

ACE2013: Statistics for Marketing and Management

Dr. Lee Fawcett

Semester 2: 2013—14

Welcome to ACE2013: Part 2

Welcome! I've decided to call this part of the course "Statistical Business Modelling" – from now until the Summer, we will review and extend on the ideas introduced in the last part of MAS1403:

1. Correlation and Regression
2. Time series and Forecasting
3. Business Modelling

Contact details

My name: Dr. Lee Fawcett

Office: Room 2.07 Herschel Building

Phone: 0191 222 7228

Email: lee.fawcett@ncl.ac.uk

www: 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 much more difficult. But don't worry, we will use the Minitab package extensively, especially in the first two topics. I will probably use Minitab in every lecture. *You will be expected to interpret computer output in the exam.*

Differences between this course and MAS1403

You will have more to do in lectures

I will expect you to take more notes in lectures, and really listen carefully! You can't coast through this course! And always have your calculator to hand.

You will have more to do outside of lectures

You should read your notes before coming to lectures. I know I said this last year, but this is vital this year!

Exam is not open-book

Self-explanatory. Though a formulae sheet will be included in the exam paper and we'll tell you what you'll need to memorise.

Timetable of events

Lectures

- Every week – annoyingly, the time/venue seem to change alot!
- Two hours – please don't be late, and don't turn up for half the lecture!

Computer practicals

- Every fortnight, starting in the third week of term
- Again – time/venues seem to change
- Register!

Timetable of events

Revision

One two hour revision session at the end of the Semester, in place of the lecture.

Office hour

I will always be available to see students on Wednesdays, 4–6, in my office.

Assessment

Exam (60%)

- 2 hours in May/June
- Covers the entire course
- Not open-book!

CBAs (20%)

- Four CBAs – you've already done two of them
- Semester 2 deadlines on the course website

Practical work (20%)

- Some questions in each practical session will be “starred”
- You will hand solutions to all of these starred questions in at the end of term (May)

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

Motivating example: Product placement

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.

Motivating example: Product placement



Motivating example: Product placement



Motivating example: Product placement



Motivating example: Product placement



Motivating example: Product placement

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.

Motivating example: Product placement

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”.

Motivating example: Product placement

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
- The product's prior exposure might influence the viewer's ability to recall the brand

Motivating example: Product placement

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.

Statistics has a vital role to play here.

We return to this example in Section 1.5.

Introduction

In this part of the course, we will:

- re-visit correlation and simple linear regression from MAS1403;
- extend these ideas to consider **multiple linear regression**;
- consider **non-linear regression**;
- consider what happens when we have a binary response using **logistic regression**;
- think about how to 'build' the most suitable model for our data.

Throughout, we will use the computer package Minitab extensively.

Bivariate data

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.

Bivariate data

These data could have arisen from:

- a random sample of n individuals from a population;
- 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.

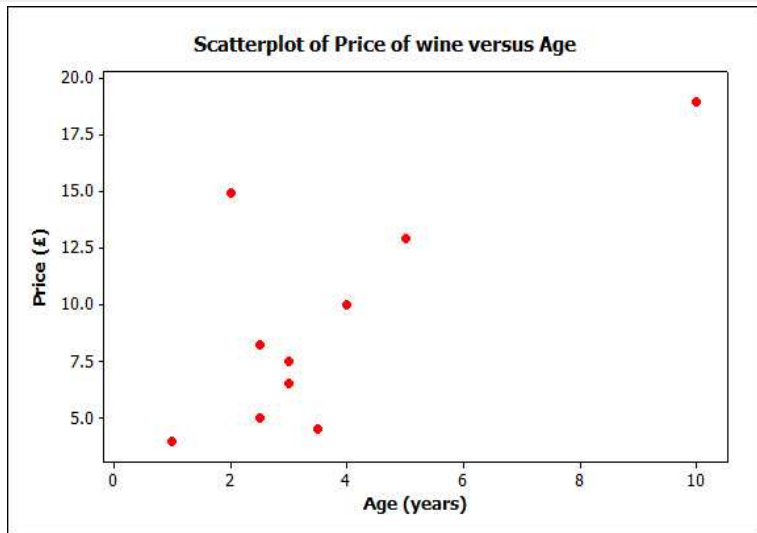
Example: Price of wine

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.

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

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

Example: Price of wine



Example: Price of wine

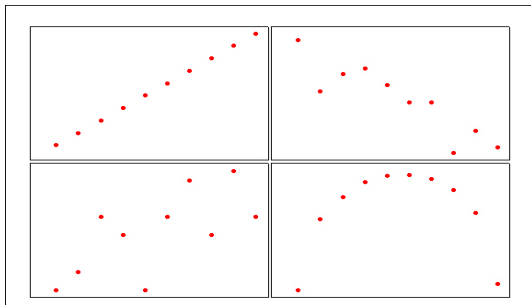
Looking at the scatter plot (and maybe just the raw data themselves!), what can you say about the relationship between age of wine and price?

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

Quantifying the relationship: Correlation

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

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



Quantifying the relationship: Correlation

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 subjective; *quantifying* such relationships *numerically* can circumvent this problem of subjectivity.

Quantifying the relationship: Correlation

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

The formula for r is

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

where

$$S_{XY} = \left(\sum xy \right) - n\bar{x}\bar{y},$$

$$S_{XX} = \left(\sum x^2 \right) - n\bar{x}^2 \quad \text{and}$$

$$S_{YY} = \left(\sum y^2 \right) - n\bar{y}^2,$$

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

Quantifying the relationship: Correlation

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

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

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

Can you estimate the value of r for the wine age/price data? And for the four datasets shown in Figure 1.2?

Quantifying the relationship: Correlation

Using the wine price/age data, we can calculate the value of r . Thinking back to MAS1403, the easiest way to do this is to draw up a table:

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

Quantifying the relationship: Correlation

Then we have:

$$\begin{aligned}\bar{x} &= \frac{36.5}{10} \\ &= 3.65 \quad \text{and}\end{aligned}$$

$$\begin{aligned}\bar{y} &= \frac{92.58}{10} \\ &= 9.258.\end{aligned}$$

Quantifying the relationship: Correlation

We can now calculate S_{XY} , S_{XX} and S_{YY} :

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

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

Quantifying the relationship: Correlation

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

Quantifying the relationship: Correlation

Thus,

$$\begin{aligned} r &= \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} \\ &= \frac{81.433}{\sqrt{55.525 \times 222.0344}} \\ &= \frac{81.433}{111.0336} \\ &= 0.7334. \end{aligned}$$

Quantifying the relationship: Correlation

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

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

For information, the value of r for the plots in Figure 1.2, from top-left and moving clockwise, is: $r = 1, -0.899, 0.699$ and 0.064 .

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

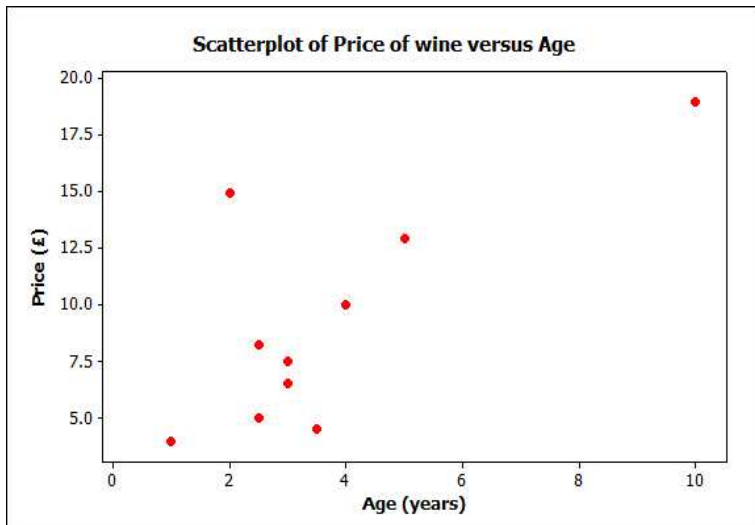
Modelling the relationship: simple linear regression

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.

Modelling the relationship: simple linear regression



Modelling the relationship: simple linear regression

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

The problem is, everyone’s line of best fit is bound to be slightly different, and so everyone’s predictions will be slightly different!

The aim of regression analysis is to find the very best line which goes through the data in a completely objective way.

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

Modelling the relationship: simple linear regression

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**.

Modelling the relationship: simple linear regression

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.

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.

Modelling the relationship: simple linear regression

For the wine data, we have

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \\ &= \frac{81.433}{55.525} \\ &= 1.467 \quad \text{and}\end{aligned}$$

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 9.258 - 1.467 \times 3.65 \\ &= 3.903.\end{aligned}$$

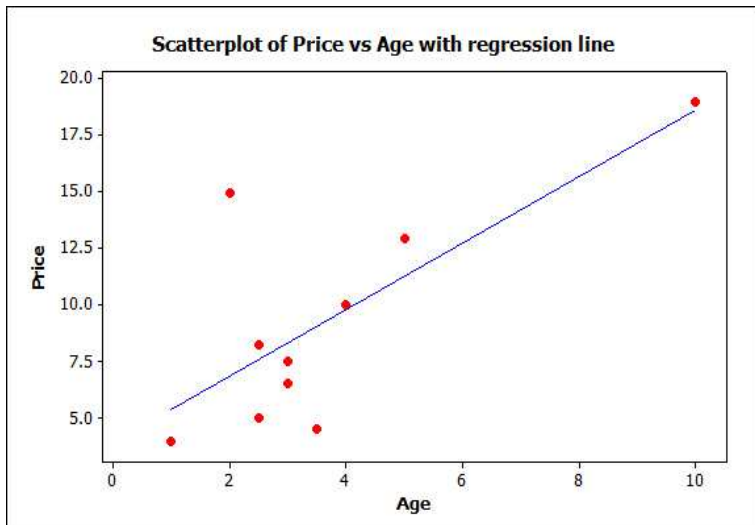
Modelling the relationship: simple linear regression

Thus, the regression equation is

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

The plot in Figure 1.3 shows the scatter diagram for the wine data again, but now with the regression line superimposed.

Modelling the relationship: simple linear regression



Modelling the relationship: simple linear regression

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

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

Based on 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.

Modelling the relationship: simple linear regression

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.

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.

Clicking OK gives the regression output shown in your notes; let's do this now in Minitab.

Minitab

Using 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:

Minitab

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.

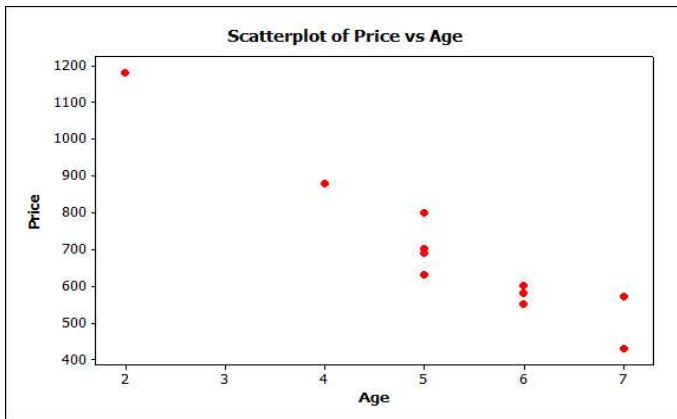
Testing the strength of a relationship

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.

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

Testing the strength of a relationship



The sample correlation coefficient between age and price is $r = -0.957$.

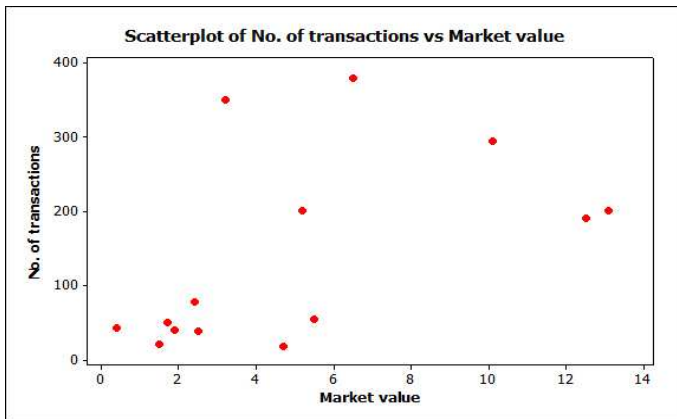
Testing the strength of a relationship

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.

Market value	6.5	5.2	0.4	1.7	1.9	2.4	3.2
Transactions	380	200	42	50	40	78	350
Market value	4.7	10.1	12.5	13.1	5.5	2.5	1.5
Transactions	18	295	190	200	55	38	20

Testing the strength of a relationship



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

Testing the strength of a relationship

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
- One 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 $r = 0.515$).

Testing the strength of a relationship

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;
- 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”?

Testing the strength of a relationship

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 .

Testing the strength of a relationship

Just like the sample mean ($\bar{x}$) is an estimator of the **population** mean (usually denoted μ), the sample correlation coefficient (r) is an estimator of the **population** correlation coefficient (usually denoted ρ).

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.

Revision: hypothesis testing

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

Revision: hypothesis testing

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 .

Revision: hypothesis testing

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

Testing the strength of a relationship

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:

$$H_0 : \rho = 0 \quad \text{versus}$$

$$H_1 : \rho \neq 0$$

Testing the strength of a relationship

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:

Minitab

Testing the strength of a relationship

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.

Since $p = 0.000$, which is less than 0.01,

- we have **strong** evidence against H_0 ;
- therefore we **reject** H_0 and go with H_1 ;
- H_1 says $\rho \neq 0$, i.e. there *is* a significant association between the age of a car and its sale price.

Testing the strength of a relationship

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:

Minitab

Testing the strength of a relationship

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 ;
- this is not enough to reject it, so we **retain** H_0 ;
- there is insufficient evidence to suggest a significant association between market value and number of transactions

Testing the strength of a relationship

Look at the Minitab output on **page 13** for the correlation between age and price for the wine data; use the p -value to check the significance of the association here.

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

Since the p -value is 0.016 (or 1.6%), and so lies between 1% and 5%:

- We have **moderate** evidence against H_0 ;
- We therefore **reject** H_0 in favour of H_1 ;
- There is evidence in our sample to suggest a significant relationship between age and price of wine.

Testing the significance of the slope

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 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 and we can conclude that the value of X does not influence the value of Y .

Testing the significance of the slope

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 .

Testing the significance of the slope

The null and alternative hypotheses are now:

$$H_0 : \beta_1 = 0 \quad \text{versus}$$

$$H_1 : \beta_1 \neq 0$$

- If we **retain** H_0 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 then we would conclude that the slope term *is* important in our model, and so X *is* a significant predictor of Y .

Testing the significance of the slope

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

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

Page 13 of these lecture notes gives the regression output from Minitab for the wine data, which we will now obtain again.

Minitab

Testing the significance of the slope

Minitab tells us that the estimated slope term using the data in our sample is $\hat{\beta}_1 = 1.4666$.

- 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**)
- The test statistic is just the estimated coefficient divided by its standard error ($1.4666/0.4805$)...
- ... which gives $t = 3.05$

Testing the significance of the slope

This has a p -value of 0.016, or 1.6%. Thus,

- we have **moderate** evidence against H_0 ;
- we **reject** H_0 in favour of H_1 ;
- the slope of our regression line **is** significant, and so age *is* an important predictor of price.

What about the rest of the output?

The Minitab output gives $S=3.58129$.

Recall that the linear regression model here is

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

- The assumption is that $\epsilon \sim N(0, \sigma^2)$ (see diagram on board)
- Minitab has estimated σ to be $\hat{\sigma} = 3.58129$
- Thus, $\epsilon \sim N(0, 12.826)$
- This just gives us an idea of the variability of our data points about the regression line. The bigger the value of $\hat{\sigma}$, the more 'scatter' we have!

What about the rest of the output?

The Minitab output also gives $R\text{-Sq}=53.8\%$.

R^2 measures the percentage of variability in the Y data that is explained by X .

- If all our data lie on a straight line, X tells us everything about Y , with no deviations from the line, and so $R^2 = 100\%$
- The closer R^2 is to 100%, the better!
- 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

Words of warning

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.

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 .

Multiple linear regression

Examples

- Sales of a product may depend on:
 1. product's unit price, *and*
 2. amount of advertising expenditure, *and*
 3. the price of a competing product
- Number of fatal accidents may depend on:
 1. number of registered vehicles on the road, *and*
 2. the price of petrol

The first example has *three* predictor variables, the second has *two*.

Example: Back to the price of wine

In the wine example, we used **simple linear regression** to investigate the capability of age in predicting the price of wine.

Surely there are other things that can influence the price of a bottle of wine?

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)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rain (X_2)	126	121	125	106	107	112	124	105	116	108
Temp (X_3)	16	20	17	18	18	22	19	15	21	20

Example: Back to the price of wine

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 the 'random error' term, and $\epsilon \sim N(0, \sigma^2)$
- The β 's are parameters that need to be estimated
- But now we have *four* β 's:
 - β_0 can be thought of as the intercept term as before
 - β_1 is the 'age coefficient'
 - β_2 is the 'rainfall coefficient'
 - β_3 is the 'temperature coefficient'

Example: Back to the price of wine

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 for the simple linear regression model, but this requires knowledge of matrix algebra which many of you won't have.

Anyway, Minitab can perform the calculations for us, and I will demonstrate this now.

Minitab

Example: Back to the price of wine

Thus, the full (multiple) regression model is:

$$Y = \underbrace{-22.5}_{\text{constant}} + \underbrace{0.807X_1}_{\text{Age}} - \underbrace{0.0004X_2}_{\text{Rainfall}} + \underbrace{1.55X_3}_{\text{Temperature}} + \underbrace{\epsilon}_{\text{random error}},$$

where

- $\epsilon \sim N(0, 1.80760^2)$
- X_1 represents the age of a bottle of wine
- X_2 represents the total rainfall during the growing season
- X_3 represents average afternoon temperature.

Example: Back to the price of wine

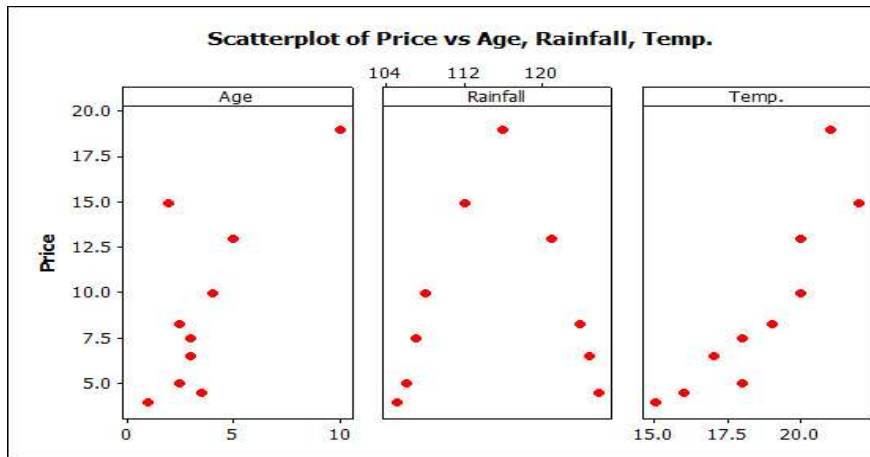
The estimated coefficients of the model indicate the direction of the relationship between the price of a bottle of wine and each of the corresponding predictors.

For example:

- $\hat{\beta}_1 = 0.807$ is positive: this indicates a positive relationship between age and price;
- $\hat{\beta}_2 = -0.0004$ is negative: this indicates a negative relationship between rainfall and price;
- $\hat{\beta}_3 = 1.55$ is positive: this indicates a positive relationship between temperature and price.

However, producing simple scatter plots of each predictor variable (age, rainfall and temperature) against the response variable (price) can help to inform our model.

Example: Back to the price of wine



Example: Back to the price of wine

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.

Example: Back to the price of wine

In fact, what we see is a **non-monotone** relationship, and possibly a **non-linear** relationship, which both increases with rainfall *and* decreases.

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!

Testing the importance of our predictor variables

Recall that our multiple linear regression equation is

$$Y = \underbrace{-22.5}_{\text{constant}} + \underbrace{0.807X_1}_{\text{Age}} - \underbrace{0.0004X_2}_{\text{Rainfall}} + \underbrace{1.55X_3}_{\text{Temperature}} + \underbrace{\epsilon}_{\text{random error}} .$$

However, do we really need all three predictor variables in the model?

Maybe just two – or one of them – would do just as good a job at predicting the price of a bottle of wine.

The simpler the model the better!

Testing the importance of our predictor variables

Testing the importance of Age as a predictor

Age is variable X_1 , which has coefficient β_1 . Our hypotheses are:

$$H_0 : \beta_1 = 0 \quad \text{versus}$$

$$H_1 : \beta_1 \neq 0.$$

The p -value for this, as given in the Minitab output, is 0.030 (or 3%). Since this lies between 0.01 and 0.05 (1% and 5%),

- we have **moderate** evidence against H_0 ;
- 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**.

Testing the importance of our predictor variables

Testing the importance of Rainfall as a predictor

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

$$H_0 : \beta_2 = 0 \quad \text{versus}$$

$$H_1 : \beta_2 \neq 0.$$

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 ;
- we **retain** $H_0: \beta_2 = 0$;
- **rainfall is NOT important in our model.**

Testing the importance of our predictor variables

Testing the importance of Temperature as a predictor

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

$$H_0 : \beta_3 = 0 \quad \text{versus}$$

$$H_1 : \beta_3 \neq 0.$$

The temperature coefficient β_3 has a p -value of 0.002 (or 0.2%). Since this is less than 1%,

- we have **strong evidence against** H_0 ;
- 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**.

Testing the importance of our predictor variables

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

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

Minitab

Testing the importance of our predictor variables

Notice that the regression equation has 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 also 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.

Testing the importance of our predictor variables

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.

We could now use this model to make predictions.

Making predictions

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?

Making predictions

The regression equation of the final model is

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

Substituting $X_1 = 7$ and $X_3 = 18.5$ into this equation gives:

$$\begin{aligned} Y &= -22.6 + 0.806 \times 7 + 1.55 \times 18.5 \\ &= 11.717, \end{aligned}$$

so we could sell this wine for about £11.72 per bottle. Notice that we didn't use the rainfall figure of 117mm in our calculation as this was found not to be an important predictor (and so was dropped from the model).

An overall test of the model

As well as the regression equation, the p -values for each predictor variable and the R^2 statistics, Minitab also gives output associated with the *overall fit of the model*.

This is often referred to as the **F-test**; the hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \dots = 0 \quad \text{versus}$$

$$H_1 : \text{at least one of the parameters is not zero}$$

In the final model for the wine example, we were left with β_1 and β_3 in our model, and so we have

$$H_0 : \beta_1 = \beta_3 = 0 \quad \text{versus}$$

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

An overall test of the model

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.

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!

An overall test of the model

Here, we see the p -value is very small (0.000 to three d.p.!) and so

- we have **strong** evidence against H_0 ;
- therefore we **reject** it and go with H_1 ;
- some, or *all*, of the predictor variables used in the fit are useful in predicting the price of a bottle of wine!

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).

Minitab

Checking the model

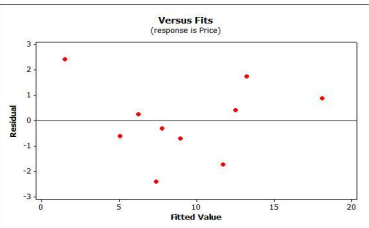
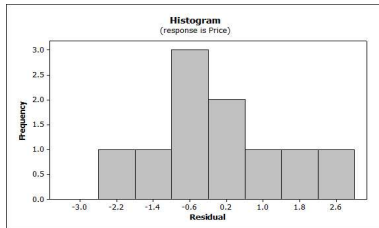
We have used

- t -tests to check the importance of each predictor variable
- The F -test to check the *overall* fit of the model

These tests both rely on $\epsilon \sim N(0, \sigma^2)$. We need to check this!

Minitab

Checking the model



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).

Indicator variables

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

Indicator variables

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.

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.

Example: Back to the price of wine

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)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rain (X_2)	126	121	125	106	107	112	124	105	116	108
Temp (X_3)	16	20	17	18	18	22	19	15	21	20
NZ? (X_4)	0	1	0	0	0	1	0	0	1	0

Example: Back to the price of wine

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?

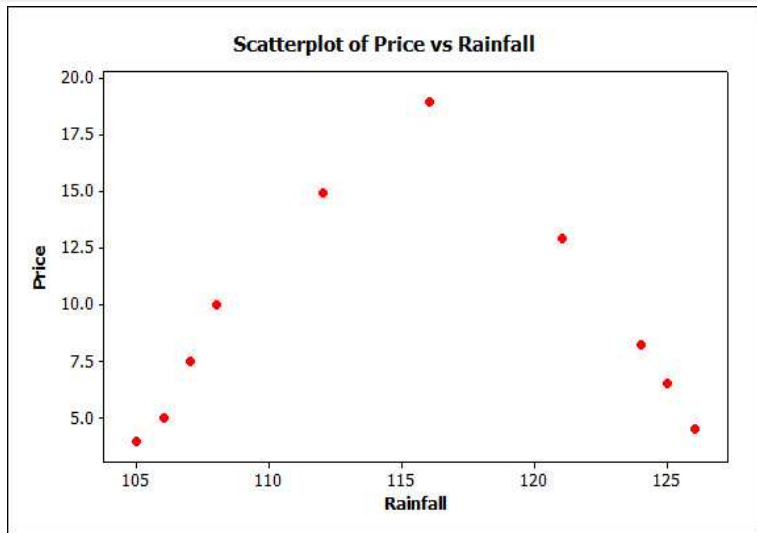
Minitab

Non-linear regression

In all of the regressions performed thus far on the wine sales dataset, the **rainfall** variable has always been excluded from the model.

However, as Figure 1.11 shows, there is clearly a relationship between rainfall and price.

Non-linear regression



Non-linear regression

Suppose we are interested in *only* the relationship between rainfall and price.

How can we proceed?

- A simple linear regression is not appropriate: we have both an increasing and decreasing relationship
- There appears to be a curved relationship to the left, and perhaps the right, of 117mm.

Quadratic graphs

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

1. $y = x^2$

2. $y = -x^2$

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

Non-linear regression

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** we also need rainfall² (X^2) in another column (say column C3).

Minitab

Non-linear regression

So our regression equation is

$$Y = -1658 + 28.97X - 0.1252X^2$$

Also notice that both Rainfall and Rainfall² are important predictor variables in the model, since both β_1 and β_2 have small p -values.

So we have found a regression model which caters for the non-standard relationship between price and rainfall!

Non-linear regression

We can also use Minitab to produce a scatterplot with the quadratic regression equation superimposed.

Minitab

Non-linear regression

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

The regression equation is

$$Y = -1658 + 28.97X - 0.1252X^2.$$

Substituting $X = 125$ into this gives:

$$\begin{aligned} Y &= -1658 + 28.97 \times 125 - 0.1252 \times 125^2 \\ &= 7, \end{aligned}$$

i.e. £7.

Logistic regression

We now return to the motivating example on page 6 of these notes.

Motivating example: Product placement



Motivating example: Product placement



Motivating example: Product placement



Motivating example: Product placement



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.

Logistic regression

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, 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*.

Logistic regression

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

Logistic regression

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 ?

Minitab

The logistic regression equation

In simple linear regression, we know 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

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

The logistic regression equation

In **logistic regression**, statistical theory, as well as practice, has shown that the relationship between $E[Y]$ and X is better 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, $E[Y]$ provides a **probability** that $Y = 1$ given a particular value for the predictor variable X .

The logistic regression equation

Because of the interpretation of $E[Y]$ as a probability, the **logistic regression equation** is often written as:

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

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.

Minitab

The logistic regression equation

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

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

The logistic regression equation

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

$$\begin{aligned}\Pr(Y = 1|X) &= \frac{e^{-2.99147+0.0445355 \times 11}}{1 + e^{-2.99147+0.0445355 \times 11}} \\ &= 0.0757.\end{aligned}$$

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. We have:

$$H_0 : \beta_1 = 0 \quad \text{versus}$$

$$H_1 : \beta_1 \neq 0.$$

From the Minitab output, the p -value associated with the Salary predictor variable is 0.010, or 1%.

- We have **moderate** evidence against H_0
- We **reject** H_0 in favour of H_1
- There is evidence to suggest that the time at which a *Mars Bar* is placed in a film **is** an important predictor of whether or not a cinema-goer will be able to recall the brand.