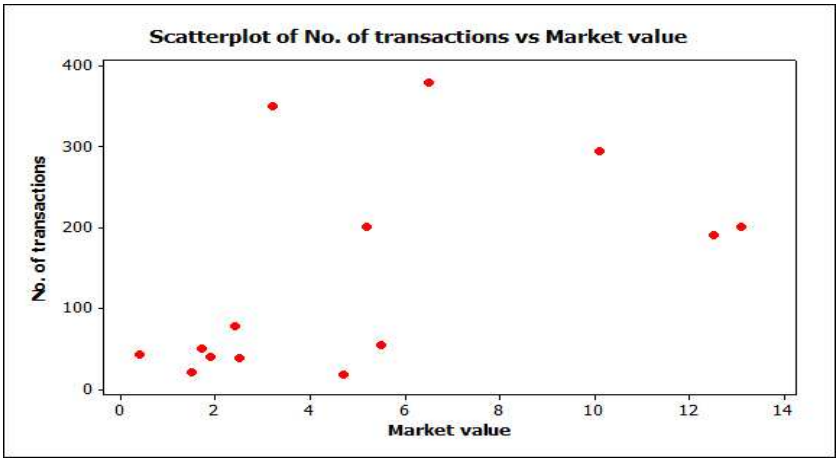
Example 7.1

The following table shows the total market value (X) of 14 companies (in £million) and the number of stock exchange transactions (Y) 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 (X)	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	1.5
Transactions (Y)	380	200	42	50	40	78	350	18	295	190	200	55	38	20

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

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


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

(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 7.2

Test the significance of the correlation coefficient you calculated for the wine age/price data in Section 7.3.

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. For instance, the sales of a product may depend on the product's unit price, as well as the amount of advertising expenditure and the price of a competing product (three predictor variables), or the number of fatal accidents during a time period may depend on the number of registered vehicles on the road and the price of petrol (two predictor variables). 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*.

Most of the work covered in this section will be demonstrated via `Minitab`. However, before we start, we shall think about how to determine whether a predictor variable is an *important* predictor variable.

7.6.1 Back to the simple linear regression model

The regression output given by `Minitab` also allows us to check the significance of the slope in our regression equation. Recall that the simple linear regression model is given by

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

where β_0 represents the y–intercept of our regression line (the point on the y–axis at which the regression line intersects) and β_1 represents the slope of the regression line. If there is little or no (linear) relationship between X and Y, then not only will the correlation coefficient be close to zero, but so too will the slope term β_1. If the slope term is zero, then X drops out of the above linear regression model (since $\beta_1 X = 0 \times X = 0$, and so we are left with $Y = \beta_0 + \epsilon$) and we can conclude that the value of X does not influence the value of Y. In reality, we do not know the true value of β_1; from our data, we have the *estimated* value $\hat{\beta}_1$, and so we proceed with a hypothesis test for the population slope β_1 in the same way we did for the population correlation coefficient ρ. The null and alternative hypotheses are now:

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

If, based on our data and the test statistic, we retain H_0, then we would conclude that the slope term β_1 is not significantly different from zero and thus X is not an important predictor of Y. If we reject H_0 and thus go with the alternative hypothesis

H_1 then we would conclude that the slope term *is* important in our model, and so X *is* a significant predictor of Y. Again, `Minitab` can be used to this end.

Recall that for our wine data, the estimated linear regression equation is

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

This regression analysis can be performed in `Minitab`, giving the following output:

Regression Analysis: Price versus Age

```
The regression equation is
Price = 3.90 + 1.47 Age


Predictor    Coef  SE Coef     T      P
Constant    3.905    2.088  1.87  0.098
Age        1.4666   0.4806  3.05  0.016


S = 3.58129   R-Sq = 53.8%   R-Sq(adj) = 48.0%
```

`Minitab` tells us that the estimated slope term using the data in our sample is $\hat{\beta}_1 = 1.4666$ (to 4 d.p.); obviously, this is specific to our dataset and will vary from sample to sample, but the theory suggests that this will vary with standard deviation 0.4806 (the "standard error" of our estimator); the test statistic is just the estimated coefficient divided by its standard error (1.4666/0.4805) which gives $t = 3.05$, and this gives the answer to step 3 of our hypothesis test. `Minitab` then uses this test statistic to obtain a p–value for this test, which is 0.016, or 1.6%. We can now interpret the p–value in the usual way:

- We have *moderate* evidence against H_0 (since p lies between 1% and 5%)
- Reject it and go with the alternative H_1
- The alternative suggests that $\beta_1 \neq 0$, and so there *is* a significant slope term in our model

So, not only is there a significant correlation between age and wine, but age is an important *predictor* of the price of a bottle of wine.

7.6.2 Extending the simple linear regression model

In Section 7.2 we discussed that the price of a bottle of wine might not only depend on the age of the wine; other factors may have a part to play – for example, the amount of rainfall during the grape–growing season. In fact, the country and region of origin of the ten bottles of wine that were randomly selected were sourced, and data were then also collected relating to the total observed rainfall, and average daily afternoon temperature, during the grape growing season in the year the wine was produced. The full dataset is shown in the table below:

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^oC)	16	20	17	18	18	22	19	15	21	20

Notice that we've labelled the predictor variables X_1, X_2 and X_3; the main response variable – the price of a bottle of wine – is still Y. 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. In fact, β_1 is the parameter that multiplies the age of the bottle of wine, β_2 is the parameter that multiplies the amount of rainfall observed, and β_3 is the parameter that multiplies the average temperature. Notice that this is a natural extension of the simple linear regression model using age only as presented in Section 7.4.

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. We can enter the data from the table above in the first four columns of a `Minitab` worksheet. Then, using `Stat`–`Regression`–`Regression`, and entering the column containing the prices as the `response` and the other columns containing Age, rainfall and temperatures as the `Predictors`, `Minitab` will perform the multiple linear regression analysis estimating all the parameters in our model. Doing this gives the (edited) output shown overleaf.

Regression Analysis: Price versus Age, Rainfall, Temp.

```
The regression equation is
Price = - 22.5 + 0.807 Age - 0.0004 Rainfall + 1.55 Temp.

Predictor      Coef  SE Coef      T      P
Constant     -22.54    10.54  -2.14  0.076
Age          0.8065   0.2853   2.83  0.030
Rainfall   -0.00042  0.07383  -0.01  0.996
Temp.        1.5538   0.3117   4.99  0.002

S = 1.80760   R-Sq = 91.2%   R-Sq(adj) = 86.8%
```

Thus, the full (multiple) regression model is:

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

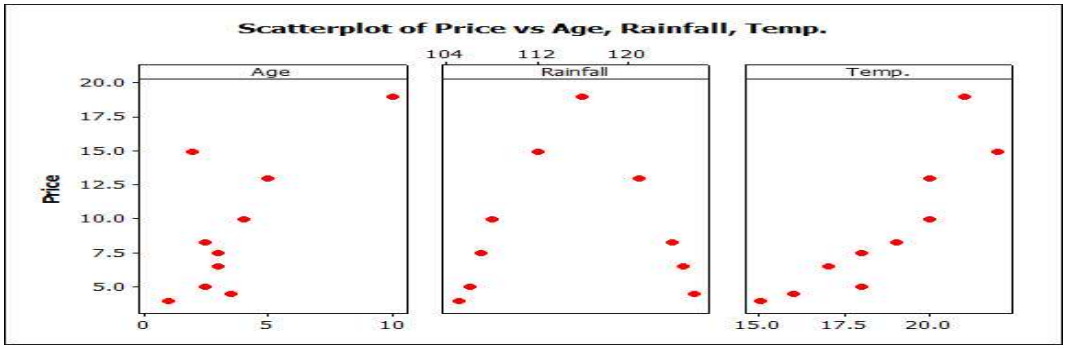
where X_1, X_2 and X_3 represent the age of a bottle of wine, the total rainfall during the growing season and the corresponding average afternoon temperature (respectively). The estimated coefficients of the model indicate the direction of the relationship between the price of a bottle of wine and each of the corresponding predictors. For example:

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

We need to be careful when doing this, however. If there are strong correlations between the predictor variables this could lead us to making incorrect conclusions about these relationships. This is known as *multicolinearity*. In this case, however, the observations made above shouldn't be too surprising:

- You might expect the price of a bottle of wine to increase as its age increases: vintage wines, for example, are usually quite expensive!
- Our model suggests that as rainfall during the growing season increases, the value of a bottle of wine from that region decreases: the more rain, the less sun, and so the lower the quality of the grapes!
- Our model also suggests that the higher the average temperature during the growing season, the higher the price of a bottle of wine from that region: again, you might expect this, as the higher the temperature, the more sunshine we have!

However, producing simple scatter plots of each predictor variable (age, rainfall and temperature) against the response variable (price) can help to inform our model. Such scatter plots have been produced in `Minitab`, and are shown below.



Notice that, in agreement with our model, there are positive linear relationships between age/price and temperature/price. However, our model suggests a negative linear relationship between rainfall/price, and the the left–hand side of the scatter plot for rainfall and price doesn't seem to match up with this. In fact, what we see is a *non–monotone* relationship, and possibly a *non–linear* relationship, which both increases with rainfall *and* decreases. This sort of relationship could actually be sensible: grapes need a certain amount of rainfall during the growing season, but too much can be detrimental to the quality of the grapes produced; thus, there might be an "optimal" amount of rainfall necessary for producing high–quality grapes for wine – too little and the grapes will be lower in quality, and too much could also produce lower quality grapes. The lower the quality, the cheaper the bottle of wine!

Since there is a non–standard relationship between rainfall and price, we might question using rainfall in our model, or perhaps think of more complex models which would be more appropriate for such a relationship. *This highlights the importance of the humble scatter plot!*

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 187 and can be used in a similar way.

Testing the importance of Age as a predictor

For example, let us once again consider the importance of **Age** in our model. Age is variable X_1, which has coefficient β_1. Our hypotheses are:

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

The p–value for this, as given in the `Minitab` output, is 0.030 (or 3%). This lies between 0.01 and 0.05 (1% and 5%), and so we have *moderate evidence against* H_0. Thus we *reject* H_0 and go with the alternative H_1; β_1 is significantly different from zero, and so *age appears to be important in our model*.

Notice that the p–value for Age in the multiple linear regression (0.030) is different to that obtained in the simple linear regression (0.016, page 180). This is because in a multiple linear regression, each variable is tested *in the presence* of the other variables.

Testing the importance of Rainfall as a predictor

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 rainfall coefficient β_2 has a p–value of 0.996 (or 99.6%). Since this is very high, and certainly above 10%, we have *no evidence against* H_0. Thus we *retain* H_0: $\beta_2 = 0$ and so *rainfall is NOT important in our model*.

Testing the importance of Temperature as a predictor

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 temperature coefficient β_3 has a p–value of 0.002 (or 0.2%). Since this is very small, and certainly less than 1%, we have *strong evidence against* H_0. Thus we *reject* H_0 and go with the alternative H_1; β_3 is significantly different from zero, and so *temperature appears to be important in our model*.

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. Doing so gives the (edited) output shown overleaf.

Regression Analysis: Price versus Age, Temp.

```
The regression equation is
Price = - 22.6 + 0.806 Age + 1.55 Temp

Predictor      Coef  SE Coef      T      P
Constant   -22.589    4.964  -4.55  0.003
Age         0.8061   0.2553   3.16  0.016
Temp.       1.5540   0.2855   5.44  0.001

S = 1.67352   R-Sq = 91.2%   R-Sq(adj) = 88.6%
```

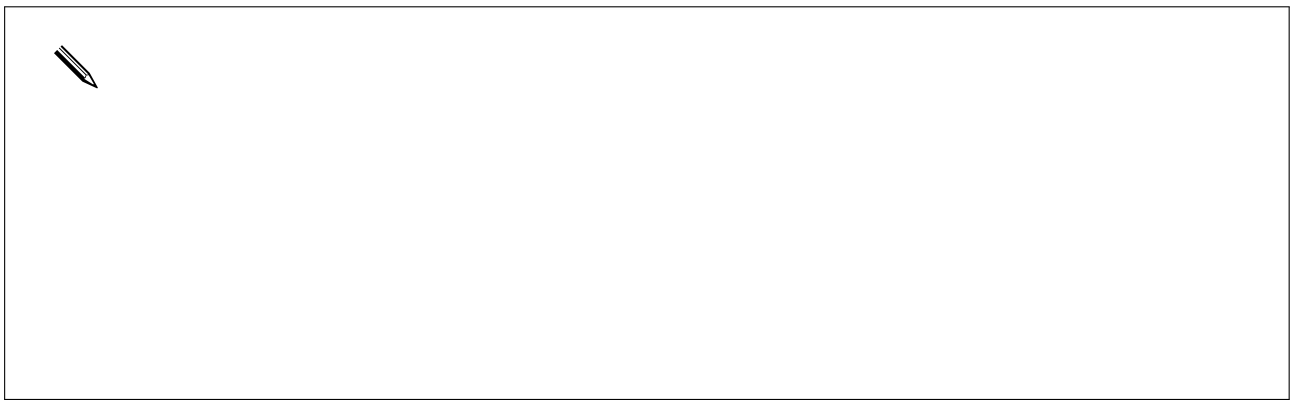
Notice that the regression equation has now changed, and only includes age and temperature. We now have:

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

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 (there is strong evidence to include temperature and moderate evidence for age).

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.

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°C and the total amount of rainfall was 117mm. How much, per bottle, might this wine sell for?



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. For simplicity, suppose you have the simple linear regression model

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

If we observe a perfect relationship between X and Y, that is, all of our points lie perfectly on a straight line, then if we know X, we know Y, as the Y value could just be "read off" from the regression line. In this case, we would say that X explains 100% of the variation in Y, and the corresponding value of the R^2 statistic would be 100%.

Thus, the closer the value of the R^2 statistic to 100%, the better the model. We can compare the R^2 statistic for the various fits we have tried out for the wine data:

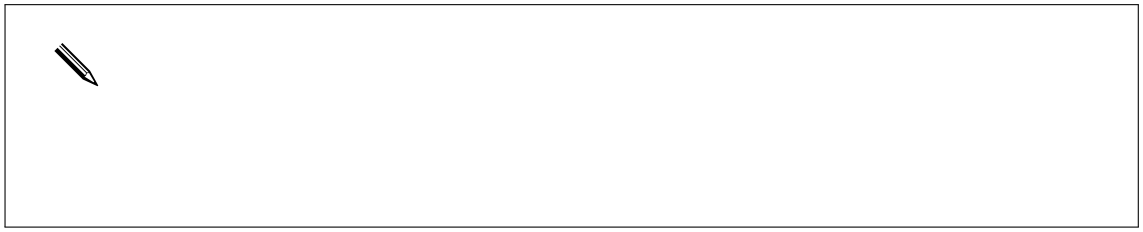
Model	R^2
Simple – including just "Age" (page 185)	53.8%
Multiple – including "Age", "Rain" and "Temp" (page 187)	91.2%
Multiple – including "Age" and "Temp" (page 190)	91.2%

From this, we can clearly see the effect of including more than just "Age" in the model as a predictor of price: the R^2 value has increased sharply from 53.8% to 91.2%. Also notice the effect of removing "Rainfall" from the model; that is, this has had no effect at all, at least on the R^2 statistic (to 1 d.p., anyway), further confirming that the rainfall variable is not an important linear predictor in the model. Generally, the higher the R^2 statistic the better; however, we are prepared to allow small reductions in R^2 if it means we can remove a non–significant predictor variable.

Example 7.3

On a small island the government would like to predict the number of mortgage loans issued by the state mortgage company (Y) from: the amount of personal income in millions of local currency (X_1), the interest rate (X_2) and the year (X_3).

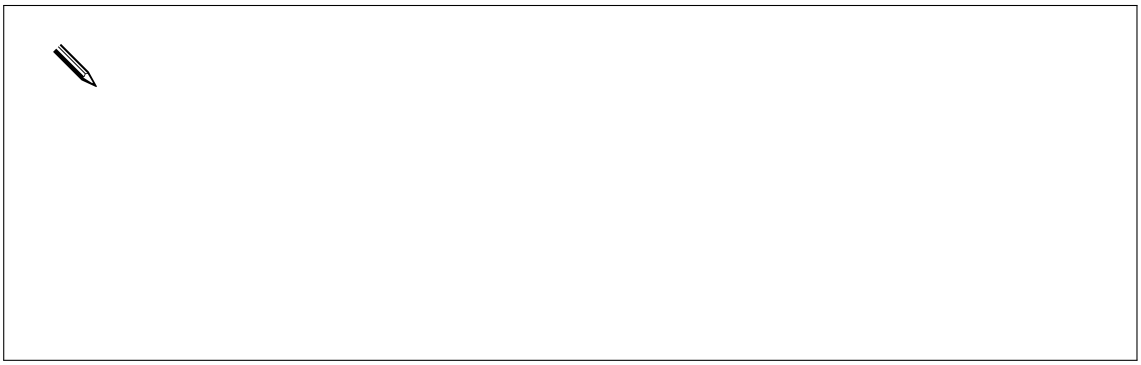
1. Write down the "full" multiple linear regression model that might be suitable for this scenario.



To estimate the parameters of the model the government collects the following data over a 10–year period:

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Income	3.2	3.3	3.4	3.5	3.7	3.8	3.9	4.1	4.3	4.6
Interest	7.0	7.5	7.5	8.0	7.0	7.0	6.0	5.5	5.0	4.5
Mortgages	6,253	6,516	4,678	6,743	8,586	7,087	10,386	13,591	13,649	16,717

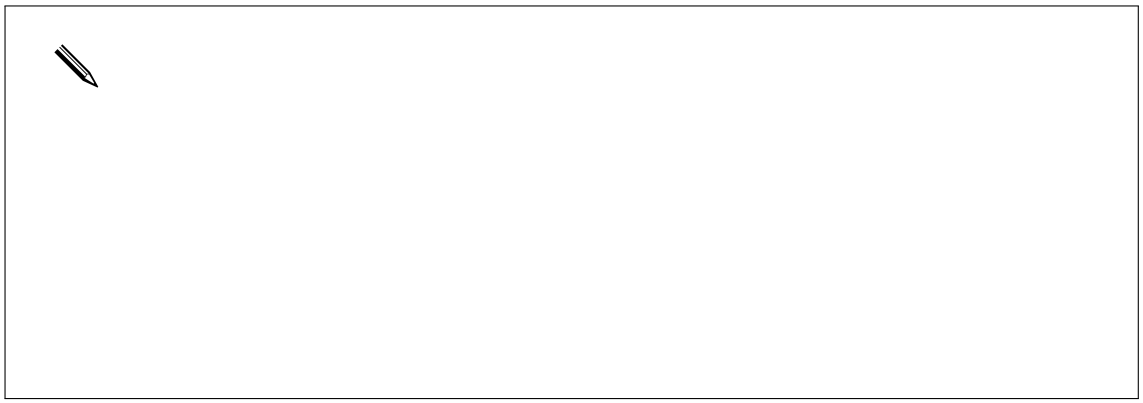
2. Use `Minitab` to fit the model in (**1**). Summarise the regression output in the space below, making sure you write down the fitted model.



3. Which, if any, of the variables *do not appear to be important predictors* in your model?

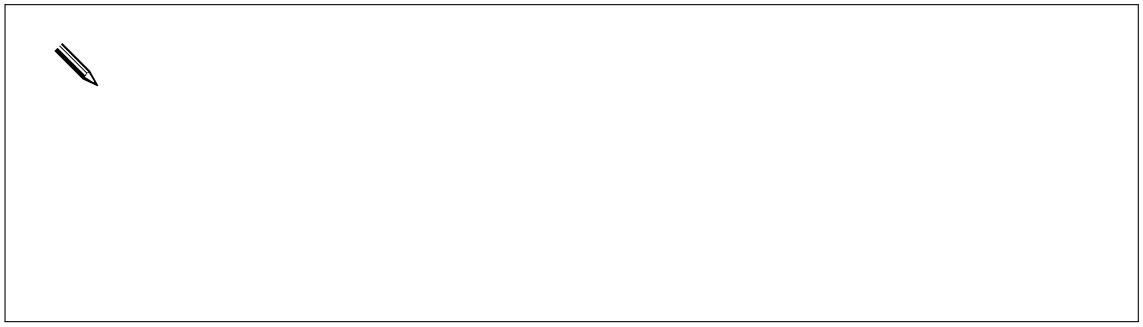
4. Remove the least important variable from your model, and re–fit using `Minitab`. Again, summarise the regression output in the space below, making sure you write down the fitted model. *[Note: Never remove more than one variable at a time! Always remove only the variable with the highest p–value (provided this is greater than 0.05) and then re–fit!]*

5. Repeat step 4 until you have removed all variables that do not appear to be important predictors in your model. Write down your "final" regression equation in the space below:



6. Comment on the R^2 statistic in your regression analyses.

7. This year (2016), the average income is forecast to rise to 4.9 million units of the local currency and the interest rate is set to rise to 5%. How many mortgages can the state mortgage company expect to issue this year?



7.7 Chapter 7 practice questions

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

Hint: You may use the following summaries:

$$\sum_{i=1}^{12} x_i = 1326 \qquad \sum_{i=1}^{12} y_i = 157.8$$

$$\sum_{i=1}^{12} x_i^2 = 148308 \qquad \sum_{i=1}^{12} y_i^2 = 2125.52 \qquad \sum_{i=1}^{12} x_i y_i = 17695.1$$

2. "Northern Lights", the main electrical supplier to homes in Northern Sweden, believes that the time a customer takes to pay their electricity bill is related to the size of their bill. To investigate, their research team randomly selected 10 customers and recorded the size of their bill (x, in pounds) and the time it took to pay this bill (y, in days). The results are show below.

x	400	105	205	150	460	250	315	420	100	300
y	35	15	18	20	30	22	25	34	10	20

(a) Produce a scatterplot for these data, and comment on the relationship between the two variables. Don't forget to label your plot.

(b) The following summaries have been obtained for the above data:

$$\sum x = 2705 \qquad \sum y = 229$$

$$\sum x^2 = 885275 \qquad \sum y^2 = 5839 \qquad \sum xy = 70720$$

Using these summaries,

(i) calculate the sample correlation coefficient, and comment;

(ii) perform a linear regression analysis, and obtain the linear regression equation. Plot this regression line on your scatter diagram in part (a).

(c) Northern Lights levy a "late–payment charge" if a customer takes longer than 30 days to pay their bill. Mr. Adams' bill for 2012 is £375. Will he incur this late payment charge?

(d) Another customer, Miss Bloggs, has a large electricity bill of £520. Why should we be cautious about using the regression equation obtained in part (b) (ii) to predict how long she will take to pay her bill?

3. The marketing team at *Marks & Spencer* are investigating the effectiveness of three types of advertising currently used: localised direct mailing (e.g. flyers posted through letterboxes), newspaper advertising and TV advertising. The team collect one week's data from 25 randomly selected stores, recording:

$$
\begin{aligned}
y &= \text{Weekly gross sales (£ thousand)} \\
x_1 &= \text{Weekly local expenditure on direct mailing (£ thousand)} \\
x_2 &= \text{Weekly local expenditure on newspaper advertising (£ thousand)} \\
x_3 &= \text{Weekly local expenditure on TV commercials (£ thousand)} \\
x_4 &= \begin{cases} 1 & \text{if the store has a foodhall} \\ 0 & \text{if the store does } \textit{not} \text{ have a foodhall} \end{cases}
\end{aligned}
$$

(a) Which variable(s) do you think are *indicator variables*?

(b) `Minitab` was used to fit a multiple linear regression model to these data; the resulting (edited) output for the "full" model is shown below. Look at it, and then answer the following questions.

Regression Analysis: y versus x1, x2, x3, x4

```
The regression equation is

y = ?_____________________________________


Predictor     Coef  SE Coef      T         P

Constant     82.93    10.93   7.59     0.000
x1           2.894    3.837   0.75     0.459
x2          -2.232    4.355  -0.51     0.614
x3         13.1891   0.9647  13.67     0.000
x4           9.182    3.367   2.73     0.013


S = 5.72401   R-Sq = 96.2%   R-Sq(adj) = 95.5%
```

(i) Complete the regression output above by filling in the blank indicated by the question mark (`y = ?`).

(ii) Which variable should be removed before a reduced model is fitted? Explain your answer, with reference to the null hypotheses:

$$H_0 \;:\; \beta_j = 0 \qquad j = 1, \ldots, 4.$$

(c) The (edited) output from `Minitab` shown below represent the "final" model. Look at it, and then answer the following questions overleaf.

Regression Analysis: y versus x3, x4

```
The regression equation is

y = ?_______________________________________


Predictor      Coef    SE Coef        T          P

Constant     81.796      7.072    11.57      0.000

x3          13.2602     0.9329    14.21      0.000

x4         ?________     3.218     2.66      0.014


S = 5.58254   R-Sq = 96.1%   R-Sq(adj) = 95.7%
```

(i) Complete the regression output by filling in the blanks indicated by the question marks (`?`).

(ii) Briefly explain what has happened between the fit of the "full" and "final" models.

(iii) Give a practical interpretation of including x_4 in the model.

(iv) Comment on the change in the R^2 statistic from the "full" to the "final" model.

(v) Next week, *Marks & Spencer* will spend £5000 on TV advertising in the Northeast of England. The Newcastle branch of *Marks and Spencer* has a foodhall. What can we expect the weekly gross sales to be for this branch of the store?