9 Predicting the Unpredictable or Explaining the Inexplicable: Regression Models

This chapter shows how you can use some data on one or more variables (for example people's diet and exercise habits) to try to predict the value of a further variable (for example on their general health). Models to do this are called 'regression models'. We start with the single variable case and then move on to 'multiple regression models' which use several variables to make the prediction. We also look at how we can assess the accuracy of regression models, the circumstances under which they may be misleading and how they can be used; often the aim is to try to *explain* something rather than to predict it.

▶ 9.1 Introduction: predicting earnings on the Isle of Fastmoney

The data in Table 3.1 shows that the male students drank an average of 13.4 units of alcohol on the Saturday, whereas the females only drank 3.3 units. This means that if we wanted to predict how much a randomly chosen student would drink next Saturday, a reasonable guess might be 13.4 units if the student was male and 3.3 if female. However, we would not expect this guess to be at all accurate. If we had more information, like the student's age and other variables, we should be able to make a much more accurate prediction. But we still wouldn't expect to be 100% accurate.

This is the subject of this chapter: making predictions like this as accurately as possible. I'll illustrate the methods we can use drawing examples from an imaginary island known as the Isle of Fastmoney. This island used to have another name, but ever since a couple of students at the college on the island did a small survey and found that high earners on the island could run substantially faster than those earning less, the older name has been forgotten in favour of the name coined by the local paper, the Isle of Fastmoney. Every year on the island there is a 10 km race. Most of the population enter, and all finishers are given a medal with their name and time engraved on it.

It's taken very seriously and the times are generally pretty fast. The survey found, much to everyone's surprise, that there was a negative correlation (Section 3.7.3) between the time individuals recorded in the race and their annual earnings, especially if allowances were made for age and sex. There was a definite tendency for the higher earners to record faster times in the race. Other surveys since have found a similar pattern.

In the last couple of years, pressure from a cycle manufacturer on the island has resulted in a cycle race being added to the island's calendar. The results show a very similar relationship with earnings: the high earners seem to be able to cycle faster. The most recent survey was based on a sample of 50 male islanders and 50 female islanders. These were drawn at random from the population of the island aged between 20 and 60 who entered both the running and the cycle race (about 10 000 people). This is a 'stratified' sample because we draw the males and the females separately and ensure that we have the right proportion of each (50%) in the final sample.[128]

The data in Table 9.1 shows eight males and eight females drawn, randomly again, from this sample. (Table 9.1 is on the web as *iofm16.xls*, and the whole sample as *iofm.xls*.) How can we predict earnings from the race times and the other variables in Table 9.1? And can these predictions help us

Table 9.1 Subsample of data from the Isle of Fastmoney

Refno	*Sex*	*Sexn*	*Age*	*Height in cm*	*Run 10 km*	*Cycle 10 km*	*Earn000e*
1	F	1	42	170	61	24	106
2	F	1	53	165	87	35	17
3	F	1	27	161	71	28	18
4	F	1	53	170	83	34	11
5	F	1	23	161	56	20	15
6	F	1	53	168	69	29	40
7	F	1	32	162	67	28	25
8	F	1	51	167	66	29	156
9	M	0	46	174	70	27	37
10	M	0	49	187	53	23	143
11	M	0	33	180	64	28	12
12	M	0	48	182	77	33	20
13	M	0	49	177	58	21	79
14	M	0	24	189	48	17	6
15	M	0	31	176	50	20	59
16	M	0	35	190	52	21	66

Sexn is a numerical coding of sex: 1 for female and 0 for male. The Run 10 km and Cycle 10 km columns are the race times to the nearest minute. Earn000E is earnings in thousands of euros. Data is in *iofm16.xls*.

understand the factors which determine earnings on the island? These questions are the subject of this chapter.

9.2 Straight line prediction models: linear regression

The simplest form of model is based on a single category variable. Does Table 9.1 suggest that we can predict anything about islanders' earnings from their sex (M or F in Table 9.1)**?**

Not much. The average earnings of the females in Table 9.1 are 48.5 thousand euros, and the corresponding figure for males is 52.75. But this is a slight difference, which would provide a very unreliable prediction.

The next simplest possibility is to base a prediction of earnings on a single number variable, such as the time in the 10 km run. Figure 9.1 shows the relationship between these two variables as a scatter diagram (Section 3.7.1). This figure seems to show the expected correlation between the two variables: most of the high earners seem to have fairly fast (low) race times. But this relationship does not look strong enough to make a useful prediction.

This is not the case with relationship between the run time and the cycle time in Figure 9.2. This figure suggests that it would be possible to get a reasonable estimate of the time an individual took for the cycle race from their time in the run. For example, what would your predictions be for two runners: one with a run time of 60 minutes, and the other with a time of 80 minutes**?**

Figure 9.1 Relationship between 10 km run time and earnings (based on sample of 16)

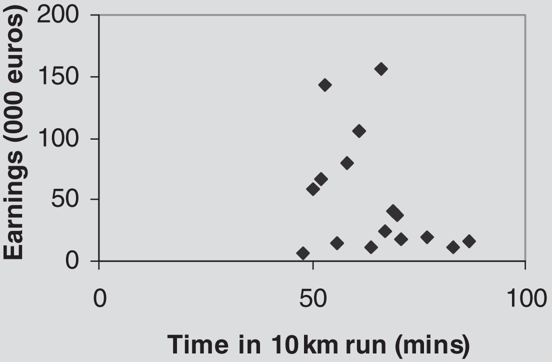
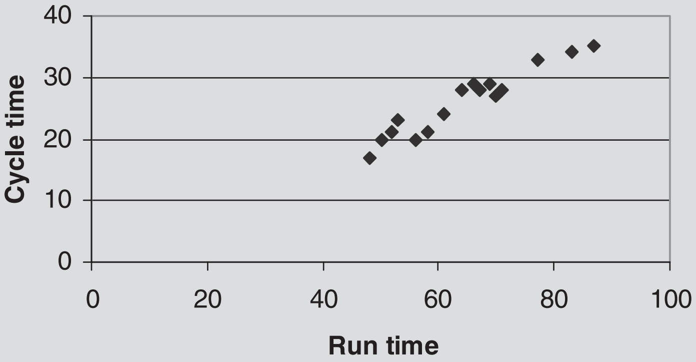


Figure 9.2 Relationship between 10 km run time and 10 km cycle time (based on sample of 16)



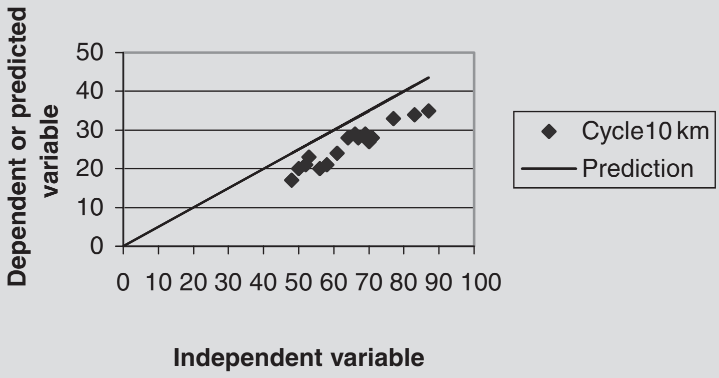
Runners who finished in about 60 minutes took about 20–25 minutes in the cycle race. Similarly, the graph suggests that a time of 80 minutes for the run is likely to imply a time of 30–40 minutes for the cycle ride. The slightly scattered pattern emphasises the obvious point that we should not expect these predictions to be exact.

A prediction can be represented on the graph as a line. Figure 9.3 shows a prediction line superimposed on Figure 9.2. It's obviously *not* a good prediction line, but it will do to explain how prediction lines work, and how we can decide how good they are. The vertical axis (cycle time) is the variable we are trying to predict, so we have to assume it is *dependent* (in some sense) on the other variable, the run time. This is known as the 'independent variable' because we assume we have a way of measuring this independently of the other variable.

The prediction line in Figure 9.3 makes a very simple prediction: that the cycle time will be half the run time. For example, the prediction for a runner who takes 60 minutes is that they take 30 minutes for the cycle race. The line consists of all the points for which this is true. The half (0.5) is called the 'slope' (gradient) because it corresponds to the steepness of the line. In this case the line goes up half a unit for every one unit along the horizontal axis. If the slope was 2.5, it would go up 2.5 units for every one unit along, and it would look steeper. (The important terminology is summarised in Table 9.7 below.) What is the prediction from the line in Figure 9.3 for a runner who takes 80 minutes**?**

Using this line, a run time of 80 minutes corresponds to a cycle time of 40 minutes. Alternatively, using the rule on which the line is based, cycle time

Figure 9.3 One line for predicting 10 km cycle time from 10 km run time (based on sample of 16)



Intercept = 0; Slope = 0.5; MSE = 40.8.

is half the run time. How good are the predictions made by the line in Figure 9.3**?**

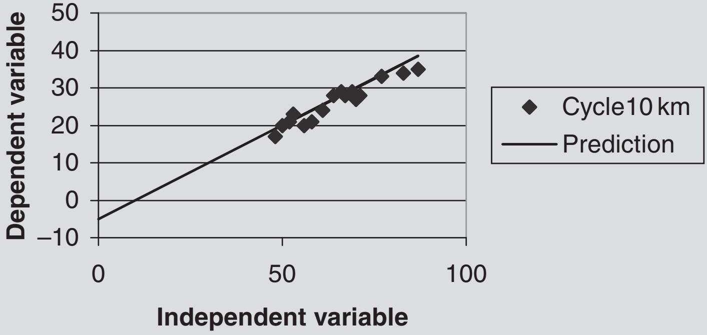
They all look too high. The line is consistently above the points representing the actual race times. To compensate for this, we need to lower the line. We can do this by changing the starting point.

The prediction line in Figure 9.3 starts from zero: it crosses the vertical axis at the point where the dependent variable is zero. The line in Figure 9.4 starts lower, at the point where the dependent variable is –5. This is known as the 'intercept' because it intersects the vertical axis at this point. It is the value of the dependent variable where the independent variable is zero. What is the intercept of the prediction line in Figure 9.3**?**

The intercept is 0. The line in Figure 9.4 looks a better predictor than the one in Figure 9.3. What we need is a way of measuring how accurate a prediction line is likely to be. Then we should be in a position to find the most accurate possible line.

I'll explain the way this is done in relation to Figure 9.3. The basic idea is to work out the error we would make in using the prediction line to predict each of the cycle times, and then work out a sort of average error. Some of the results are in Table 9.2. The Excel workbook *pred1var.xls* uses this method to work out prediction lines.[129] The last row in Table 9.2 (no. 14) corresponds to the point on the left of the cluster of points in Figure 9.3. The run time is 48 minutes, so the predicted cycle time is half this, 24 minutes. The actual cycle time is 17 minutes, so the error in the prediction is 24 minus

Figure 9.4 A better line for predicting 10 km cycle time from 10 km run time (based on sample of 16)



Intercept = –5; Slope = 0.5; MSE = 3.9.

Table 9.2 Calculation of error and square error for Figure 9.3

Refno	*Run 10 km*	*Cycle 10 km*	*Prediction*	*Error*	*Square error*
1	61	24	30.5	6.5	42.25
8	66	29	33	4	16
14	48	17	24	7	49

Only three of 16 rows shown. Slope = 0.5; intercept = 0.

17 or 7 minutes. This 7 minutes corresponds to the vertical gap between the point and line in Figure 9.3. The smaller these errors are, the closer the line will fit the points and the more accurate any prediction is likely to be.

To assess the overall performance of the line, we need some measure of average error. The obvious way to do this is perhaps to use an ordinary mean. However, this leads to a few problems (see Exercise 9.9.4); the preferred measure is the 'mean square error' (MSE). To work this out you square the errors and then take the mean of these squares. For the last row in Table 9.2, the square error is 49, and the mean of the three square errors in the table is 35.75. If we include all 16 people in Table 9.1, the MSE comes to 40.8.

Figure 9.4 is based on an intercept of –5 (instead of 0). What would the three errors and square errors in Table 9.2 be with this intercept**?**

The intercept is now –5, so to work out the predictions you need to add this on. The answers are in Table 9.3. Note here that the middle error is

Table 9.3 Calculation of error and square error for Figure 9.4

Refno	*Run 10 km*	*Cycle 10 km*	*Prediction*	*Error*	*Square error*
1	61	24	25.5	1.5	2.25
8	66	29	28	–1	1
14	48	17	19	2	4

Only three of 16 rows shown. Slope = 0.5; intercept = –5.

negative because the prediction is too small (the point lies beneath the line), but the square of a negative number is positive (see Note 18). The mean of the three square errors in Table 9.3 is 2.42. The overall MSE from all 16 people is 3.9. The fact that the MSE for Figure 9.4 is much less than for Figure 9.3 reflects the fact that the prediction line in Figure 9.4 is much more accurate.

Can we do better? Can we find a line with an even lower MSE? To see if this is possible, we can use the Excel Solver,[130] which will find the values of the slope and the intercept which lead to the lowest possible MSE. In this case, Solver tells us that the lowest possible MSE is 2.2, which is achieved by setting the slope to 0.45 and the intercept to –3.0. A value of 2.2 for the MSE is better than 3.9, although the graph (Figure 9.5) does look very similar.

The prediction line in Figure 9.5 is called a 'least squares' prediction line, or a 'best fit' line. It's also known as a 'regression line' for reasons that are less obvious.[131] What prediction does this regression line give for the cycle time for someone who takes 50 minutes in the 10 km run**?**

It's clear from Figure 9.5 that the answer is about 20 minutes. To get the exact prediction from the line, we need to add the intercept (–3.0) to the slope (0.45) multiplied by the run time:

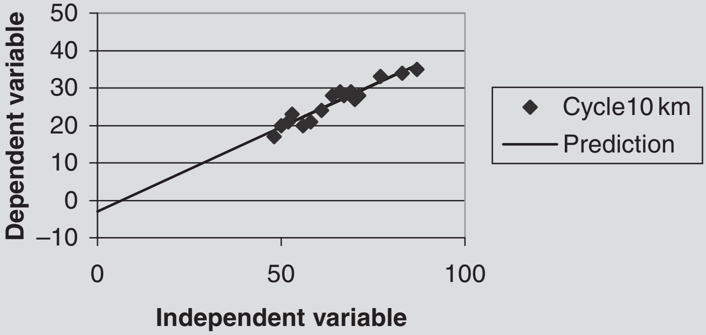
Prediction = –3.0 + 0.45 × 50 = 19.5 minutes.

The intercept in Figure 9.5 is the cycle time for a person who can do the run in zero time. This comes to minus 3 minutes. Does this make sense**?**

This intercept does not much make sense as a prediction. The regression line only makes sensible predictions towards the right of Figure 9.5 as this is where the data is.

The slope determines the amount of extra time we need to add on to take account of the run time. The slope of 0.45 means that if you have two people whose run times differ by one minute, their cycle times are likely to differ by 0.45 minutes.

Figure 9.5 Best fit line for predicting 10 km cycle time from 10 km run time (based on sample of 16)



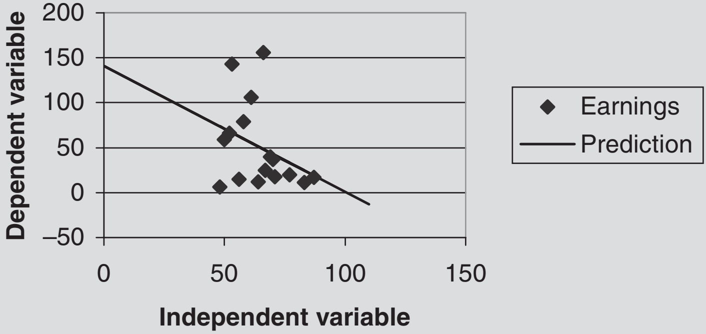
Intercept = –3; Slope = 0.45; MSE = 2.2; R squared = 92%.

We'll now turn to the original prediction problem: predicting earnings from the 10 km run time (Figure 9.1). Exactly the same method (using the spreadsheet *pred1var.xls*) leads to the prediction line in Figure 9.6. The slope here is negative (–1.4): the line goes downhill, indicating that people with bigger (slower) race times earn less. The value of the MSE for Figure 9.6 is a lot more, 1893. How useful do you think this prediction line is? What problems can you see**?**

There are lots of difficulties: this is a good illustration of many of the problems with regression, which we will look at later. Compared with Figure 9.5, the scatter in the diagram suggests that the prediction is likely to be very inaccurate. The ends of the line seem particularly unrealistic: it is unlikely that people with run times longer than 100 minutes will earn negative amounts, for example. Earnings are likely to depend on the other variables too (for example age), so a prediction which ignores these is unlikely to be very accurate. And the sample of 16 is unlikely to be sufficient to come up with reliable conclusions. We'll see how we can do better in Section 9.4.

We've seen how to use the Excel Solver to work out regression lines. It is also possible to work out the slope and intercept by means of mathematical formulae which are built into Excel and SPSS. They are the basis of the Excel functions forecast (which gives the predictions), slope and intercept.[132] There is also a Regression tool (see Section 9.5). You should find these functions give identical answers to the Solver method, and they are certainly easier to use. The main advantage of the Solver method is that it makes it clear what is going on. It is also more flexible: you can adjust the method to try rather different types of model (for example Exercise 9.9.4).

Figure 9.6 Best fit line for predicting earnings from 10 km run time (based on sample of 16)



Slope = –1.4; MSE = 1893; R squared = 11.5%.

▶ 9.3 Assessing the accuracy of prediction models

9.3.1 A measure of error reduction: R squared

The errors – the difference between the prediction line and the actual values – look a lot smaller in Figure 9.5 than in Figure 9.6. This suggests that a prediction based on Figure 9.5 is likely to be much more accurate than one based on Figure 9.6. The measure we used – MSE – seems to confirm this: the MSE is 2.2 for Figure 9.5 and 1893 for Figure 9.6. However, this comparison is not entirely fair. The values of the dependent variable in Figure 9.6 (earnings) are larger and more spread out than those in Figure 9.5 (10 km cycle time), so any errors are likely to be larger for this reason. There is a simple way round this. This is to think in terms of the *proportional reduction in mean square error* which the prediction produces. To do this, we need to start with the error before we have the prediction line. If you were faced with the task of predicting an individual's 10 km cycle time from the data in the cycle 10 km column of Table 9.1, *without* using any of the other variables to help you, what would your best guess be**?**

The obvious answer is the mean. (The median would be another possibility – see Section 3.4.2 – but I'll use the mean as this is the conventional approach.) The mean time for the cycle race comes to 26 minutes. What would the square error be if we were using this 26 minutes as our prediction for person 14 (Table 9.1)**?**

The actual time was 17 minutes, so the error is 9 minutes, and the square error is 81 minutes. This is a lot more than the 2 in Table 9.3, and the square error from the best fit line in Figure 9.5, which comes to 2.7. (Overall, of course, the best fit line is better than the model in Table 9.3, but it is not better for this particular point.) The mean square error from taking this prediction of 26 for all points is 27.6. This quantity is called the 'variance' of the cycle times and is one of the standard statistics built into software.[133] It is a measure of how spread out the cycle times are. The method of working it out is the same as for the standard deviation (Section 3.4.4) except that you don't take the square root in the last step. The variance is the square of the standard deviation.

So the MSE without the prediction line (the variance) is 27.6, but with the prediction line in Figure 9.5 this is reduced to 2.2. What is the proportional reduction in error**?**

The reduction in error is 25.4 (27.6 – 2.2) which is 25.4/27.6 or 92% of the original error. This prediction line reduces the MSE by 92%. It is likely to be a good prediction. How would expect Figure 9.6 to compare**?**

The variance of the earnings is 2139, and the MSE from the prediction line is 1893. The prediction obviously hasn't made much difference: the proportional reduction (R squared) is 0.115 or 11.5% {(2139 – 1893)/2139}.

The proportional reduction in mean square error from a regression line is also known as 'R squared'. This is because it turns out to be the square of the Pearson correlation coefficient (Section 3.10), and a standard symbol for this is *r*. This should seem reasonable because the closer the correlation coefficient is to zero, the lower R squared will be, the more scattered the scatter diagram (Section 3.7), and the less accurate any prediction line is likely to be.

Go back to the three scatter diagrams in Figure 3.7. For each of the three diagrams, what would the regression line look like, and what would R squared be**?**

For the first two diagrams (with a correlation of +1 and –1), the prediction line would follow the points exactly, and R squared would be +1 or 100%, indicating that the error has been reduced by 100%. The predictions are completely accurate. (Remember that the square of –1 is +1.) In the third case, the variable on the horizontal axis is obviously of no help in predicting the other variable. This means that the prediction line would be a horizontal line at the level of the mean of the dependent variable. As the correlation is zero, R squared will also be zero, which confirms that the prediction does not reduce the error at all. R squared gives a good indication of the likely value of a regression line. There are, however, snags; we'll come to some of the problems in later sections.

9.3.2 Confidence intervals

In the last section we saw how R squared can be used to measure how useful a regression line is. A slightly more direct question is to ask how big the error is likely to be. The first approach to this is to use the MSE. We saw that this is a bit like the variance, and its square root is a bit like the standard deviation. The MSE for the prediction of earnings in Figure 9.6 is 1893, so the 'standard deviation' is 44 (thousand euros, of course). This is the standard deviation of the errors, often called the 'standard error' of the prediction. If the error distribution is normal (Section 5.6), we can use this to estimate a confidence interval for the error:[134] the 95% interval will extend from −2 sds to +2 sds, or −88 000 euros to +88 000 euros. For example, the prediction from Figure 9.6 for someone who can do the run in 90 minutes is that they will earn about 15 000 euros. The confidence interval derived from this is that we can be 95% sure that this runner will earn something between 103 000 euros and losing 73 000 euros (since 15 − 88 is negative). This is not an impressive prediction, but the pattern in Figure 9.6 should not encourage you to expect a good prediction.

What is the equivalent confidence interval for the error from Figure 9.5 (with an MSE of 2.2)**?**

The standard error comes to 1.5 so the 95% confidence interval extends from 3 minutes below the prediction line to 3 minutes above it. This is a much more acceptable error for a useful prediction.

There are a few difficulties with this method. It fails to take account of the fact that the regression line itself may be inaccurate because the sample is small. It depends on the errors being normally distributed (see Section 5.6). And it's based on an average error; it ignores the fact that the predictions are likely to be more accurate for some values than for others.

The next approach, bootstrapping, gives us a way of largely avoiding the first two problems (but not the third). I'll start with a slightly different question: how accurate is the estimate of the slope? The slope is an important piece of information in its own right. The slope of −1.4 means that running a minute quicker corresponds to, on average, an extra income of 1.4 thousand euros. But how accurate is the estimate of the slope based on such a small sample?

To find out, we can use the bootstrap method (Chapter 7). The method is just like the method we used to estimate the confidence interval for a correlation (Section 8.4), except that here we are analysing the slope. The principle is to resample from the sample of 16 with a view to seeing how much the slope estimated from samples of 16 will vary between samples. (You will probably need to read Chapter 7, and Section 8.4 to follow this fully.)

When I used `resample.xls` to work out the 95% confidence interval for the slope of Figure 9.6,[135] it came to −3.1 to +0.4. This indicates that the slope is very uncertain: there is a reasonable chance it may even be positive. This confidence interval is based on simulation with only 200 resamples; if you do it on your computer the answer may be slightly different. It is also possible to get confidence intervals for the slope from the regression procedures in Excel and SPSS (Section 9.5).

Making a prediction from the regression line in Figure 9.6 is a bit of a waste of time, so I'll return to Figure 9.5 to illustrate how we can find a confidence interval for the prediction error. (Again, this builds on Chapter 7.) We need to draw resamples of 17 from the data in Table 9.1. The first 16 are used to set up the regression model, and the 17th is used to make a prediction and then find the error, that is, the difference between the actual value and the prediction. When I used `resample.xls`,[136] the first resample of 17 led to an error of −1.5, the next to an error of +1.3 and so on. The 95% confidence interval extended from −2.8 to +2.5 (based on results from 200 resamples.) As expected, this is similar to the −3 to +3 minutes we got above. The resampled answer should be more realistic because it avoids making assumptions about the errors being normally distributed and it takes account of sampling error, the likelihood of differences between one sample and another.

▶ 9.4 Prediction models with several independent variables: multiple regression

The prediction line for earnings in Figure 9.6 is based on only one variable, the 10 km run time. We are likely to be able to get a better prediction using more than one variable. Earnings, for example, seem to increase with age, so age should help to improve the prediction. The regression model can easily be extended to include any number of independent variables. We could include all five independent variables in Table 9.1. There are, however, good reasons for not doing this, which I will explain in due course. The variables I'll use for the first model, to try to get a better prediction of earnings, are Sexn, Age and Run 10 km.

It's easy to draw a line on a graph to represent the prediction from one variable (for example Figures 9.5 and 9.6). This is not possible with more than one variable, so the prediction can't be represented by a line on a graph. The phrase 'prediction line' no longer makes sense. Instead, we can refer to a prediction, or regression, 'model': this is simply the rule for making the prediction. (The prediction lines in Figures 9.5 and 9.6 are also models.)

The first column in Table 9.1 gives sex coded as F or M. For the regression model, we obviously need to code this numerically: the convenient coding scheme is to use 0 for one sex and 1 for the other. This gives the variable Sexn, which is known as a 'dummy' variable (because it's not a genuine numerical variable). This trick can always be used for category variables which can take one of two values. (The situation where there are three or more values, for example the three courses in Table 3.1, is a little more complicated.[137])

The procedure is now, more or less, identical to before. The interpretation of the answer is, however, a little more complicated, so I'll adjust the variables to make the conclusions more user-friendly. (This adjustment is not essential: you will get the same answers without it, but they may be harder to visualise.)

The intercept in Figure 9.5 is the predicted 10 km cycle time for someone who takes zero time in the 10 km run. This makes little sense! We could have improved matters by giving the independent variable, the run time, another base, say 50 minutes. The new variable would then have been *time above 50 minutes*. The first person in Table 9.1 with a time of 61 minutes would score 11 minutes on this new variable, and the 14th, with a time of 48 minutes, would score –2 minutes. If we had done this, the predicted times, and the slope, would have been unchanged. (If you don't believe this, try it.) The intercept would obviously have been different (19.5 minutes), but the value now has a reasonable interpretation: the 10 km cycle time corresponding to a 50 minute run time.

I'll take 50 minutes as the 'base value' for the new run time variable. A similar difficulty will occur with the age: I'll take 20 years as the base here. (It doesn't matter what base you take from the point of view of the final predictions: they will always be the same.)

The 'base case' is now someone who scores zero on all three independent variables, that is, a male (Sexn = 0) aged 20 with a 10 km run time of 50 minutes. There may or may not be such a person in the sample, but such a person could exist. You should be able to imagine him. The equivalent of the intercept for a multiple regression is the predicted value for this base case: the 'base prediction'.

Table 9.4 shows how a three variable prediction model works. Each of the three independent variables has its own slope: I have (arbitrarily) chosen 3 for the three slopes and the base prediction. Table 9.4 is based on `predmvar.xls`. I'll take the first row as an example. There are four components to the model. The first is the base prediction, that is, the prediction for our base case. We've arbitrarily set this to 3. The next component is due to sex. The first person is female, which means she is predicted an extra 3 units of income. We get this by multiplying the slope (3) by the value in the Sexn column (1). For a male, this component would be zero because males are coded as 0. Next we come to the age component. She is 42, so her age over 20 is 22 years, and the component due to age is 3 × 22 or 66, because we have assumed a slope of 3: income rises by 3 for each extra year of age. Similarly, the final component from her 10 km run time comes to 33 (3 × 11). And adding up all four components gives a final prediction of 105. This is very close to her actual earnings of 106, but this is just luck!

Notice that the underlying assumption is that we can start off from a prediction for the base case, and then add on separate components for each

Table 9.4 Calculation of three variable prediction model (three slopes and base prediction equal to 3)

					Prediction model				
		Data				*Additional component for:*			
Ref	*Sexn*	*Age 20+*	*Run 50+*	*Earnings*	*Base prediction*	*Sexn*	*Age 20+*	*Run 50+*	*Prediction*
1	1	22	11	106	3	3	66	33	105
8	1	31	16	156	3	3	93	48	147
14	0	4	–2	6	3	0	12	–6	9

Only three of 16 rows shown. Age 20+ means age over 20; Run 50+ means run time over 50 minutes. Overall MSE is 8981.

independent variable. Furthermore, we're assuming that each component can be derived by multiplying some number, the slope, by the value of the variable. Do you think this is realistic**?**

Probably not, but the question is whether it's good enough to be useful. We'll return to this question in Section 9.6. What is the mean square error of this model for the three cases in Table 9.4**?**

The square errors are 1, 81 and 9 with a mean of 30.3. However, these three cases fit the model much better than the rest: the overall MSE is 8981. The variance of the earnings is only 2139, so this model has managed to *increase* the error. Not surprisingly – in view of the fact that we set everything to 3 quite arbitrarily – this is a very bad model. To get it to fit better, we must use the Solver to find the three slopes and the base prediction which lead to the lowest possible MSE. The result of doing this is in Table 9.5. Whatever values of the three slopes and the base prediction you try, you will not manage to make the MSE lower than the value in Table 9.5 (677).

Note that the MSE (677) is now a lot less than it was for the model we started with (8981). It's also less than it was for the best fit model for predicting earnings from the run only (1893). The prediction fits better, as we expected. This is reflected in a bigger R squared (0.68 compared with 0.115 for the earlier model), which can be calculated in exactly the same way. We can use this model to make some predictions. What earnings would you predict for a 60-year-old female who can do the run in 52 minutes**?**

The prediction from the model is $13.9 + 33.2 + 4.0 \times 40 + (-4.3) \times 2 = 198$. What difference would it make if her run time was 48 minutes**?**

The only difference is that her time is now below the base of 50 minutes,

Table 9.5 Three variable regression model for earnings based on subsample of 16

Base prediction (all independents = 0)	13.9
Slope – sexn	33.2
Slope – age over 20	4.0
Slope – run time over 50 minutes	–4.3
MSE (mean square error)	677
RMSE (square root of MSE)	26
R squared	0.68
Number of cases (people)	16

so she scores –2 minutes, the component from the run becomes positive,[138] and the prediction becomes 215.

Which of the three independent variables – sex, age and run time – do you think has the biggest influence on earnings**?**

You might have said sex, on the grounds that this has the biggest slope in Table 9.5. Or you might think age because this has the biggest component (4.0 × 40 = 160) in the prediction above. Or you might think run time on the grounds that if she was a more typical 60-year-old who took, say, 90 minutes for the run, her run component would be –4.3 × 40, or –172. This would make the run component the biggest of the three. The point here is that the question is difficult to answer, because it depends on what you mean. Be careful and don't jump to unwarranted conclusions.

As well as using the model to make predictions, you can also use the slopes as an indicator of the way things work. Are the slopes in Table 9.5 what you would expect**?**

The age slope (4.0) indicates that, on average, the older inhabitants of the island earn more, at a rate of 4000 euros per year older. The negative slope for the run time indicates that the larger (slower) the race time, the less people earn, with one minute slower corresponding to a decrease of earnings of 4300 euros. This is in line with the original survey (see Section 9.1). The slope for sexn indicates females are likely to earn 33 200 euros more than similar males. Notice that this gives a slightly different picture from the single variable analyses in Section 9.2. The males in Table 9.1 earn slightly more, on average, than the females, so the slope we would expect for sexn is negative. Similarly, the negative slope for the run time is not so pronounced (–1.4 instead of –4.3). Why the differences**?**

The multiple regression slope for the run time (–4.3) takes account of the other variables. For example, there is a tendency for the fast runners to be younger and male. Being faster seems (on this island) to be associated with

Table 9.6 Five variable regression model for earnings based on subsample of 16

Slope – sexn	14.2
Slope – age	3.8
Slope – height	–1.4
Slope – run time	–8.3
Slope – cycle time	8.9
MSE	518
R square	0.76

Note that the slope for age is the same as the slope for age over 20; similarly for run time and run time over 50.

higher earnings, but being younger and being male are associated with lower earnings. The multiple regression tries to separate these influences. The single variable regression obviously can't do this.

So why not put all five variables in? If we do this, the results are Table 9.6. I have used the unadjusted variables for this regression: this make no difference to the slopes. The main oddity here is the slope for the cycle time (+8.9). This seems to indicate a strong tendency for the slower cyclists to earn more. As the fast cyclists tend to be the same people as the fast runners (Figure 9.2), this seems very odd indeed.

It is this strong correlation that is responsible for the problem. The run time and the cycle time both provide very similar information, so the regression can pick up this information from one, the other or some combination of both, depending on small quirks of the data. The model is unstable; the slopes may be quite different with another, very similar, sample. This is an important principle: it is important not to include strongly correlated independent variables in a multiple regression.

The other point to watch is that adding another variable is *always* likely to reduce the error (MSE) and increase the R squared (which has improved from 0.68 for the three variable model to 0.76). R squared can never go down if you add another variable, because Solver could always find the previous solution by setting the new slope to zero. This means that any variable, even one with no relation at all with the dependent variable, may appear to help. It is important to remember the context, and only add variables if it is reasonable to expect them to add some useful information.

It is a good idea to look at the confidence intervals for the slopes to see what the likely error is (see Section 9.3.2). The bootstrap method for deriving confidence intervals for multiple regression requires more sophisticated software than *resample.xls*,[139] so you need to use the Regression tool in

Excel or SPSS, as explained in the next section. The confidence intervals for the slopes in the models we have just set up (using the Excel Regression tool) are very instructive. For the three variable model, the 95% confidence interval for the sexn slope is –5 to 72. In other words, it could well be negative. On the other hand, the interval for the run time is –6.3 to –2.3. This is quite a wide interval, but it's all negative. We can be reasonably sure that the 'true' relationship is a negative one. This is not the case with the single variable regression: the confidence interval here was from –3.1 to 0.4. Turning to the five variable model, the confidence interval for the run slope is now much wider (–14.0 to –2.7) and that for the cycle time is even wider (–2.9 to 20.5). This confirms what we said above about the five variable model being unreliable.

▶ 9.5 Using the regression procedures in Excel and SPSS

You will find that SPSS uses a lot of jargon, and some of it does not correspond to the terms used in Excel. Table 9.7 summarises the important concepts, and gives some of the alternative names used for them. In Excel, use Tools – Data analysis – Regression for a full set of statistics and various scatter plots; these are useful for checking for any patterns that may indicate problems with the model (see Section 9.6). In SPSS use Analyze – Regression – Linear. You will need to click on Statistics then Confidence intervals if you want confidence intervals and Plots if you want plots.

As well as slopes, intercepts and confidence intervals, Excel and SPSS will give you *p* values (otherwise known as 'significance' or 'sig' levels); one for each of the slopes, one for the intercept and an overall one for the whole model. These are explained, in general terms, in Chapter 8.[140] The *p* value for the whole model is based on the null hypothesis that there is no relationship between the independent variables and the dependent. The *p* values for the slopes and the intercept give you some, but not all, of the information provided by the confidence intervals (Section 9.3.2). I would suggest focusing on the confidence intervals.

For a single independent variable, you can also use the Excel functions, forecast, slope and intercept (see Section 9.2), and if you right click on the points of a scatter diagram, you will be invited to add a trendline to the graph. The linear one is the regression line explained in Section 9.2; there are also some other options.

Table 9.7 Regression terminology

Concept	*Alternative names*
Regression model	Prediction model Least squares model Best fit model
Independent variable A variable used to help predict values of the dependent variable	Xs or X values Predictor variable Explanatory variable
Dependent variable The variable whose values are predicted or explained	Ys or Y values Predicted variable Explained variable Response variable
Error The difference between the actual value of the dependent variable and the value predicted by the model	Residual
Mean square error (MSE) The mean (average) of the squares of the errors for the predictions for the data on which the model is based	
R squared The proportional reduction in MSE provided by the model. For a single independent variable, R squared is equal to the square of the (Pearson) correlation coefficient. R squared = 1 suggests a perfect prediction; R squared = 0 suggests a useless one	Proportion of variance explained by the model
Slope Each independent variable has a slope indicating its impact on the prediction of the dependent variable. The predictions from two values of the independent variable separated by one unit will differ by the value of the slope	X coefficient Regression coefficient β (beta)* b
Intercept The predicted value if all independent variables are zero	Constant Base prediction

* Although some books reserve this term for the 'standardised' coefficients, see, for example, Lewis-Beck (1993: 56).

▶ 9.6 Things to check for with regression

If you've got some data comprising a few cases, and at least two variables, you will have no difficulty in building a regression model. But it may not make much sense. Always ask yourself the following questions.

Is the underlying model reasonable?

There are three checks you can do here. First you can have a look at the graphs. Figure 9.5 suggests that a straight line prediction is reasonable. Figure 9.6 suggests that it isn't. It's more difficult to do this with multiple regression, but both Excel and SPSS will produce graphs (plots) which give you some idea of any major discrepancies.

The second check is R squared. This, remember, tells you the extent to which the model has managed to reduce the error in making a prediction (MSE), so low values indicate that the model is not very useful. The value of 11.5% for Figure 9.6 indicates this model is much less useful than Figure 9.5, with its R squared of 92%. R squared depends on the number of variables and the number of cases, so be careful about using it to compare models with different numbers of variables or cases.

The third check is to see if the model makes sense in terms of your understanding of the situation. Look back at the model in Table 9.5. This implies that being female is worth an extra 33.2 thousand euros. The figure is the same regardless of age and run time. Similarly, the run time slope of –4.3 seems to imply that cutting one minute off your run time leads to an extra 4.3 thousand euros (on average). This is assumed to be the same for fast runners and slow runners, for males and females and for the young and the old. This can really only be a guess. You can use R squared as a crude measure of how good a guess it is.

What about sampling error? Is the sample large enough for the model to be reliable?

Check the confidence intervals for the slopes and the standard error of the prediction (Section 9.3.2). These should reveal how accurate estimates are likely to be. If you are using the Excel Regression tool or SPSS, strictly, the method for estimating these depends on the assumption that the errors should not show any strong relationship with any of the variables. You can check this using the plots provided by the package. If everything looks random, you should be OK. Any obvious pattern and the answers may be misleading.

Should you leave any variables out? Or include any extra ones?

The main thing to check is that there are no very high correlations (that is, approaching +1 or –1) among the independent variables (see Section 9.4). If there are, you should leave out one of the correlated variables. This is the reason for excluding the cycle time (strongly correlated with run time) and height (strongly correlated with sex) from the first model for earnings (Table 9.5).

You then need to come to a judgement about whether extra variables contribute enough to be worth putting into the model. Remember that the model will always squeeze something out, even if there's nothing useful there. SPSS has some built-in procedures for making this decision, search for Help on Stepwise.

▶ 9.7 Cause, effect, prediction and explanation

Let's return to the story of the Isle of Fastmoney. Over the years, the knowledge of the relationship between earnings and 10 km run time has had an impact on the life of the island. Ambitious people, keen to earn a lot, typically take their running seriously and train extensively. And employers have taken to using the relationship as a guide to recruitment: they are keen to employ fast runners and are willing to pay them more. In fact, they have largely ceased to take any notice of academic qualifications. Potential students have responded to this by staying away from the college and enroling at the gym. The economic performance of the island does not seem to have suffered, but the college is nearly bankrupt!

In Section 3.9 we saw the difficulties of using data from surveys to come to conclusions about causal relationships. The Isle of Fastmoney illustrates these difficulties well. There are many plausible hypothesis about causal relationships here:

- *Hypothesis 1*: People who can run faster can move around faster, work faster and get more done, so are likely to earn more.
- *Hypothesis 2*: A high salary means people have more money to spend and can afford a better diet or the right drugs, and employ the best training methods, so are likely to do better in the race.
- *Hypothesis 3*: There is a psychological quality of single-mindedness which can be applied both to work and running. Some people have it; others don't. Those who have it both run faster and earn more.
- *Hypothesis 4*: Employers *believe* that fast runners are more productive and so are prepared to pay them more.

All these hypotheses are consistent with the data. We've no way of knowing how much truth there is in each of them, which is a pity, because the four hypotheses have very different implications. For example, suppose Hypothesis 4 is true and the other three hypotheses are false. What would you do if you were an employer on the island**?**

It would be tempting to seek employees with poor run times in order to avoid paying exorbitant salaries for fast runners who are not likely to be any more productive.

If, on the other hand, there is some truth in Hypothesis 1, this would not be a good idea because slow runners are likely to work more slowly. But if Hypothesis 1 is false and Hypothesis 3 is true, it might still be worth employing fast runners because running fast is an indicator of determination. This is despite the fact that running fast is of no direct benefit.

The situation is even more confusing than this discussion suggests if we think there may be *some* truth in *all* the hypotheses: a fuzzy approach to truth (Section 4.6) seems more helpful than the crude assumption that each hypothesis must be true or false.

What regression allows us to do here is to *control*, or make allowances or adjust, for any variables we know about. The model in Table 9.5 tells us the relationship between earnings and 10 km run time if we make allowances for age and sex. Whatever age and sex we consider, the model predicts that a difference in run time of 1 minute will correspond to a difference in earnings of 4.3 thousand euros, with the negative sign indicating that the two differences go in opposite directions. This is just a model. We haven't got the data to make the comparison for real, so the regression enables us to compare hypothetical scenarios. Don't forget that regression models are based on assumptions which are seldom fully realistic (Section 9.6). They are guesses. Treat them with caution.

Could we use regression to test the truth of the four hypotheses above**?**

If we had data on some of the variables mentioned – diet, single-mindedness and so on – we may be able to derive some regression coefficients (slopes) to come to some tentative conclusions about the truth of each hypothesis, in a fuzzy sort of way. However, we would need to be careful. Observing a relationship between two variables can *never* prove conclusively that one causes the other (Section 3.9). To demonstrate this possibility, we would need to do an *experiment* (Section 10.1.3 and Exercise 10.5.2).

In this chapter I've described regression models as methods for making predictions. This is convenient for explaining how they work. In practice, however, we are often more interested in the understanding that the models give us, than in making predictions. On the Isle of Fastmoney, the predictions of earnings are useful because of the insights they give into how to make money. Similarly, a model of the relationship between smoking and lung cancer is more likely to be used to understand what causes lung cancer, than to predict who is going to die and when. This is recognised in some of the alternative regression terminology (Table 9.7). The independent variables are also called 'explanatory' variables because they help to explain the variations in the dependent variable. The errors are called 'residuals' because they are the bit left over which you can't explain. And R squared can be described as the proportion of the variation in the dependent variable which is *explained* by the regression model.

9.8 Similar concepts

Models like Table 9.5 are normally written in symbols:

$$y = 13.9 + 33.2x_1 + 4.0x_2 - 4.3x_3$$

where x_1 is sexn, x_2 is age over 20, x_3 is run time over 50 minutes and y is predicted earnings. The general form is:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3$$

The SPSS procedure General linear model (Analyze – General linear model – Univariate) combines the idea of regression with that of 'analysis of variance' (see Section 8.8). Independent category variables are called 'factors', and number variables are referred to as 'covariates'. Sex would be a factor in explaining income on the Isle of Fastmoney and age would be a covariate. The 'effect' of the factor sex is 33 (Table 9.5): this is another way of describing the slope of a dummy variable. If you click on Options – Display means, SPSS will produce means of female and male earnings 'adjusted' to particular values of the covariates. This provides essentially the same information as the slope for sexn in Table 9.5.

This procedure also allows you to analyse 'interactions' between the effects of the independent variables. For example, the effect of sex (on earnings) may be different for different age groups. If this is so, the model in Table 9.5, which assumes that you can measure the slope for each independent variable irrespective of the values of the other variables, may be unrealistically simple. You will need to refer to the SPSS manuals to see how SPSS deals with interactions.

The single variable regression model in Section 9.2 is called a 'linear' model, because the prediction line is straight (not because it's a line, as you might expect). The multiple regression model in Section 9.4 is of a similar form and is also described as linear. It is possible to use exactly the same least squares argument to build other, non-linear models. If you right click on the points of an Excel scatter diagram, you will be invited to add a trendline: there are several non-linear possibilities on offer.

A recent alternative to multiple regression is an 'artificial neural network' (ANN). These are designed to mimic the way natural neurons in the brain recognise patterns. They provide a way of making predictions without assuming any sort of linear model.

▶ 9.9 Exercises

9.9.1 Predicting earnings with the full sample of 100

In Section 9.2 we saw how to predict 10 km cycle times and earnings from the 10 km run time, and in Section 9.4 we saw how to set up a model to predict earnings from three of the variables (Table 9.5). These are all based on a subsample of 16. Obviously more accurate results would be possible if we used the whole sample of 100 in *iofm.xls*.

(a) Use *pred1var.xls* and *iofm.xls* to set up a model for predicting cycle time from run time. Check you understand how the spreadsheet works. How closely do the results agree with those in Section 9.2? How fast do you think you could run 10 km? Use your estimate to work out a predicted cycle time.

(b) Do the same thing for the prediction of the earnings from run time. Use the model to predict your earnings if you lived on the island. How much extra would your predicted earnings be if you could run 5 minutes faster? You should find R squared is less than in (a). What does this indicate?

(c) Use *predmvar.xls* to do the same for the prediction of earnings from sexn, age, and run time. Use the model to predict your earnings if you lived on the island. How much extra would your predicted earnings be if you could run 5 minutes faster?

(d) Finally, use the Excel Regression tool, or *resample.xls*, to work out confidence intervals for the slopes you found in (a), (b) and (c). Any comments?

9.9.2 Understanding drinking habits

Suppose you wanted to explain the amount students drink in terms of the other data in *drink.xls*. What's the best model for doing this and how good is it? Which are the most useful predictor variables?

9.9.3 Predicting returns on the stock market from what's happened in the past

A considerable amount of research has been done to see if it is possible to predict returns on the stock market from past data (see Exercise 3.11.5 for an explanation of what is meant by 'returns'). One such study[141] produced a regression model to predict the return which investors would receive from investing in a particular company's shares for a period of four years, from the return they would have received if they had invested in the same shares over the previous four years. The data on which the model was based were the returns for a sample of large companies over consecutive periods of four years. The regression coefficient cited was –0.112, and the value of R squared was 0.0413. What do these results mean? What implications do they have for investors? What else would you like to know?

9.9.4 Why square the errors?

You may have wondered why we square the errors and work out the mean square error in Section 9.2. Wouldn't it be easier to ignore any negative signs and just take the mean of the errors? This would be the mean absolute error (MAE). The snag with doing this is that you may end up with lots of different 'best fit' models, some of which may be intuitively unreasonable. To see the problem, try using both methods with this data:

Values of independent variable:	0, 0, 1, 1
Values of dependent variable:	1, 5, 3, 7

It is quite easy to do this without a computer: just draw a scatter diagram, sketch some possible best fit lines, and work out MSE and MAE (try a line with intercept 3 and slope 2, and another with intercept 1 and slope 6). Alternatively use *pred1var.xls*.[142]

9.9.5 What's the relationship between the model for predicting B from A, and the model for predicting A from B?

You should be able to see the answer by experimenting with different patterns of data. I would start with the data in Section 9.9.4 above, then try the patterns in Figure 3.7. The worksheet *reg2way.xls* may help.

9.10 Summary of main points

Suppose you have some data consisting of at least two number variables, for each of a number of cases. You can use this data to build a regression model for predicting values of one of the variables (the dependent variable) from values of one or more of the others (the independent variables). You can then use your model to make predictions (often very rough ones) for new cases and help you understand how the dependent variable depends on the independent variables(s). Take care of regression models: many of them are less sensible, accurate or reliable than they may look.