



ACE2013

**Statistics for
Marketing and Management**

Part 2: Statistical Business Modelling

Semester 2, 2013—2014

Lecturer: Dr. Lee Fawcett

School of Mathematics & Statistics

Introduction to ACE2013 Part 2: Statistical Business Modelling

Welcome to Part 2 of ACE2013: Statistics for Marketing and Management. I've decided to call this part of the course "Statistical Business Modelling", as this title sets the scene for the material we will cover:

1. Correlation and regression

- Review of correlation and regression from MAS1403
- Multiple linear regression
- Regression diagnostics
- Non-linear regression
- Analysing a binary response: logistic regression

2. Time series and forecasting

- Review of time series and forecasting from MAS1403
- Moving Average (*MA*) processes and Auto-Regressive (*AR*) processes
- Box-Jenkins approach to model-building

3. Business modelling – one topic from:

- Linear programming
- Planning projects
- Quality control

Each topic will start with a quick review of what we've already covered in MAS1403; I think this is important to refresh your memories and provide background to the new material we will cover. However (unlike Part 1 of this course) we will concentrate on the new material, introducing more realistic statistical techniques and (hopefully) reinforcing the idea that Statistics has a place in the world of Business and Marketing!

Lecturer contact info

Name: Dr. Lee Fawcett
 Office: Room 2.07, 2nd floor of the Herschel Building
 Phone: (0191) 222 7228
 Email: lee.fawcett@ncl.ac.uk
 Web: www.mas.ncl.ac.uk/~nlf8
 Access: Open-door policy – just knock

Differences between this course and MAS1403

- *We will cover fewer topics in more detail*
Notice there's only three main topics. In MAS1403 we covered a new topic every week, in one lecture and one tutorial! In ACE2013, each topic will take three or four weeks to complete, and there's two lectures per week.
- *We will use the computer extensively*
Modelling real-life more realistically usually means the maths gets more difficult. But don't worry, we will use the **Minitab** package extensively, especially in the first two topics. I will use **Minitab** in *every* lecture, and there will be fortnightly computer practicals. You will have to interpret computer output in the exam.
- *You will have more to do in lectures*
I will expect you to take more notes in lectures, and there will be less "spoon-feeding" and "hand-holding"! Make sure you make supplementary notes and always work hard to try to follow what I'm doing. This is more important than in first year, since the concepts are more difficult!
- *You will have more to do outside of lectures*
You should read your notes before coming to lectures.

Timetable of events

- *Lectures*
One *two hour* lecture every week – unfortunately, the time/place seems to change, so keep an eye on the schedule overleaf.
- *Computer practicals*
One per fortnight – in *odd teaching weeks*, starting in teaching week 3: again, keep an eye on the schedule overleaf.

Assessment

- *Formal University examination: 60%*
Two hour exam in May/June covering the entire course. We will have time to run through last year's exam, as well as a specimen paper.
- *Computer Based Assessments: 20%*
There are four of these; you've already done two of them. There will be two more in Semester 2; see below for deadlines. In total, they will count for 20% of the course.
- *Practical work: 20%*
There will be some assessed work in each of the fortnightly practical sessions; you will hand in the solutions to these at the end of term (see below for deadlines).

Full week-by-week schedule

Teaching week	Event	Date	Time	Place
1	Lecture	Tues 28th Jan	12–2	Bedson 2.76
2	Lecture	Tues 4th Feb	12–2	Bedson 2.76
3	Practical	Thur 13th Feb	5–6	Stephenson Tree
	Lecture	Fri 14th Feb	3–5	KGVI LT4
4	Lecture	Tues 18th Feb	4–6	Stephenson F13
5	Lecture	Tues 25th Feb	12–2	Bedson 1.75
	Practical	Tues 25th Feb	5–6	Herschel PC
6	Lecture	Tues 4th Mar	12–2	Bedson 1.75
7	Lecture	Tues 11th Mar	12–2	Bedson 1.75
	Practical	Tues 11th Mar	5–6	Herschel PC
8	Lecture	Tues 18th Mar	12–2	Bedson 1.75
9	Lecture	Tues 25th Mar	12–2	Bedson 1.75
	Practical	Thur 27th Mar	1–2	Herschel PC
<i>Easter holidays</i>				
10	Lecture	Tues 29th Apr	12–2	Bedson 2.76
11	Lecture	Tues 6th May	12–2	Bedson 2.76
	Practical	Tues 6th May	5–6	Herschel PC
12	Revision	Tues 13th May	12–2	Bedson 2.76

Deadlines!

- CBA3: Friday 14th March 2014 (23:59). This CBA will open in both practice and assessed modes on Monday 24th February 2014.
- CBA4: Friday 9th May 2014 (23:59). This CBA will open in both practice and assessed modes on Monday 24th March 2014.
- Practical work: Monday 13th May 2014 (4pm). Solutions to *all* assessed practical completed during the term should be handed in together, word-processed, to the Maths & Statistics General Office on the 3rd floor of the Herschel Building. *Work submitted elsewhere will not be marked!*

Enough of that, let's start some work...

Chapter 1

Correlation and regression

1.1 Introduction

This Chapter reviews the techniques we used last year in MAS1403 for assessing the relationship (if any) between two variables. We will re-visit correlation and simple linear regression, where a *regression equation* is formulated to enable us to make predictions of one variable (the *response* or *dependent* variable) given the level of another variable (the *predictor* variable). We will then extend these ideas to consider *multiple linear regression*, where we allow for more than one predictor variable in the regression equation, as well as *non-linear regression* for curved relationships and *logistic regression* where the response variable is binary (can only take one of two values). We will use **Minitab** to formulate our models and consider the process of *model-building*; we will also use **Minitab** to check the assumptions underpinning many of the models considered. The work in Chapter 2 will rely heavily on an understanding of the material in this Chapter.

1.2 Review of correlation and regression: MAS1403

In MAS1403 (Chapter 6, Semester 2) we thought about how we might analyse bivariate data using correlation and simple linear regression techniques. In fact, we first encountered bivariate data right at the start of MAS1403 (Chapter 2, Semester 1) when we looked at *scatter diagrams* or *scatter plots*. Suppose our data consist of n pairs of observations on two variables X and Y , i.e. we have data of the form:

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

Our variables X and Y might be height and weight, or market value and number of transactions, or maybe temperature and sales of ice cream, respectively. 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 in analysing such bivariate data is *always* to plot the data on a scatter diagram.

Motivating example

Product placement, sometimes known as *embedded marketing*, is a form of advertising where branded goods or services are placed in a context usually devoid of advertisements – such as movies, music videos and TV shows.



Advertisers might be interested in the relationship between the chance of a viewer being able to recall the brand and, for example: (i) the number of times the product appeared; (ii) how many minutes into the show the product appeared; (iii) the type of show in which the product appeared. For example, we might expect the chance of a viewer being able to recall the brand to increase as the number of times the brand appears increases (*positive* relationship); similarly, we might expect a product placement near the end of a film to be more easily remembered than one right at the start. The type of show might also have an effect: for example, studies have revealed that product placement in violent or sexually explicit films is less effective than in comedies or “chick flicks”.

The advertiser might be interested in other relationships: for example, younger people might be more likely to recall a brand than older people watching the same film (i.e. a *negative* association with age); females might be more likely to recall a brand than males watching the same film; and the product’s prior exposure might also influence the viewer’s ability to recall the brand.

Being able to quantify, and model, such relationships is crucial for companies interested in using this form of marketing to advertise their product; they want to make sure their product has the best chance of being remembered, and so need to place it in the best film/TV show that will maximise this chance. We return to this example in Section 1.5.

Example: Price of wine

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. Table 1.1 below shows the price of 10 randomly selected bottles of wine from www.tanners-wines.co.uk, an online wine merchant. Also shown in Table 1.1 is the age of each wine selected. Figure 1.1 shows a scatter plot of Price against Age; we produce this by plotting points with x and y co-ordinates given by the observed values for X and Y , i.e. $(3.5, 4.50)$, $(5, 12.95)$, $\dots$, $(4, 10.00)$. Doing so gives the plot shown in Figure 1.1. 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

Table 1.1: Age of wine (in years) and corresponding price (in £)

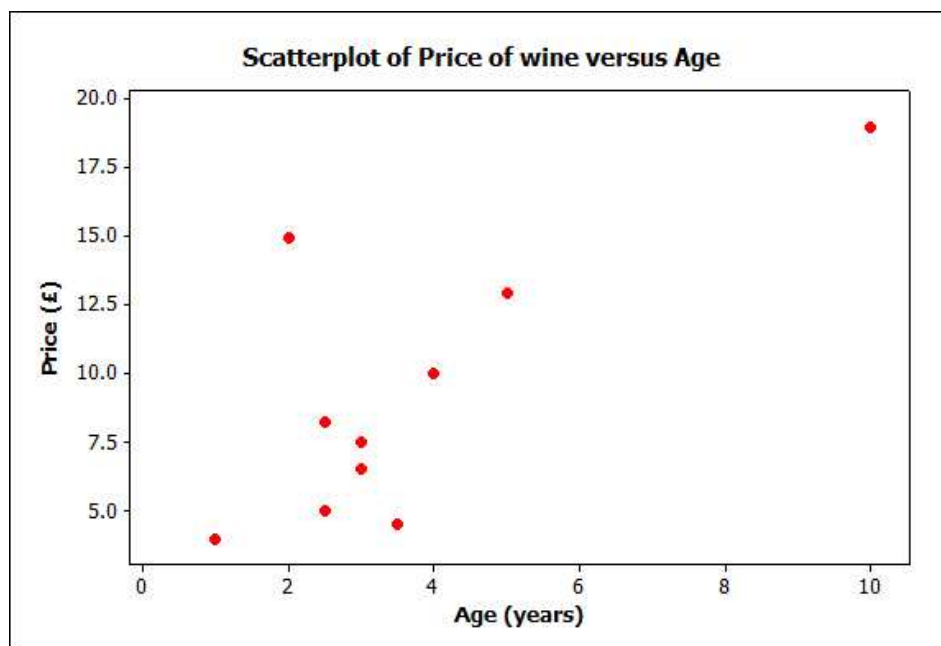


Figure 1.1: Scatter plot of price of wine (£) against age (years)

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?



1.2.1 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?

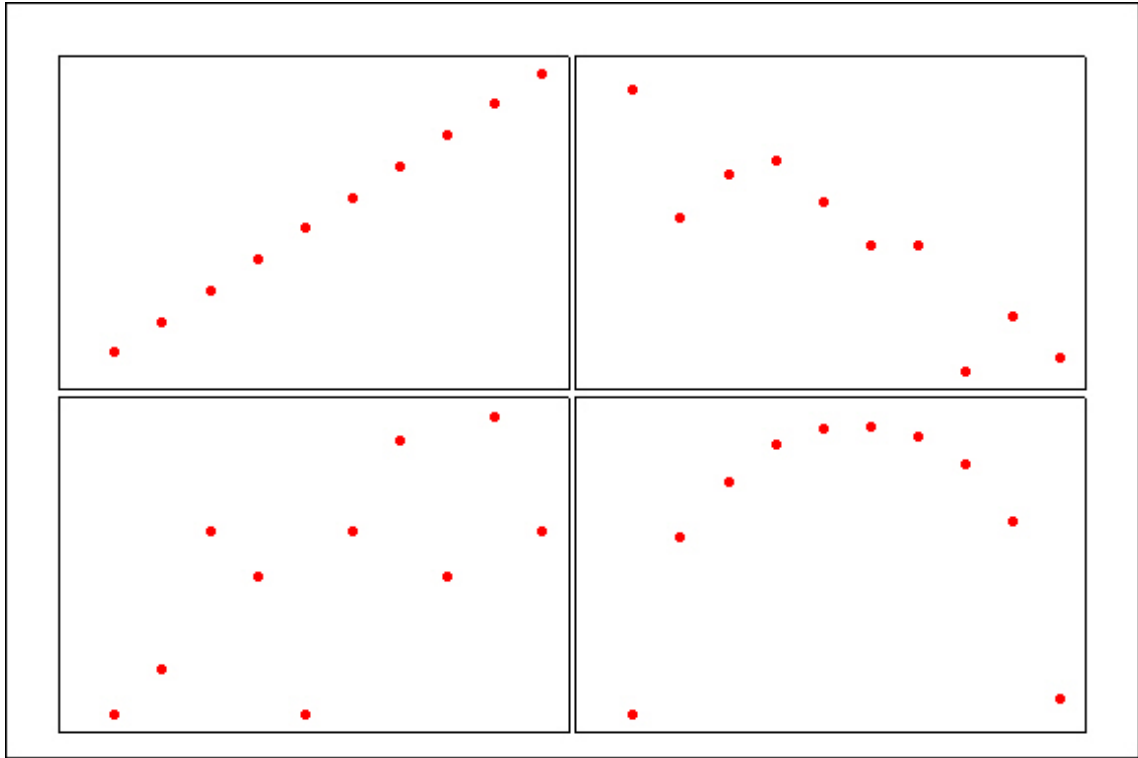


Figure 1.2: Scatter plots showing different relationships

Scatterplots such as the one in the bottom left-hand corner of Figure 1.2 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

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

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 $+1$. 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. Based on this, can you estimate the value of r for the wine age/price data? And for the four datasets shown in Figure 1.2?

Since we have the data for the plot shown in Figure 1.1, we can calculate the value of r here. Thinking back to MAS1403, the easiest way to calculate the correlation coefficient (other than using a computer!) 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:





...continued...

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 in Figure 1.2, 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.

1.2.2 Modelling the relationship: simple 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 (Figure 1.1). 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 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 completely objective way. We do this through the *regression equation*. Recall from MAS1403 that the simple linear regression equation takes the form

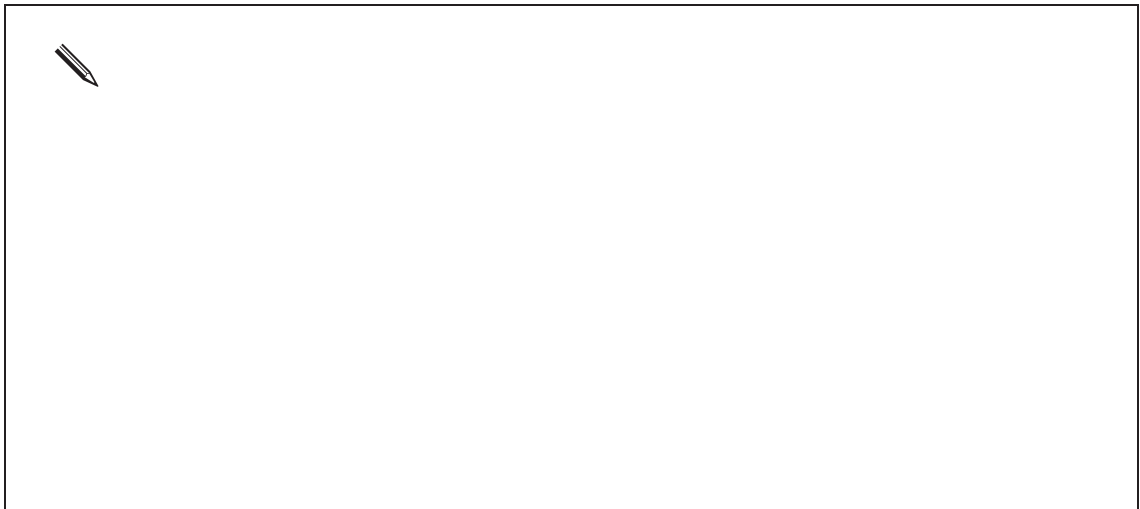
$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where Y is the response variable and X the predictor variable, and ϵ (“epsilon”) is a “random error” with zero mean and constant variance. 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*. The values of β_0 and β_1 which give rise to the ‘best’ regression line, i.e. the line which minimises the residuals, are

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \quad \text{and} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x},\end{aligned}$$

where S_{XY} and S_{XX} are as before. The ‘hats’ on β_0 and β_1 are there to remind ourselves that we have *estimated* β_0 and β_1 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. Note that in MAS1403, we used α and β where we now have β_0 and β_1 (respectively). It will become apparent later in this Chapter why we have changed the symbols used.



The plot in Figure 1.3 shows the scatter diagram for the wine data again, but now with the regression line superimposed. Again, you will be reminded in the next computer practical class how to produce such a plot in **Minitab**. 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 the data given in Table 1.1, 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. Recall from MAS1403 that 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. We can also interpret the regression equation in the following way: for every one year increase in age, the selling price of a bottle of wine increases by about £1.47.

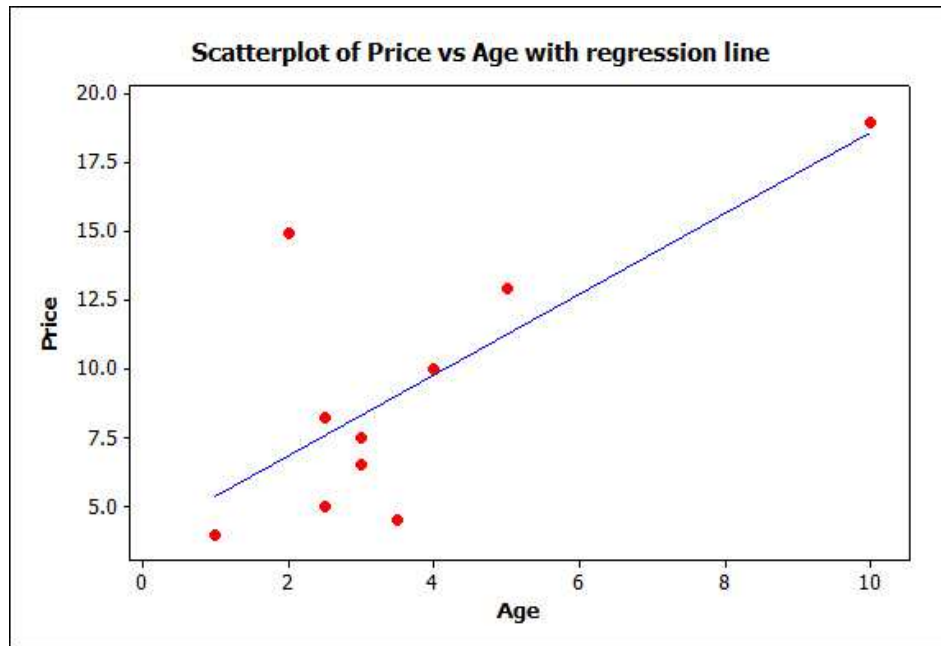


Figure 1.3: Scatter plot of price of wine (£) against age (years), with regression line superimposed

1.2.3 Using Minitab

If we enter the wine data into two columns of a Minitab worksheet, then click on **Stat**–**Regression**–**Regression**, we can enter Price as the **Response** variable and Age as the **Predictor** variable (see Figure 1.4).

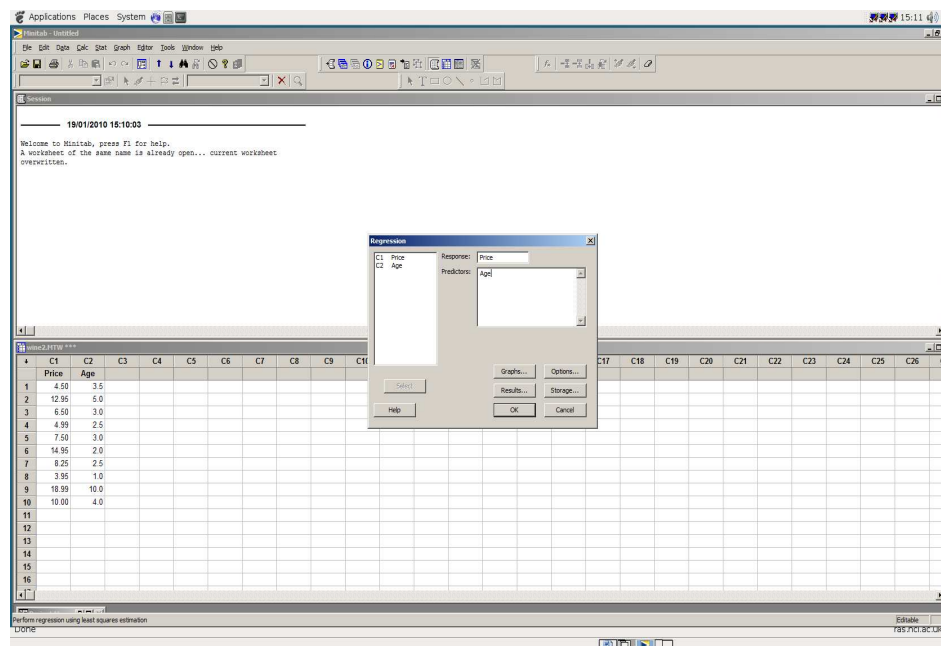


Figure 1.4: Minitab screenshot showing the Regression option

Clicking OK gives the following (edited) 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%

You might remember this from Practical 2 in Semester 2 of MAS1403. You can see that the regression equation, as calculated by Minitab, is

$$Y = 3.90 + 1.47X$$

which, to 2 decimal places, is exactly the regression equation we obtained by hand. We will return to this example later on in this Chapter to discuss the rest of the output produced by Minitab. We can also use Minitab to obtain the correlation coefficient r . Clicking on Stat-Basic Statistics-Correlation, and entering Price and Age in the Variables box, gives:

Correlations: Price, Age

Pearson correlation of Price and Age = 0.733

P-Value = 0.016

1.2.4 Testing the strength of a relationship

The aim of this Section is to bring us up-to-date with correlation and regression from MAS1403. Before moving on to further topics in regression, we will complete our revision of correlation and simple linear regression by thinking about how we can check the *significance* of any relationship between our two variables.

Example 1: car sales data

The following data shows the age in years (X) and the second-hand price (£ Y), of a sample of 11 cars advertised in a local paper. Figure 1.5 shows a scatter plot of these data.

X	5	7	6	6	5	4	7	6	5	5	2
Y	800	570	580	550	700	880	430	600	690	630	1180

You should try this yourself (by hand *and* in Minitab), but the sample correlation coefficient between age and price is -0.957 .

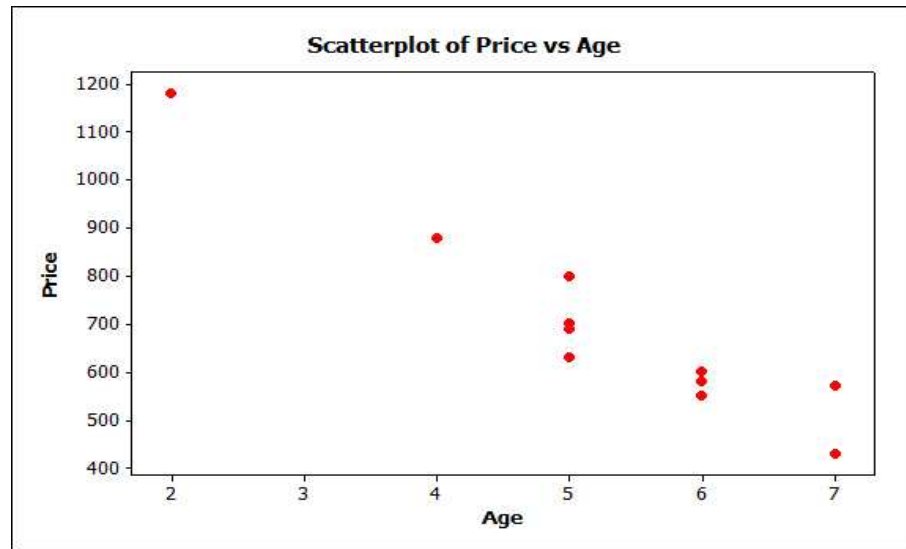


Figure 1.5: Scatter plot for the cars data

Example 2: stock exchange data

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. Figure 1.6 shows a scatter plot for these data.

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	1.5
Transactions	380	200	42	50	40	78	350	18	295	190	200	55	38	20

Again, you should be able to calculate the sample correlation coefficient as $r = 0.515$.

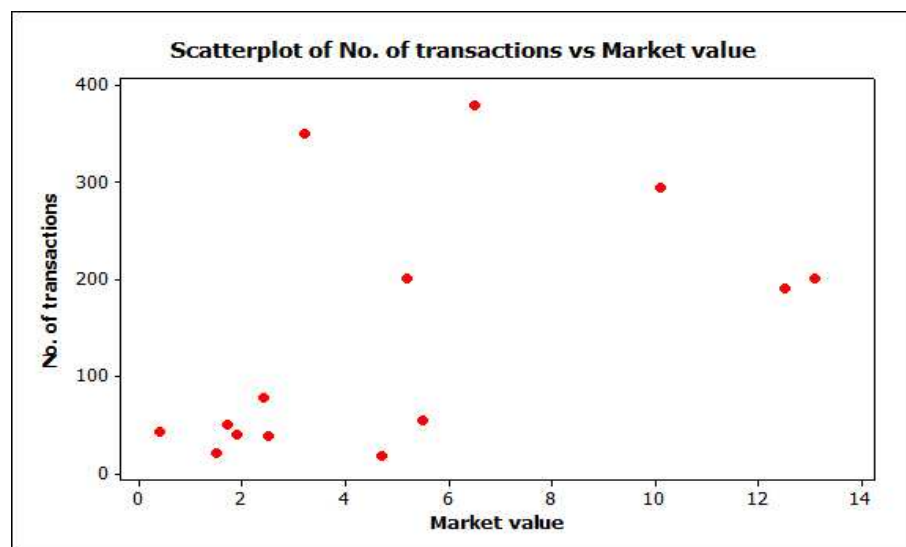


Figure 1.6: Scatter plot showing market value versus number of stock exchange transactions for 14 companies

You might agree that both scatter plots in Figures 1.5 and 1.6 indicate a relationship between the pairs of variables involved. One is negative (as the age of a second-hand car increases, its price decreases), and the other is positive (as the market value increases, *generally* the number of stock exchange transactions also increases). Indeed, this is what the calculated sample correlation coefficients tell us ($r = -0.9570$ and $r = 0.515$). The negative correlation coefficient for the car sales data is very close to -1 , indicating a very strong, linear relationship between X and Y . However, the positive correlation coefficient for the stock exchange data is less convincing; it seems far enough away from zero to suggest there *is* a relationship, but it is also probably too far away from $+1$ to indicate that this relationship is significant. So how can we proceed here? Is there a relationship or not? If so, is it really anything to “write home about”?

One way of determining whether or not a relationship between two variables is *statistically significant* is to perform a hypothesis test for the correlation coefficient. In each of the two examples above, we have calculated a *sample* correlation coefficient, since we have used the (limited) information from our (small) samples to ascertain whether or not a relationship exists between X and Y . Just like the sample mean ($\bar{x}$) is an estimator of the *population* mean (usually denoted μ and pronounced “mu”), the sample correlation coefficient (r) is an estimator of the *population* correlation coefficient (usually denoted ρ and pronounced “rho”). Just because our sample correlation coefficient r might indicate a strong linear relationship between our variables, this doesn’t automatically imply that there is a strong linear relationship between these variables in the population. In fact, r will vary from sample to sample, and we should really try to capture this variability.

Recall, from MAS1403, the five steps of *any* hypothesis test:

1. State the null hypothesis (H_0)
2. State the alternative hypothesis (H_1)
3. Calculate a test statistic
4. Use the test statistics from (3) to obtain a p -value, or at least a range for this p -value
5. Form your conclusion

Recall that the p -value summarises the hypothesis test, and informs the decision that you make (either reject or retain the null hypothesis). The p -value can be thought of as the probability of observing our data, or anything more extreme than this, if the null hypothesis is true. Therefore, the smaller the p -value, the less likely it is that we would observe the data we have if the null hypothesis is true, and so the more evidence there is to reject the null hypothesis. But how small is “small”? The standard convention is to consider anything smaller than 0.05 (or 5%) as being small enough to reject H_0 . In fact, in MAS1403, we considered the following interpretations of a p -value:

p -value	Interpretation
p bigger than 10% ($0.1 \rightarrow$)	<i>No</i> evidence against H_0
p between 5% and 10% ($0.05 \rightarrow 0.1$)	<i>Slight</i> evidence against H_0 ; not enough to reject it
p between 1% and 5% ($0.01 \rightarrow 0.05$)	<i>Moderate</i> evidence against H_0 ; reject H_0
p less than 1% ($\rightarrow 0.01$)	<i>strong</i> evidence against H_0 ; reject H_0

Table 1.2: Conventional interpretations of the p -value

If there is no (linear) relationship between our variables in the population, then this means the population correlation coefficient is zero, i.e. $\rho = 0$. If there really *is* a (linear) relationship between our variables, be it positive or negative, then $\rho \neq 0$. This forms the basis of a hypothesis test to check the significance of our sample correlation coefficient:

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

The next step is to calculate the test statistic; then obtain the p -value; then use Table 1.2 to form a conclusion. We can use **Minitab** to do this. For example, let's suppose the car sales data are stored in columns **C1** and **C2** (age and price, respectively), and the stock exchange data in columns **C3** and **C4** (market value and number of transactions, respectively). Then we can click on **Stat-Basic Statistics-Correlation**; entering **C1** and **C2** in the **Variables** box, and then clicking OK, gives the following output:

Correlations: Age, Price

Pearson correlation of Age and Price = -0.957

P-Value = 0.000

Notice that **Minitab** does not give the test statistic, just the p -value (0.000 to three decimal places). In fact, we only calculate the test statistic to get the p -value anyway, and **Minitab** gives us this automatically. Thus, we go from steps 1 and 2 (the hypotheses) directly to step 4 (p -value). **Minitab** tells us the p -value is 0.000; it might not be exactly zero, but it is zero to three decimal places. We can now use Table 1.2 to form our conclusion: since $p = 0.000$, which is less than 0.01, we have strong evidence to reject H_0 . Thus, we accept the alternative hypothesis H_1 , which says $\rho \neq 0$, i.e. there *is* a significant association between the age of a car and its sale price. Now this might seem fairly obvious, since the sample correlation coefficient ($r = -0.957$) is so close to -1 , and so it's quite likely that this will imply a significant association between these variables in the population. However, the hypothesis test is designed to take the variability of the sample correlation coefficient into account, and having done so, there is strong evidence to suggest a significant correlation between age and price in the population as a whole.

Clicking on **Stat-Basic Statistics-Correlation**, and entering **C3** and **C4** in the box **Variables** (for market value and number of transactions) and then clicking **OK**, gives the following output for the stock exchange data:

Correlations: Market value, No. of transactions

Pearson correlation of Market value and No. of transactions = 0.515
P-Value = 0.060

Notice here that our p -value (0.060, or 6%) lies between 0.05 and 0.1 (5% and 10%) and so, according to Table 1.2, we only have *slight* evidence against H_0 – not enough to reject it. So in this case, although our sample correlation coefficient suggests a positive association between market value and number of transactions (0.515), the p -value suggests that this is not far away enough from 0 to imply a significant association in the population as a whole. Thus we retain $H_0 : \rho = 0$.

Look at the **Minitab** output on page 17 for the correlation between age and price for the wine data; use the p -value to check the significance of the association here.



Testing the significance of the slope term β_1

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:

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

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 accept the alternative 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

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

Page 18 of these lecture notes gives the regression output from **Minitab** for the wine data, which is repeated below:

Regression Analysis: Price versus Age

The regression equation is

$$\text{Price} = 3.90 + 1.47 \text{ Age}$$

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

$$S = 3.58129 \quad R\text{-Sq} = 53.8\% \quad R\text{-Sq(adj)} = 48.0\%$$

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%. As always, we can now use Table 1.2 to interpret: Since this lies between 1% and 5%, we have *moderate* evidence to reject H_0 , and so we 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. Thus, Age *is* an important predictor of the price of a bottle of wine.

What about the rest of the output?

Notice from the above **Minitab** output we are told that $s = 3.58129$; an assumption of this simple linear regression model is that ϵ , which represents the “scatter” about the regression line, follows a Normal distribution with zero mean and constant variance σ^2 ,

i.e. $\epsilon \sim N(0, \sigma^2)$. Minitab estimates the standard deviation of this Normal distribution, which in this case is $s = 3.58129$. The vertical “gaps” between the regression line and the observed data points are known as the *residuals* $\epsilon_1, \epsilon_2, \dots, \epsilon_n$, and so, in our wine example, these residuals have estimated (constant) variance of $3.58129^2 = 12.826$. The diagram below should help to illustrate this.



The value R^2 estimates the percentage of variability in the Y -data that is explained by X ; basically, the closer this is to 100%, the stronger the relationship between X and Y . Here we see that about 54% of the variability in wine price is explained by the age of the wine.

Words of warning

Just because a fitted regression model tells us that X is useful in predicting Y , it doesn't mean that X *causes* Y . 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 buy 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.

It should also be emphasised that the hypothesis test for the slope is only valid if the assumptions made at the start of this Chapter are true, i.e. that the correct model for our data is

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

where $\epsilon \sim N(0, \sigma^2)$. Later in this chapter we will consider ways of assessing the viability of our regression model.

1.3 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 be predictable from 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*.

Example: Back to the price of wine

Recall the example of Section 1.2 where we investigated the relationship between the price of wine and its age. However, at the top of page 6 it was discussed that the price of a bottle of wine might also depend on other factors, for example, the amount of rainfall during the 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 below in Table 1.3:

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

Table 1.3: Price of bottles of wine, with corresponding Age, total rainfall and average temperature during growing season

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 is

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

as before, ϵ is our “random error” term, and is assumed to be Normally distributed with zero mean and variance σ^2 , 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 1.2.2.

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 1.2.2 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 Table 1.3 in the first four columns of a **Minitab** worksheet (see Figure 1.7). 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 below Figure 1.7.

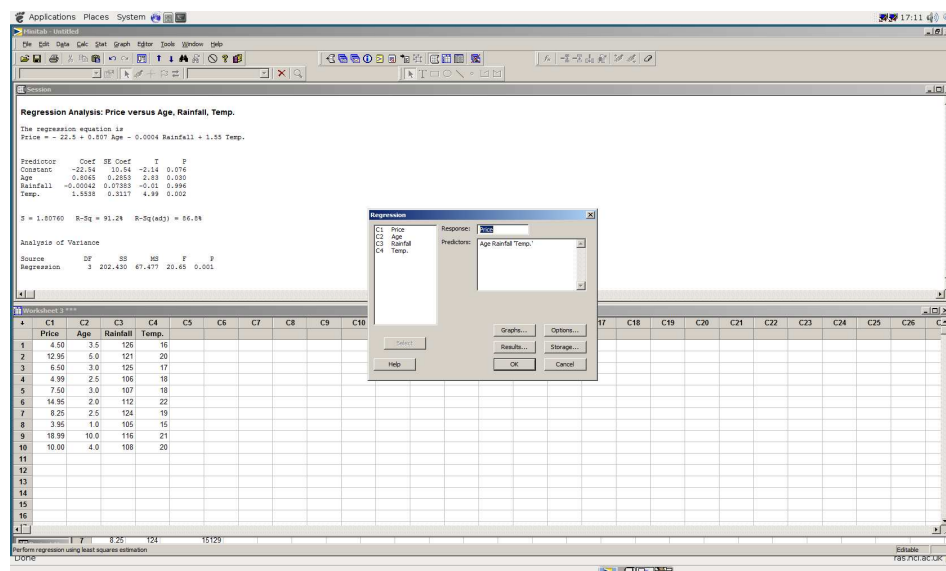


Figure 1.7: Minitab screenshot showing the Regression option

Regression Analysis: Price versus Age, Rainfall, Temp.

The regression equation is

$$\text{Price} = -22.5 + 0.807 \text{ Age} - 0.0004 \text{ Rainfall} + 1.55 \text{ Temp.}$$

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

$$S = 1.80760 \quad R\text{-Sq} = 91.2\% \quad R\text{-Sq(adj)} = 86.8\%$$

Thus, the full (multiple) regression model is:

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

where $\epsilon \sim N(0, 1.80760^2)$, and 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).

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 in Figure 1.8.

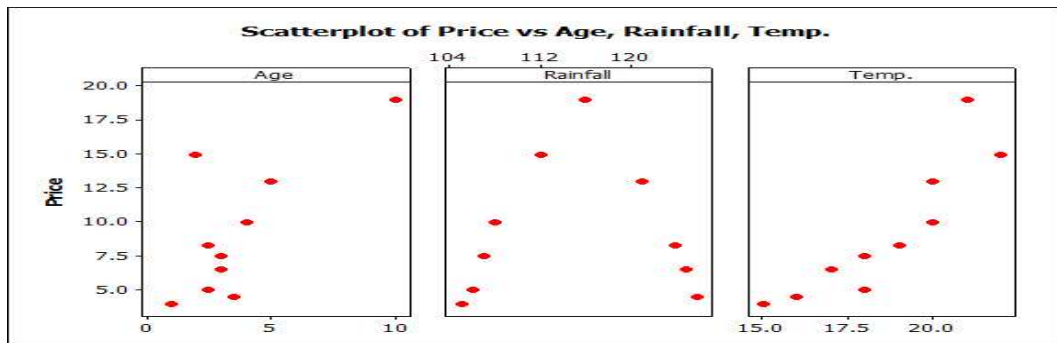


Figure 1.8: Scatter plots for the wine data

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!*

1.3.1 Testing the importance of our predictor variables

Recall Section 1.2.4, 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 1.2.4 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 bottom of page 20 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:

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

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 accept 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 17). 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:

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

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:

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

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 accept 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 below:

Regression Analysis: Price versus Age, Temp.

The regression equation is

$$\text{Price} = -22.6 + 0.806 \text{ Age} + 1.55 \text{ Temp}$$

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

$$S = 1.67352 \quad R\text{-Sq} = 91.2\% \quad R\text{-Sq}(\text{adj}) = 88.6\%$$

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, and $\epsilon \sim N(0, 1.67352^2)$. 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).

Notice that the R^2 value in this analysis is 91.2%, which is exactly the same as before. Thus, excluding rainfall has not resulted in a deterioration of this statistic and the amount of variation in Y explained by X .

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?



1.3.2 The F -test: An overall test of the model

Notice that the regression outputs shown in these notes so far have been *edited*; if you’ve been watching closely in class, or if you’ve tried some of this stuff in **Minitab** yourself, you might have noticed that **Minitab** gives some more output underneath the results that we’ve looked at so far, but we’ve just ignored this. For example, the output on page 23 shows the result of fitting a multiple linear regression to the wine sales data, using the age of the wine and average afternoon temperature during the growing season as predictor variables. We eliminated the rainfall variable from the model, and were left with a model for which all the other predictor variables were significant. Now the *full* output from **Minitab** is shown below:

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%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	202.43	101.21	36.14	0.000
Residual Error	7	19.60	2.80		
Total	9	222.03			

Source	DF	Seq SS
Age	1	119.43
Temp.	1	83.00

The figure labelled F in the ANOVA table of the **Minitab** output is the test statistic for an overall test of the whole model. In this case, where we have two predictor variables (age and temperature) it tests the null hypothesis:

$$H_0 : \beta_1 = \beta_3 = 0$$

against

$$H_1 : \text{at least one of } \beta_1 \text{ and } \beta_3 \text{ is not zero}$$

Notice that when H_0 is true the model is just $Y = \beta_0 + \epsilon$ and so we can predict the price of a bottle of wine just as well without the predictor variables (age and temperature). When H_1 is true a combination of one or more of the predictor variables is useful in predicting Y . When H_0 is true the test statistic comes from the F -distribution, which you should have met last term when you studied ANOVA with Dr. Ruto. We can now refer our test statistic (here $F = 36.14$) to statistical tables to obtain a range for the p -value, or just look at the p -value as given by **Minitab**! Here, we see the p -value is very small (0.000 to three d.p.!) and so we have strong evidence to reject H_0 . Therefore we accept the alternative H_1 : some, or *all*, of the predictor variables used in the fit are useful in predicting the price of a bottle of wine, so our model *is* useful for prediction.

1.3.3 Another example

On a small island the government would like to be able 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

Table 1.4: Number of mortgages issued over a ten-year period with corresponding average personal income and interest rates

2. Use **Minitab** to fit the model in (1). Write down 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, write down 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. Perform an overall test of your “final” model using the F -test.

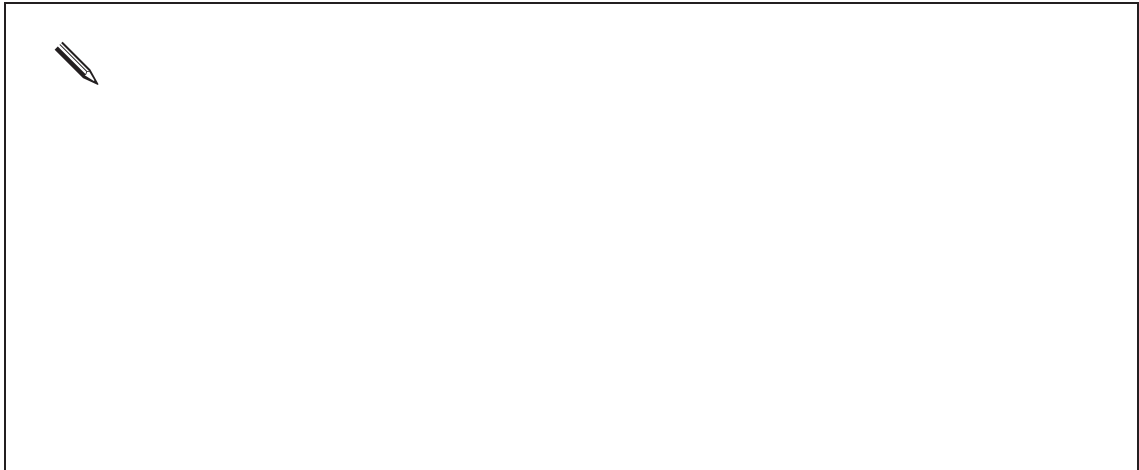


8. This year (2012), the average income is forecast to rise to 4.9 million units of the local currency and the interest rate is set to rise to 5%. How many mortgages can the state mortgage company expect to issue this year?



1.3.4 Checking the model

The t -tests we have used to check the importance of each predictor variable in the multiple linear regression, and the F -test which was used to check the model *overall*, are only valid if the assumption that the random term ϵ is Normally distributed with mean zero and constant variance σ^2 is true. The picture below might help you to visualise this:



We can attempt to check this assumption by using a graphical procedure in Minitab. In the **Stat-Regression-Regression** option, clicking on **Graphs** opens another box (see the screenshot in Figure 1.9). Checking the boxes that say **Histogram of residuals** and **Residuals versus fits** produces the graphs shown in Figure 1.10: the first can be used to check the assumption of Normality, the second can be used to check the assumption of constant variance.

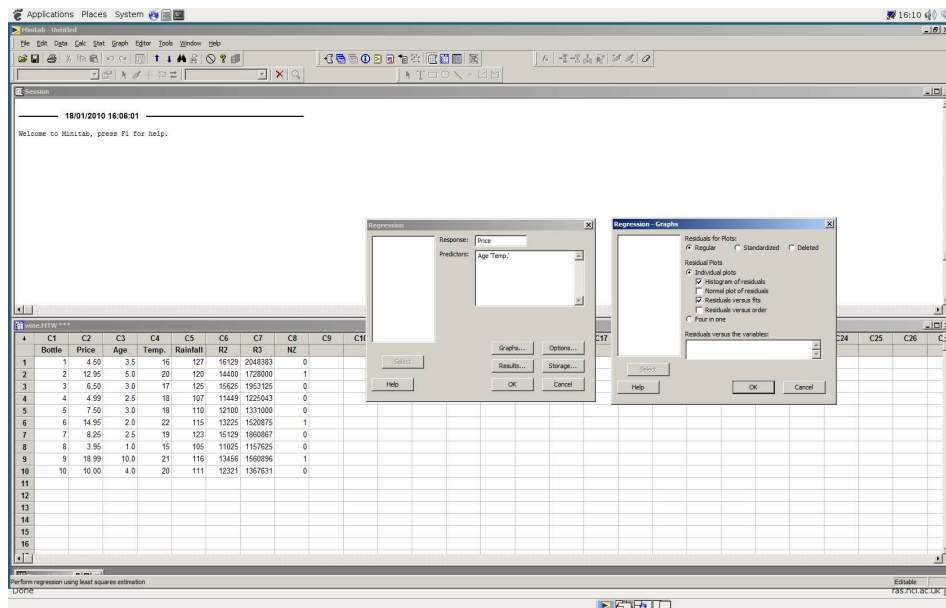


Figure 1.9: Minitab screenshot showing the **Graph** option within **Regression**

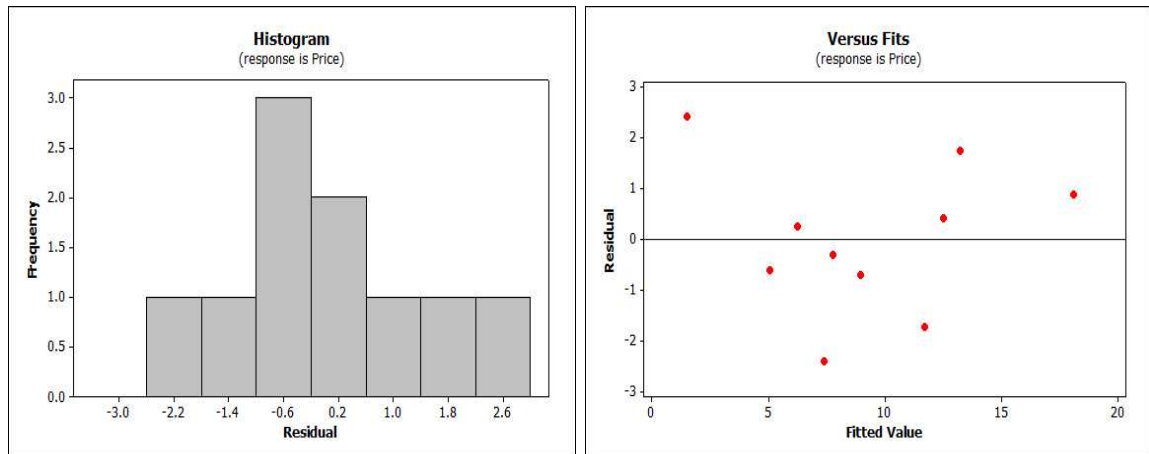


Figure 1.10: Histogram of residuals (left) and plot of residuals versus fitted values (right)

From the left-hand plot in Figure 1.10, we can see the assumption of Normality seems plausible; the histogram of the residuals peaks at about zero and then tails off reasonably symmetrically either side of this. However, with only ten data points as we have here, such plots can often be difficult to interpret. Also, getting **Minitab** to use different interval widths to construct the histogram can produce very different looking shapes! The plot on the right-hand-side in Figure 1.10 can be used to check the assumption that the residuals are distributed with constant variance. For this to be true, we would not expect to see any pattern in this plot at all, just “random scatter”. Indeed, there does not seem to be any pattern here, just a random scatter of points, and so this assumption is plausible. Thus, overall, the assumptions regarding the residuals seem to have been adequately verified.

1.4 Further topics in regression

1.4.1 Indicator variables

Sometimes qualitative (categorical) variables are used as predictor variables; this can be accomplished by using indicator variables (variables with two “states”, usually 0/1). Such data often appear in questionnaires or market research, when “tick boxes” might be used to make the questionnaire easier to complete (e.g. Male/Female, Age groupings, level of agreement with a statement). If a variable has 2 “states” (e.g. Male/Female), only one indicator variable (call it X_1) is required:

	X_1
State 1 (e.g. Male)	0
State 2 (e.g. Female)	1

A variable with 3 “states” (e.g. Agree/Not bothered/Disagree) requires two indicator variables (say X_1 and X_2):

	X_1	X_2
State 1 (e.g. Agree)	1	0
State 2 (e.g. Not bothered)	0	1
State 3 (e.g. Disagree)	0	0

Generally, a qualitative variable with k “states” requires $k - 1$ indicator variables, each taking the values 0 and 1.

Example: Back to the price of wine

Recall that, so far, we have investigated the importance of age, rainfall and temperature in the selling price of a bottle of wine at the online wine merchants `tanners-wines.co.uk`. It is believed that, in the U.K., wine from New Zealand is generally more expensive than other wines. To investigate, we now add an indicator variable to the original dataset, which takes the value 1 if the bottle of wine was from New Zealand, and 0 otherwise:

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
New Zealand? (X_4)	0	1	0	0	0	1	0	0	1	0

Table 1.5: Wine data updated to include the “New Zealand” indicator variable

Perform a regression analysis to find a suitable multiple regression model for the updated dataset. Suppose you own a vineyard in the Marlborough region of New Zealand. During the growing season in 2004, the Marlborough region of New Zealand experienced a total of 115mm of rainfall, and the average afternoon temperature was 17°C. How much can we expect a bottle of wine to sell for in the U.K., this year, in 2010?

 ...solution to last question...

1.4.2 Non-linear regression

In all of the multiple regressions performed thus far on the wine sales dataset, the rainfall variable (total rainfall recorded during the growing season, in mm) has always been excluded from the model as not being an important (linear) predictor for the price of a bottle of wine. However, as Figure 1.11 shows, there is clearly a relationship between rainfall and price:

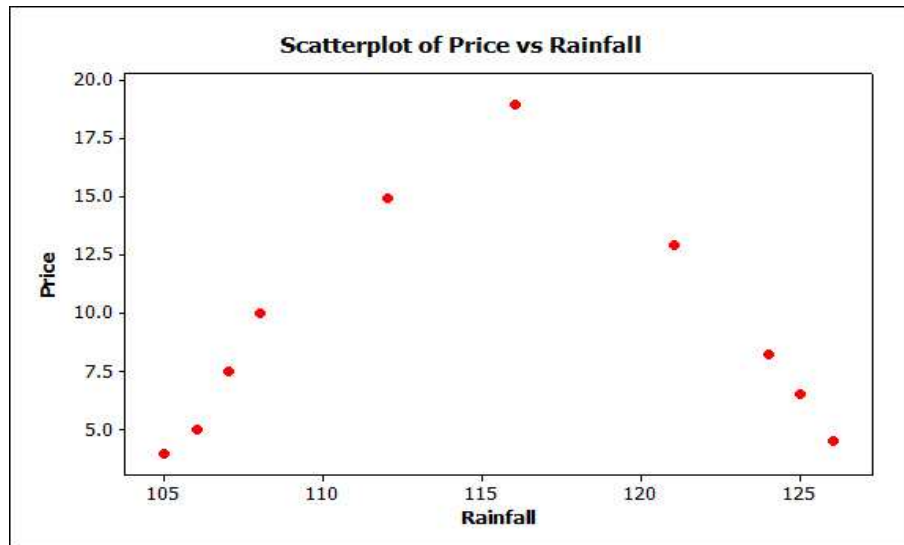


Figure 1.11: Scatter plot showing price of bottles of wine against rainfall totals

Suppose we are interested in *only* the relationship between rainfall and price (so we forget about age, temperature and whether or not the wine is from New Zealand). How can we proceed? A straightforward simple linear regression is not appropriate because we have both an increasing and decreasing relationship (price increases with rainfall until about 117mm; then price decreases). Also, there appears to be a curved relationship to the left, and perhaps the right, of 117mm.

Think back to your GCSE maths days, and think back to drawing graphs of *quadratic functions*. For example, in the space below, sketch the graphs of:

1. $y = x^2$



$$2. y = -x^2$$



$$3. y = 10 + 4x - 5x^2$$



It might be that we can capture the non-standard relationship between rainfall and price with a quadratic curve instead of a straight line. If we have price (Y) and rainfall (X) in columns C1 and C2 of a **Minitab** worksheet, then for a *quadratic regression* (as opposed to the standard *linear* regression) we also need rainfall² (X^2) in another column (say column C3). To do this, we click on **Editor-Enable commands**. This activates the command-driven interface of **Minitab**; we can now type commands into the **Session** window of

Minitab directly, by-passing the menus. Notice the **Minitab** command prompt will now appear in the **Session** window: **MTB >**. Typing

```
MTB > LET C3=C2**2
```

squares all of the X values in column **C2** (rainfall) and puts the results in column **C3**. Figure 1.12 shows a Minitab screenshot where this has been done; notice I have labelled column **C3** as **Rainfallsquared**.

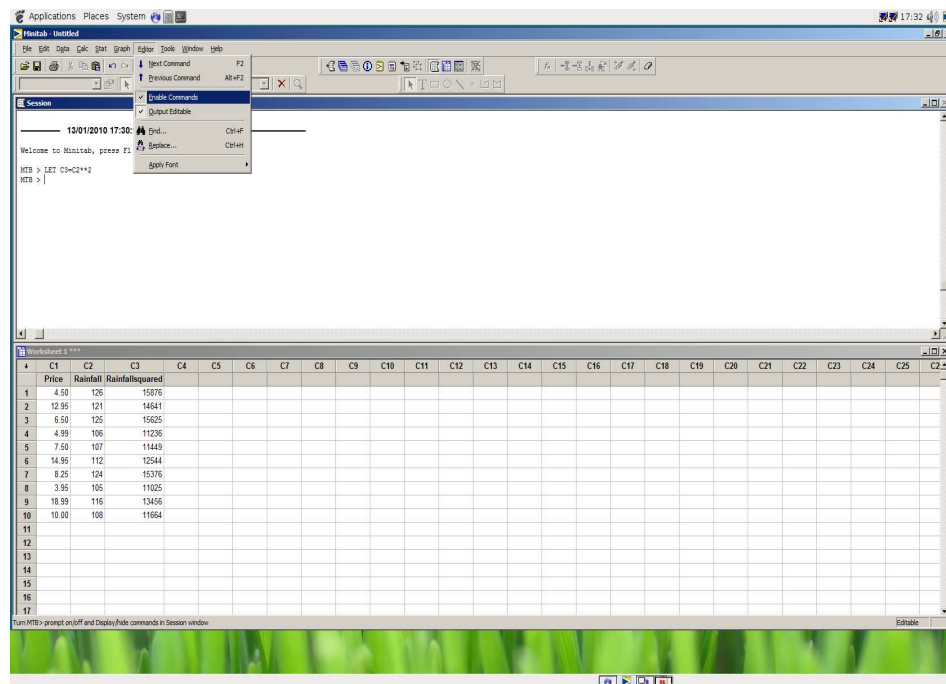


Figure 1.12: Minitab screenshot showing how to square the rainfall variable

We now perform a multiple linear regression in the usual way, with Y :Price, X_1 :Rainfall and X_2 :Rainfall²; thus, in Minitab, we click on **Stat-Regression-Regression**, and then enter **C1** as the **Response** variable and **C2** and **C3** as the **Predictor** variables. Clicking **OK** gives the following (edited) output:

Regression Analysis: Price versus Rainfall, Rainfallsquared

The regression equation is

$$\text{Price} = -1658 + 29.0 \text{ Rainfall} - 0.125 \text{ Rainfallsquared}$$

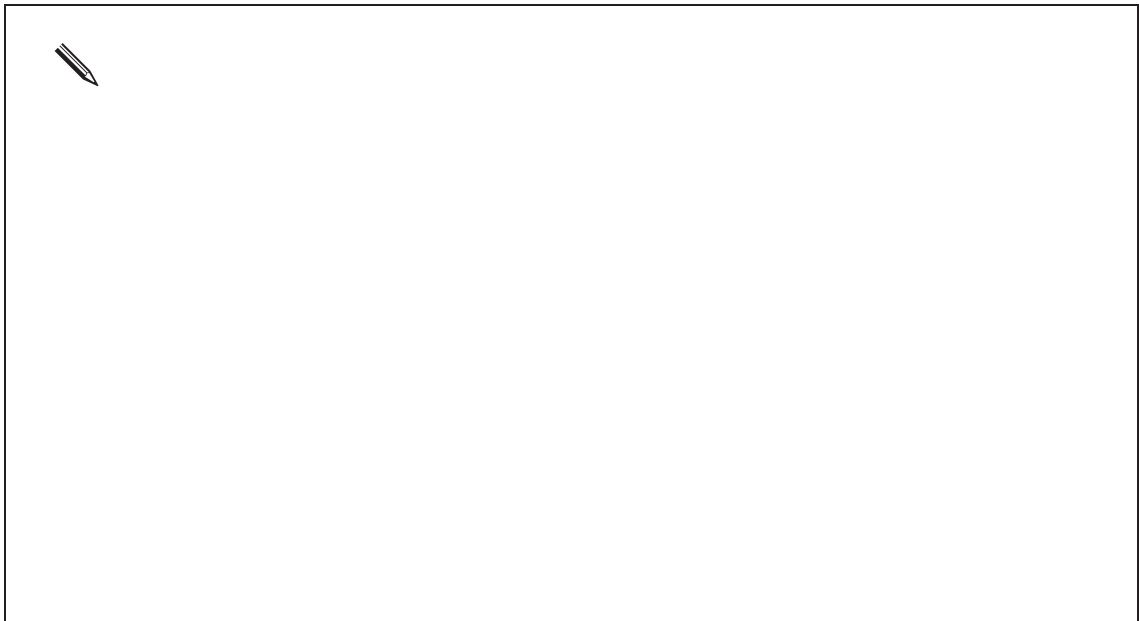
Predictor	Coef	SE Coef	T	P
Constant	-1658.1	102.6	-16.16	0.000
Rainfall	28.967	1.784	16.24	0.000
Rainfallsquared	-0.125202	0.007721	-16.22	0.000

$$S = 0.904747 \quad R\text{-Sq} = 97.4\% \quad R\text{-Sq(adj)} = 96.7\%$$

Notice that both $X_1 = \text{Rainfall}$ and $X_2 = \text{Rainfall}^2$ are important predictor variables in the model, since both β_1 and β_2 have small p -values (so small they're 0.000 to 3 d.p.!) and so we would reject both of the null hypotheses: $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. So we have found a regression model which caters for the non-standard relationship between price and rainfall!

In Minitab, clicking on **Stat-Regression-Fitted line plot** and entering **C1** as the **Response** variable and **C2** as the **Predictor** variable, *and* selecting **Quadratic** as the **Type of regression model** (see the screenshot in Figure 1.13) gives the scatter plot with the superimposed fitted quadratic regression line shown in Figure 1.14. Although this fit is not perfect, and there appears to be an outlier, the parameters in the model are significant and at least now we are not simply dismissing the predictive importance of rainfall on the price of a bottle of wine.

Suppose we observe a total rainfall of 125mm during the growing season. What price can we expect to sell a bottle of wine for?



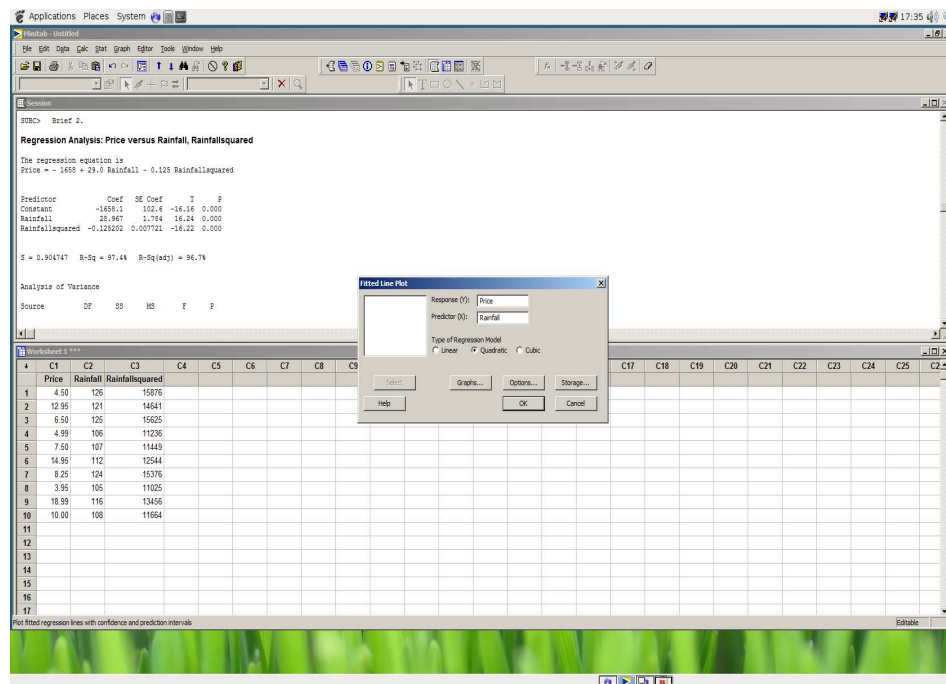


Figure 1.13: Minitab screenshot showing how to produce a fitted quadratic regression plot

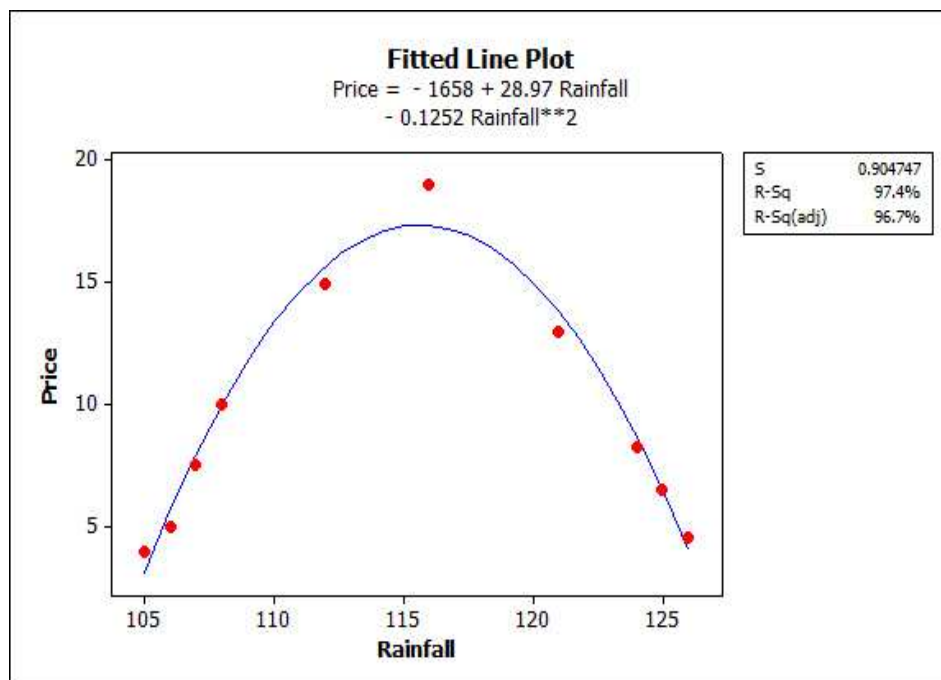


Figure 1.14: Scatter plot of price versus rainfall with superimposed fitted quadratic regression line

1.5 Analysing a binary response: logistic regression

We now return to the motivating example on page 6 of these notes. Suppose the marketing team at *Mars* are interested in the ability of cinema-goers to recall their brand using *product placement* during a film. Further, they think there might be a relationship between the chance of someone being able to recall their brand and the time in the film at which the product placement occurred. To investigate, 25 volunteers took part in a marketing experiment. Initially, the volunteers knew nothing about the aims of the experiment; after they each watched a film of length $2\frac{1}{4}$ hours, they were asked if they could recall the brand that had been “placed” in their film. The product placement for each volunteer happened at a different time during the film. The results are shown in Table 1.6 below, where X is the time, from the start of the film, at which the product placement occurred and Y takes the value 1 if the volunteer *could* recall the brand, and 0 if they *could not*.

Volunteer	1	2	3	4	5	6	7	8	9	10	11	12	13
X (minutes)	10	15	20	25	30	35	40	45	50	55	60	65	70
Y	0	0	0	0	0	0	1	0	0	1	0	1	0
Volunteer	14	15	16	17	18	19	20	21	22	23	24	25	
X (minutes)	75	80	85	90	95	100	105	110	115	120	125	130	
Y	1	1	1	1	0	1	1	1	0	1	1	1	

Table 1.6: Product placement dataset

Comment on the relationship between X and Y as shown by the data in Table 1.6.



Notice that the variable of interest here – whether a volunteer can/cannot recall the *Mars* brand – is not like the variable of interest in the wine sales example (price of a bottle of wine) or in the mortgage company example (number of mortgages approved). The variable of interest is now ‘binary’ – i.e. can only take one of two values; we use 0 for “no” and 1 for “yes”. We have already thought about how to perform a regression analysis when one of the *predictor* variables is binary (see Section 1.4.1), but not when the main *response* variable takes this form.

Why can't we use simple linear regression to predict Y using X ? Shown below is the standard Minitab output for a simple linear regression of Y on X :

Regression Analysis: Y versus X

The regression equation is

$$Y = -0.062 + 0.00831 X$$

Predictor	Coef	SE Coef	T	P
Constant	-0.0615	0.1821	-0.34	0.738
x	0.008308	0.002312	3.59	0.002

S = 0.416868 R-Sq = 35.9% R-Sq(adj) = 33.2%

The estimated simple linear regression equation is

$$Y = -0.062 + 0.00831X + \epsilon.$$

Figure 1.15 shows a scatterplot of the product placement data with this fitted regression line superimposed.

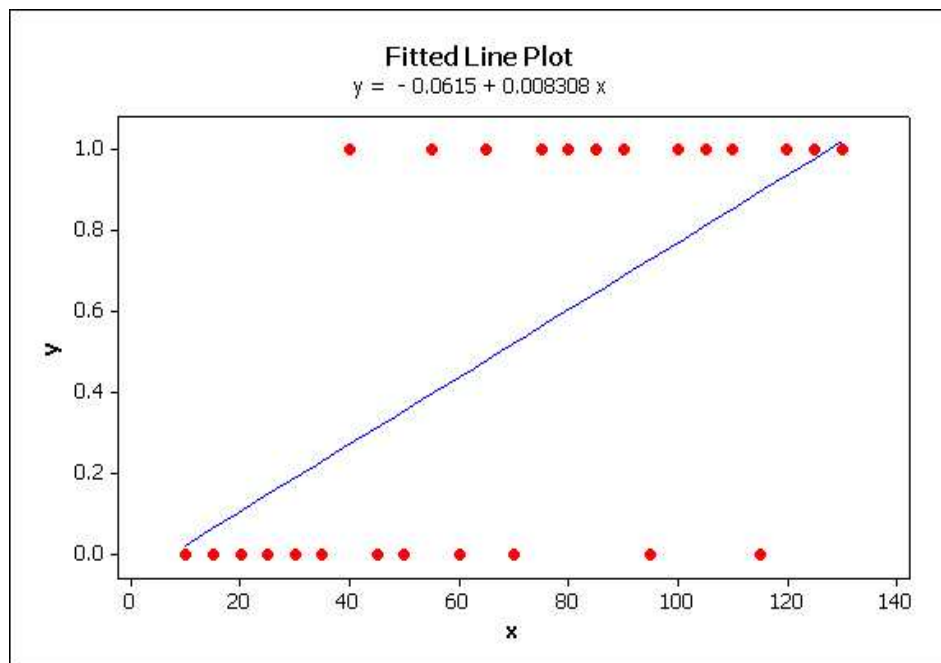


Figure 1.15: Scatter plot of *Mars* brand recall versus the time at which the product placement occurred, with superimposed fitted simple linear regression line

If, in a new film, a *Mars Bar* is ‘placed’ after 97.5 minutes, will a cinema-goer be able to recall the brand?

We could substitute $X = 97.5$ into the regression equation to get:

$$\begin{aligned} Y &= -0.062 + 0.00831 \times 97.5 \\ &= 0.748. \end{aligned}$$

But how do we interpret this? The response variable Y must either be 1 or 0 – a cinema-goer will either be able to recall the brand or they will not, there is nothing ‘in between’! One of the problems with trying to use simple linear regression when the response variable is binary is that the fitted or predicted values found using the estimated regression equation cannot arise as observations. Indeed, we could get any value at all for Y using the above regression equation! Another problem is that the residuals cannot possibly be normally distributed with constant variance. Clearly a different approach is needed, and this is now considered.

1.5.1 The logistic regression equation

In simple linear regression, we know that the regression equation is given by

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

The mean of the ϵ is assumed to be zero, and so the mean of Y ($E[Y]$ or the “expectation” of Y), is just

$$E[Y] = \beta_0 + \beta_1 X.$$

When Y is binary, statistical theory (as well as practice) has shown that the relationship between $E[Y]$ and X can better be described by the following nonlinear equation:

$$E[Y] = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

If the two values of the dependent variable Y are coded as 0 or 1, the value of $E[Y]$ in the above equation provides a *probability* that $Y = 1$ given a particular value for the predictor variable X . Because of the interpretation of $E[Y]$ as a probability, the *logistic regression equation* is often written as follows:

$$E[Y] = \Pr(Y = 1|X).$$

We can use **Minitab** to estimate the logistic regression equation – i.e. obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ – and thus estimate the probability that Y takes the value 1.

Example: Back to the product placement example

Suppose the data from Table 1.6 are stored in a **Minitab** worksheet (time of the product placement (X) in column 1 and whether or not a volunteer was able to recall the brand

($Y = 0$ for “no” and 1 for “yes”) in column 2. Then clicking on **Stat-Regression-Binary logistic regression** opens the dialogue box shown in Figure 1.16. Entering column 2 as the **Response** variable and column 1 in the **Model** box, and clicking **OK**, gives the (edited) output shown overleaf.

From this output we can see that our estimates of β_0 and β_1 are

$$\hat{\beta}_0 = -2.99147 \text{ and } \hat{\beta}_1 = 0.0445355,$$

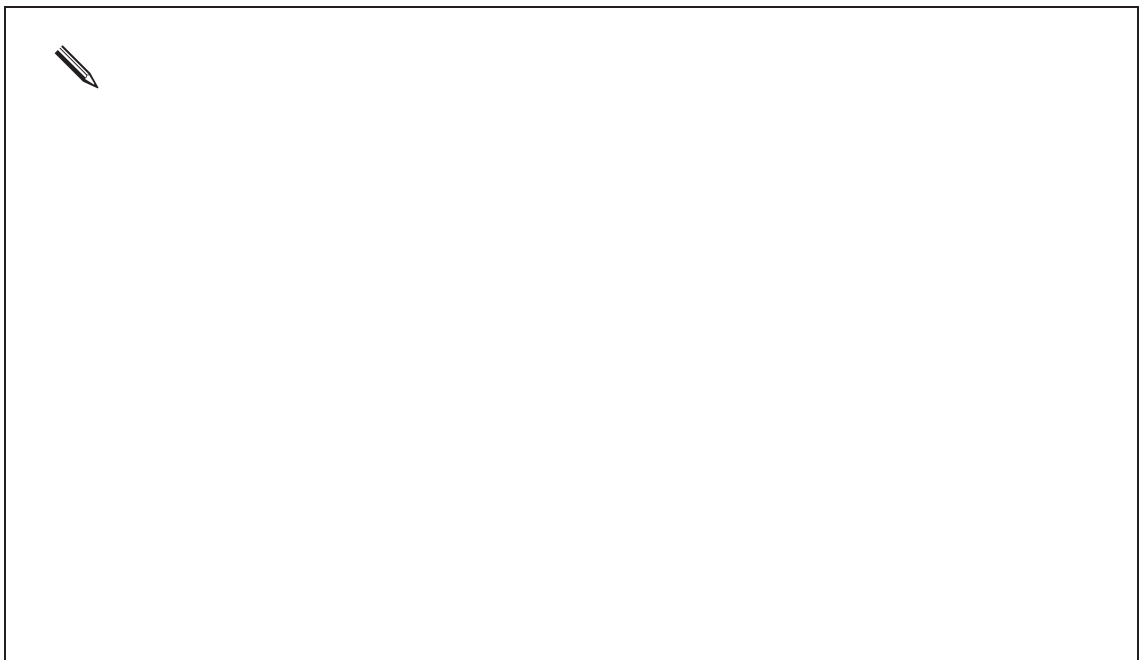
which gives an estimate of the logistic regression equation as

$$E[Y] = \Pr(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{e^{-2.99147 + 0.0445355X}}{1 + e^{-2.99147 + 0.0445355X}}.$$

From this, we can now estimate the probability that a cinema-goer will be able to remember the *Mars* brand after the film. Returning to the previous example, suppose the product is placed after 97.5 minutes. Substituting $X = 97.5$ into the above equation gives:

$$\begin{aligned} \Pr(Y = 1|X) &= \frac{e^{-2.99147 + 0.0445355 \times 97.5}}{1 + e^{-2.99147 + 0.0445355 \times 97.5}} \\ &= 0.794. \end{aligned}$$

A *Mars Bar* is placed after just 11 minutes in another film. How likely is it that a cinema-goer will be able to remember the brand?



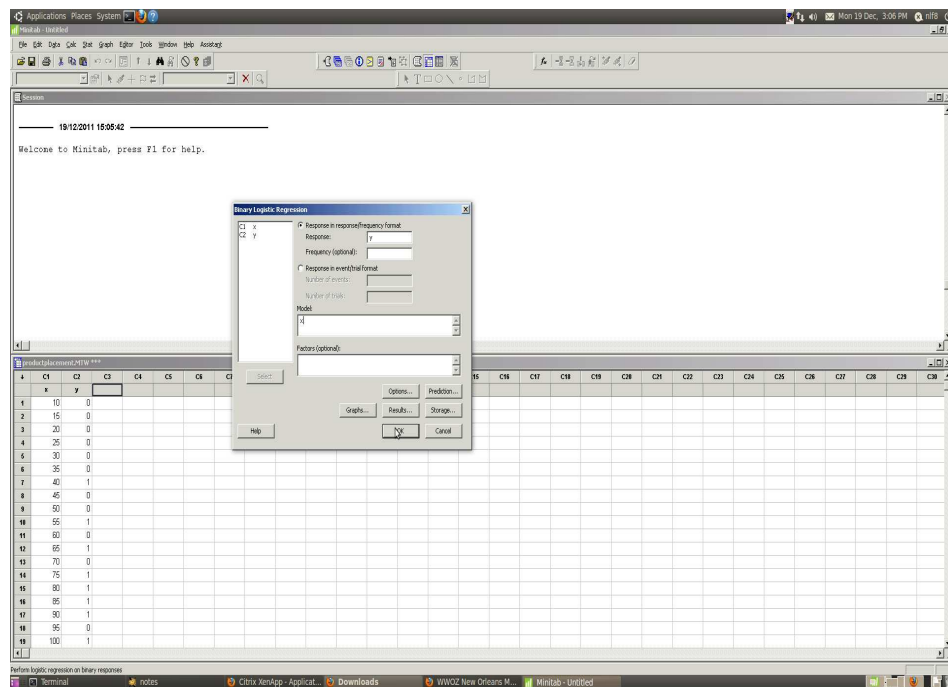


Figure 1.16: Minitab screenshot showing the Binary logistic regression option

Binary Logistic Regression: Recalled? versus time

Link Function: Logit

Response Information

Variable	Value	Count
Recalled?	1	13 (Event)
	0	12
Total		25

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-2.99147	1.28221	-2.33	0.020			
x	0.0445355	0.0173118	2.57	0.010	1.05	1.01	1.08

1.5.2 Testing the importance of the predictor variable

As in Section 1.3.1, we can use the output from **Minitab** to test the importance of the predictor variable in our logistic regression model. The parameter β_1 is the coefficient of the predictor variable X (in this case: time into the film, in minutes), and so our null and alternative hypotheses are:

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

From the **Minitab** output on the previous page, the p -value associated with the predictor variable is 0.010, or 1%. Since this is less than 5%, we have evidence to reject H_0 in favour of the alternative H_1 : it appears that the time at which a *Mars Bar* is placed in the film *is* an important predictor of whether or not a cinema-goer will be able to recall the brand after the film.

1.5.3 Logistic regression using multiple predictor variables

We can combine logistic regression with our knowledge of multiple regression to fit a *multiple logistic regression model*. For instance, in the product placement example, the probability of someone being able to recall a brand might also depend on the age of that person, the sex of that person, the type of film being watched, etc. Table 1.7 shows such information, as well as the original predictor variable, for each of the 25 volunteers in the study. Specifically: X_1 is the original predictor variable (time, in minutes, into the film at which the product placement occurred); X_2 is an indicator variable taking the value 1 if the volunteer is female/0 if the volunteer is male; X_3 is the age of the volunteer in years; X_4 takes the value 1 if the film was violent/sexually explicit and 0 otherwise.

Volunteer	1	2	3	4	5	6	7	8	9	10	11	12	13
X_1 (minutes)	10	15	20	25	30	35	40	45	50	55	60	65	70
X_2 (Sex)	0	0	0	0	1	0	0	1	0	0	0	1	1
X_3 (Age)	31	39	73	75	58	47	73	62	55	50	56	60	54
X_4 (Violent/Sexual)	1	1	1	1	1	1	0	1	0	0	0	0	1
Y	0	0	0	0	0	0	1	0	0	1	0	1	0
Volunteer	14	15	16	17	18	19	20	21	22	23	24	25	
X_1 (minutes)	75	80	85	90	95	100	105	110	115	120	125	130	
X_2 (Sex)	1	0	1	0	1	1	1	1	1	1	1	1	
X_3 (Age)	38	19	55	33	46	23	25	19	47	22	21	18	
X_4 (Violent/Sexual)	0	1	0	1	0	0	1	1	1	1	1	1	
Y	1	1	1	1	0	1	1	1	0	1	1	1	

Table 1.7: Product placement dataset, with additional potential predictor variables

1. Using logistic regression, and the techniques of multiple regression discussed in Section 1.3, find the most suitable model for estimating the probability that a cinema-goer will be able to recall the *Mars* brand after a film.
2. *Madagascar 3: Europe's Most Wanted* is a children's film due for release in 2012. If *Mars* pay to use product placement to advertise their *Mars Bar* exactly one hour into the film, estimate the probability that a 30 year old man watching the film will be able to recall the brand afterwards.





...continued...