Brief notes on statistics: Part 4 More on regression: multiple regression, *p* values, confidence intervals, etc

Michael Wood (Michael.wood@port.ac.uk)

22 October 2012

Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. The data in the tables, and the figures, are in the spreadsheet, http://woodm.myweb.port.ac.uk/stats/StatNotes.xls . For a rough, but still useful, understanding, you can ignore the numbered endnotes, which provide extra detail.

This part depends on the concepts in Parts 2 and 3, so please make sure you have read these. If you have taken in everything in Parts 2 and 3, **you should already know most of what is in Part 4**, which is why there are a lot of questions in the text.

Multiple regression

In Part 2 we saw how to make predictions based on a single independent variable. The idea of multiple regression is to make predictions based on as many independent variables as we want. This gives us a very powerful, and very widely used, method of looking at the impact of a whole range of variables on our dependent variable. If, for example, you wanted to know what has an impact on employee satisfaction, you might use pay, working hours, holidays, type of job, etc as independent variables. A multiple regression analysis would then tell you which variables have a definite impact on employee satisfaction, and how big the impact is. We'll use a rather simpler example than this to show how it works, but the principle is the same for more complicated models.

Figure 2 in Part 2 (http://woodm.myweb.port.ac.uk/stats/StatNotes2.pdf) shows a prediction for Sales based on Advertising Spend. Suppose we now learn that the products being sold are warm coats. This means we would expect better sales in cold weather, and incorporating the temperature in the model may lead to better predictions. Table 1 below shows the sort of data we might have. (Again, this is not genuine: I have made it up.)

Table 1. Advertising Spend, average temperature and sales

Advertising Spend (£)	Average temperature	Sales (£)
200	6	8000
100	8	3500
400	10	11000
600	10	12000
0	12	1500
400	15	8000

Using more than one variable for a prediction is known as *multiple* regression. The method is just the same as before (see Part 2), except that we have two or more independent variables. In symbols, the model is

$$y = c + b_1x_1 + \ldots + b_nx_n$$

Or in words (for this particular model):

PredictedSales = Constant + AdvertisingSlope x Advertising + TempSlope x Temp

Just as before, the **method of least squares (see Part 2)** is used to find the values of the constant and slopes which best fit the data. I have set up another spreadsheet to do this (http://woodm.myweb.port.ac.uk/nms/predmvar.xls), but as the principle is exactly the same as for the spreadsheet used in Part 2, I won't bother to use this here. Instead I will use the Regression Tool in Excel.

To use the Regression Tool, put the data in Table 1 in an Excel worksheet, then click Tools (or Data in Excel 2007) – Data Analysis – Regression. If Data Analysis is not on the Tools menu, you will need to install the Analysis ToolPak: click Tools – Add-ins and tick the appropriate box. (In Excel 2007 you will need to click the Office Button on the top left, and then Excel Options to find the Add-ins.)

Remember that Sales is the **Y (dependent)** variable, and for the **X (independent)** variables you need to enter the **whole block** from A2 to B7. The results from the Regression Tool are in Table 2.

Table 2. Edited version of the results of applying the Excel Regression Tool to Table 1.

Regression Statistics					
Multiple R	0.97				
R Square	0.93				
Adjusted R Square	0.89				
Standard Error	1376.69				
Observations	6				
	Coefficients	Standard Error	P-value	Lower 95%	Upper 95%
Intercept	5583.88	2114.43	0.08	-1145.17	12312.94
X Variable 1	18.22	2.82	0.01	9.23	27.20
X Variable 2	-335.61	201.31	0.19	-976.28	305.07

R squared is worked out in a similar way to a regression with a single variable. As before it indicates how closely the model fits the data. (The only difference is that it's the square of a multiple correlation, not the ordinary correlation.)

Quick question 1. Use the results in Table 2 to predict sales if the Advertising spend was 500 and the temperature was 10. (Remember that "intercept" is another word for the Constant, and the coefficients are the slopes.)

Quick question 1a. Now make the same prediction from the model in Part 2 which used just one independent variable – Advertising spend. (The slope for this model was 17.2 and the intercept was 2446.) Your answer should be different. Why?

Quick question 2. What does the value of R squared tell you about this prediction?

Quick question 2a. How does this value of R squared compare with the value for the one variable prediction (from Advertising spend only) in Part 2. Does this make sense?

Quick question 3. Use the results in Table 2 to predict sales if the Advertising spend was 500 and the temperature was –30. Do you think this prediction is likely to be reasonable?

Prediction, explanation and variation

Regression models are often used to make predictions, but they can also be used for other reasons. Another common motive for setting up a regression model is to see *how* the independent variables can be used to ***explain* the variation** in the dependent variable, and ***how much*** of this variation is explained.

The model above shows that Advertising spend has a positive impact on Sales, and Average temperature has a negative impact. This information is given by the **slopes** (coefficients) in Table 2. This means that the Advertising spend and Average temperature help to ***explain*** the variation in Sales – why some of the sales figures are higher than others.

Quick question 4. Explain what the slopes mean in words that would make sense to someone who is not familiar with regression. Does the fact that the slope for Temperature is bigger than the slope for AdvertisingSpend (if we ignore the minus sign) mean that Temperature has a bigger influence on Sales than AdvertisingSpend?

The slopes are often the most important result of a regression model. They tell you the impact each independent variable has on the dependent variable. Often, the main thing to note is simply whether the slope is positive or negative.

You can think of R squared as a measure of the accuracy of predictions from the model. A value of 1 would indicate that the predictions are completely accurate (all the variation of the dependent variable has been explained); a value of 0 would indicate that the prediction is no better than using the average as a prediction and ignoring all the independent variables. This is explained in more detail in Part 2 at http://woodm.myweb.port.ac.uk/stats/StatNotes2.pdf.

Quick question 5. How much of the variation in sales is explained by the two variables, AdvertisingSpend and AverageTemperature?

Taking account of sampling error – p values and confidence intervals

Table 1 is based on a sample of only 6 observations. Obviously another sample of 6 may give a different result. R squared does *not* take account of this *sampling error,* although Excel and other packages will calculate an "adjusted" R squared which does take some account of sampling error. We also need to consider the effect of sampling error on the slopes for the independent variables. With small samples, the slopes may be unreliable as indicators of the effects of the variables.

If we assume that the sample is a random one from a wider population (or similar to a random sample), then **confidence intervals or *p* values can be used to assess the magnitude of sampling error.** The concepts here are those described in Part 3 at http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf.

Both are included in the Excel output. The 95% confidence interval for the slope for the first variable, the Advertising spend, extends from 9.23 to 27.20. This means that, based on this small sample of data, we can be 95% confident that the true slope (i.e. the slope we would get if we had a lot more data) lies in this interval. We cannot be at all certain of the exact slope – it might be 10, or 15 or 20. But we can be more than 95% confident that this slope is positive. On the other hand, for the temperature slope, the confidence interval includes both positive and negative values, so we can't be sure, based on this small sample, that the slope for temperature is in fact negative[12].

Quick question 6. You can use Excel to produce confidence intervals other than 95% ones. How do you think 50% confidence intervals would differ from those in Table 2?

Quick question 7. P values depend on a null hypothesis (see Part 3). What are the null hypotheses for the p values for the two slopes in Table 2? What conclusions can you draw from these p values? (The p value for the intercept is usually ignored because it is not of much interest.)

Quick question 9. Imagine that you had a larger sample (say 100) which results in similar values of the slopes and R squared. What can you say about the p values? And the confidence intervals?

The Excel Regression Tool produces a number of other statistics that I won't discuss here[3].

Regression in practice

Ayres (2007), in a book with a subtitle "how anything can be predicted", gives many examples of the successful use of regression to make predictions about the quality of wine, success in sport, customer satisfaction, marital satisfaction (see Q4 under Further work below), and many other things. Multiple regression gives a straightforward method of analysing the impact of a number of independent variables. He cites evidence for **regression giving better predictions than human experts (like wine experts, marriage guidance experts, and so on) in a very wide range of domains**. Obviously, there are also many situations where experts do better than regression models, but regression models deserve to be taken seriously, both for making predictions, and for understanding which factors are influential and how they influence the outcome.

The remaining paragraphs in this section deal with a variety of extra points you should be aware of when using regression models.

The example above uses a very small sample of six observations. This is to avoid confusing you with too much detail. In practice, **samples for regression need to be larger than this, and, not surprisingly, the more independent variables you have the larger the sample should be**. The width of the confidence intervals should give you an indication of whether your sample is large enough.

Regression is mainly for dealing with number variables, but it can be extended to **category variables** as well by using the two numbers 0 and 1 to denote two categories, or the absence or presence of some characteristic – these are called *dummy* variables[4] and you will find examples in the Exercises and Further work below. Many studies use regression to analyze the effect of a long list of independent variables – both numerical and dummy (0 / 1) variables. Often the main interest is in whether each variable's slope is positive or negative. See, for example, http://woodm.myweb.port.ac.uk/stats/MultReg.pdf: the variables at the end which "= 1 if …" are dummy variables.

Sometimes, one of the independent variables is of particular interest. The other variables may be **control** variables. The purpose of multiple regression here would be to make allowances for the control variables so that the impact of the main variable is understood more realistically. For example, in Table 1, the main purpose may be to investigate the impact of Advertising spend on Sales, but it is acknowledged that temperature may have an impact so this is introduced as

control variable. The regression coefficient for Advertising spend then tells us the impact of Advertising spend on sales, assuming that the temperature does not change.

You will find that sometimes regression coefficients are **"standardized"**, and you are told **"adjusted" R squared** (e.g. see Table 2). *Standardized coefficients will tell you whether the impact of a variable is positive or negative, but you can't interpret their size as we have done above (with a single independent variable standardised regression coefficients are the same as Pearson correlation coefficients).* The adjustment of R squared is designed to take account of two problems with R squared. First, R squared from a small sample is likely to be unreasonably optimistic (to take an extreme example, with a sample of 2, R squared will *always* be 1). Second, adding another independent variable will almost always increase R squared even if the extra variable adds no useful information. However, if the sample is large, and the number of independent variables is small, R squared and adjusted R squared will be similar. *In practice, it is usually a good idea to use adjusted R squared if this is available*[5].

You will also find that the regression analyses are often set up in the management research literature by stating formal **null and alternative hypotheses**. (One of many examples is Mathur et al, 2001 at http://tinyurl.com/yshfzx.) This is unnecessary and may be counter-productive. As an alternative to their Hypothesis 1, Mathur et al (2001) could simply have started from the research question "How does multinational diversification influence the financial performance of firms?" The answer would then be the appropriate regression coefficients, and *p* values or (preferably) confidence intervals.

It is important to **check the value of R squared (or adjusted R squared)**. In some fields, like management, the values of R squared that are cited are frequently disappointingly low. Remember that this means that the variables studied do not actually explain very much, or that the predictions made are not likely to be very accurate.

It is also worth remembering that the basic idea of regression is that you can use your data to derive a model that involves adding up terms for each of the independent variables. **Quite often, common sense should tell you that the resulting model is not likely to be much good.** For example, when you used the data in Table 2 of Part 2 to predict sales for an advertising spend of £2000 (Exercise 1e), the answer you should have got was negative! The explanation is, of course, that the values of Advertising spend in the data were all between £0 and £600: the results clearly cannot be extrapolated as far as £2000. And there are also a number of more technical pitfalls with regression[6], which you should check to be completely sure that your conclusions are as credible as possible.

It is also important to realize that calling a variable "dependent" does not make it depend on the independent variables in reality. For example, according to Huff (1973: 84) you can make a prediction (which is better than chance) of the number of children born in Dutch or Danish families by counting the **storks' nests** on the roofs of their houses. But this does not, of course, prove that the storks somehow cause the babies, just that the two variables are correlated!

Quick question 10. Can you think of another explanation for this correlation?

Quick question 11. Look at a regression table in a research article. What does it tell you?

You will find many examples of regression in the research literature – e.g. Table 2 in both Glebbeek and Bax (2004 – copy at http://woodm.myweb.port.ac.uk/stats/CurveReg.pdf) and Mathur et al (2001). You will also find many more **sophisticated regression methods** for specific situations (e.g. hierarchical regression, logistic regression).

Summary

Regression models can be used to predict the value of a *dependent* variable based on values of one or more *independent* variables. A sample of data is used to set up the model, which is simply an equation for making the prediction. The *slopes* (coefficients) of this model indicate the predicted impact of each of the independent variables on the dependent variable.

The value of *R squared* (the coefficient of determination) indicates how accurate the prediction is likely to be, with an R squared value of 1 suggesting a completely accurate prediction, and a value of 0 (zero) suggesting that the model is of no help in predicting values of the dependent variable.

You should also remember that the slopes derived from small samples are not likely to be reliable. The effects of sampling error can be assessed using *p* values or confidence intervals (as described in Part 3).

A regression model that makes good predictions can also be thought of as providing a good *explanation* of the factors that cause variations in the values of the dependent variable.

Terminology

There is a confusing variety of words used in relation to regression. This is a brief guide:

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	X's or X values Predictor variable Explanatory variable
Dependent variable The variable whose values are predicted or explained	Y's 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 Coefficient of determination
Slope Each independent variable has a slope indicating its impact on the prediction of the dependent variable. If an independent variable is increased by one unit, the prediction for the dependent variable will increase by the slope of the independent variable (provided that the other independent variables remain the same).	X coefficient Regression coefficient β (beta)* b
Intercept The predicted value if all independent variables are zero.	Constant
P value See Part 3.	Significance level
Confidence interval See Part 3.	Described as “lower 95%” and “upper 95%” in the Excel output

* Although some books reserve this term for the “standardised” coefficients.

Exercises

1 Mariella Spark from the USA has just got married and landed a well paid job in an IT business in Tregynon in Wales. Her plan is to have lots of babies which her husband will look after, and the house in which this happens is very important to both of them. Having just done an MBA at Portsmouth University she decides to use multiple regression to model the factors which influence house prices in Tregynon. For her pilot study she goes to a website, specifies that she wants a house costing $700,000 or less in Tregynon, and downloads details of the 10 houses in the table below (in the Part4Data sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls).

Pilot sample of houses in Tregynon costing $700,000 or less

House	House size (sq feet)	Garden size (acres)	Location	Sale price ($000)
1	1000	0	Town Centre	569
2	2300	0.5	Town Centre	682
3	1500	1.9	Town Centre	575
4	2100	1	Town Centre	646
5	3900	1.1	Suburbs	485
6	3100	2	Suburbs	460
7	3600	1.6	Suburbs	486
8	2900	2.5	Suburbs	404
9	2000	2.6	Suburbs	427
10	3500	1.3	Suburbs	457

(a) The data suggests that houses in the town centre are more expensive. What is the difference between the average house price in the town centre and in the suburbs?

(b) She expected that big houses would cost more than small houses, so there should be a positive correlation between House size and Sale price. Calculate this correlation. Is it positive as she expected? Now draw a scatter diagram. Is this helpful?

(c) Use the Regression Tool in Excel to produce a model for predicting Sale price from the three other variables in Table 6. **(**You will need to code Location as a number variable: the usual way to do this with two categories is to code one as 0 and the other as 1.**)**

(d) Calculate the prediction of the regression model for the price of the first house. What is the error in this prediction? Now calculate the prediction for a house Mariella would like to find - one in the town centre with a House size of 4000 square feet and a Garden size of 2 acres. Can you work out the error in this prediction?

(e) What does the model tell you about the influence of the three variables on Sale price? You should write down, and think about, the X variable coefficients (slopes), the value of R squared,

and the p values or confidence intervals for the X coefficients. How do the multiple regression results compare with your answers to (a) and (b)?

(f) Which has the bigger influence on Sale price - the House size or the Garden size? Is this a sensible question?

(g) Now run the regression with just one independent variable - the Location. Do the same with the House size. Think about your results.

(h) How can Mariella use this model to help her decide which house to buy?

(i) If Mariella wanted to do some more extensive research, with a larger sample and more independent variables, what extra variables would you suggest? Remember the aim is to get as accurate predictions as possible. (If you want to have a go with a larger data set, there is one with data on American cars at http://www.amstat.org/publications/jse/datasets/kuiper.xls.)

Additional exercises

2 The spreadsheet at http://woodm.myweb.port.ac.uk/nms/predmvar.xls can be used to do multiple regression. This works in much the same way as http://woodm.myweb.port.ac.uk/nms/pred1var.xls which is explained in http://woodm.myweb.port.ac.uk/stats/LeastSquares.pdf, except that you can include more than one independent variable. Use this to do the regression analysis for this data instead of using the Regression Tool. (On the Data sheet, put the House size data in as Indep 1 starting at B8. Garden size and Location are Indeps 2 and 3, and Sale price is the Dependent in Column G. Then click on the Model sheet, and then Click Tools – Solver as before.)

Check that the results are the same. What is the mean square error in the final model? Notice that multiple regression works in just the same way as the regression with a single variable – by the method of least squares (adjusting the slope and constant to find the values which give least possible mean square error).

This spreadsheet also allows you to see the effect of the control variables on the predictions. Click on the Graph tab, and try changing the control variables to see what effect this has.

3 The two examples above both use unreasonably small samples (in the interests of keeping things simple). The data in the Turnover sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls is a "proper" data set which was used by Glebbeek and Bax (2004) to investigate the effect of staff turnover on the performance on organizations. What conclusions would you come to about this effect from this data? (Absenteeism, age and region were used as control variables.)

4 It is possible to use the Regression Tool to test the difference of two means like we did in Part 3. Try this with the medium sample in the Part 2 data in StatNotes.xls (at http://woodm.myweb.port.ac.uk/stats/StatNotes.xls). You will need to make the Institution

variable into a numerical dummy variable by letting (say) Bank be 0 and BS be 1. Write down the regression coefficient and the *p* value, and check they are consistent with your earlier results in Part 2. The Regression Tool also gives you a value of R squared, and a confidence interval.

5 The data file iofm.xls (http://woodm.myweb.port.ac.uk/nms/iofm.xls) gives data on a random sample of 100 people from the (fictional) Isle of Fastmoney. The final column (Earn000E) gives annual earnings in Euros, and two columns before this are times in minutes in a 10 km running race, and a 10 km cycling race. Sexn gives sex coded as 1 or 0 so that it can be included in the regression – this is called a *dummy* variable. The other columns should be self explanatory.

(a) Set up a model for the prediction of earnings from sexn, age, and run time. Write down the regression coefficients, the constant and R squared. Are they what you would expect?

(b) 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?

(c) What is the effect of sex? What difference to predicted earnings does the sex of the individual make? What difference does it make if an individual is one year older (with other variables being the same)?

(d) Now do a regression with just one independent variable – the run time. Compare the regression coefficient (slope) for run time in this regression to the coefficient you got in (a). It should be different. Why? How would you explain the meaning of this regression coefficient to someone not familiar with regression?

(e) Now compare the values of R squared in the two models. They should be different. Why?

(f) Now try adding a fourth independent variable – run time. Are the results sensible? (Run time and cycle time are highly correlated: the usual advice is that it is not a good idea to include both of a pair of highly correlated variables in a regression analysis. This exercise should show you why.)

6 Suppose you wanted to predict the amount students drink on Saturday night from the data on sex, age and cigarettes smoked in *drink.xls*. (This was obtained from three classes of (real) students a few years ago. Satunits means the number of units of alcohol the student drunk on the Saturday before the survey; Sununits and Monunits are defined in a similar way.) What's the best model for doing this, and how good is it?

7 According their website (http://people.ischool.berkeley.edu/~atf/dating/ accessed on 30 October 2008) a team researching online dating at the University of California intend:

> … to answer the following questions for as many data sets as we can obtain:
>
> - What predicts how many messages a given user will receive?
> - What fraction of dyadic exchanges are "successful"? What attributes differentiate "successful" from "unsuccessful" exchanges? (We can define success as the length of the exchange, or rate of transition to another medium, like telephone or off-site email.)

- What is the effect of having a photo on communication success? Of having a video clip?
- How do men and women differ in message-sending behavior? Do they value different attributes?
- What about different ethnic or cultural groups, or different regions of the country?

One of the methods they propose using is regression. What do you think they might be able to achieve? Can you foresee any difficulties?

References

Ayres, I. (2007). *Super crunchers: how anything can be predicted.* London: John Murray.

Glebbeek, A. C. & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal, 47,* No 2, 277-286.

Huff, Darrell. (1973). *How to lie with statistics.* Harmondsworth: Penguin.

Mathur, I., Singh, M., & Gleason, K. C. (2001). The evidence from Canadian firms on multinational diversification and performance. *The Quarterly Review of Economics and Finance, 41*, 561-578.

Norusis, M. J. (1993). *SPSS for Windows Base Ssystem User's Guide Release 6.0.* Chicago: SPSS Inc.

Wood, M. (2003). *Making sense of statistics: a non-mathematical approach.* Basingstoke: Palgrave.

Answers to Quick questions and some Exercises

1 PredictedSales = 5583.88 + 18.22 x 500 + (-335.61) x10 = 11337.8.

2 The value of R squared (0.93) means that the prediction is likely to be fairly accurate. A value of R squared of 100% would indicate a perfect prediction in the sense that the prediction is accurate for all observations in the data, so 93% represents a fairly accurate prediction (where accuracy is measured by MSE[7]). However, this 93% is based on the assumption that the sample exactly represents the pattern of data in the whole population; it does not take account of the fact that the sample is very small and another similar sample might give a substantially different result.

2a The value of R squared from the multiple regression is greater than for the single variable regression in Part 2 (which was 0.87). This should make sense because we are including another variable so the prediction should be more accurate.

3 The prediction is £24762.2. However, the data is all based on the temperature range 6 to 15. Using the regression model for a temperature of -30 obviously requires us to assume that the pattern is the same for this temperature. This seems unlikely, so the prediction is likely to be very unreliable.

4 The slope for AdvertisingSpend (18.22) means that the model predicts that Sales will increase by £18.22 if the AdvertisingSpend is increased by £1, *assuming* that AverageTemperature remains the same. A similar argument for AverageTemperature indicates that a one degree rise in temperature leads to a *decrease* in Sales of £335.61.

This does mean that a one degree Celsius change in temperature has a bigger impact that a £1 change in advertising, but this comparison is not really fair. Suppose that AdvertisingSpend was measured in thousands of pounds instead of pounds. This would change the regression coefficient to 18,222 which makes advertising more influential than temperature. With any comparison like this, you need to take account of the size of the change you are considering.

5 93%, or 89% using adjusted R squared (see below) which gives a slightly better estimate for the population as a whole. This is the amount of variation which is *explained by the model.* This means, of course, that the remaining 7% (or 11%) is not explained. This unexplained variation corresponds to the mean square error (see Table 5 of Part 2). (If the prediction was perfect, or everything was explained, there would be no errors, and the mean square error, or the unexplained variation, would be 0.)

6 They will be narrower.

7 The null hypotheses are that each of the coefficients are equal to zero over the population as a whole. (In other words, if lots of similar data from the same source were collected, and a regression were run using this much larger data set, the slopes would both be zero.)

Only the slope for AdvertisingSpend is significant at 5%. This means we can be reasonably sure that the AdvertisingSpend null hypothesis is false, and that AdvertisingSpend does have an impact on Sales, but the evidence is much weaker (and statistically insignificant at the 5% level) for AverageTemperature.

9 The *p* values would be smaller, meaning that the evidence for a real impact of each variable on Sales would be stronger. The confidence intervals would be narrower, because the larger samples allow us to make more accurate estimates.

10 Many possibilities here – perhaps people with lots of babies need large houses which provide more room for storks …

Exercise 1: There is a video on this question at http://youtu.be/G4lEJBCNVjw .

(a) The average price of houses in the town centre is 165 ($000) more than the suburbs.

(b) The correlation is negative, contrary to what Mariella expected. The scatter diagram explains why – there are two distinct groups of houses each of which show a positive relation between the variables. However, the *overall* relationship is negative because the bigger houses are in the town centre where they tend to be smaller and more expensive.

(c) The slope for the first X variable, House size, should be 0.049 …

(d) The prediction for the first house should be 587.9 ($000) and the error is 18.9. The predicted price for the house Mariella would like is 723.4. You cannot work out the error because there is no similar house in the data.

(e) The regression model predicts that an extra square foot of house size costs about an extra $50, and that houses in the town centre cost about 229 more. At first sight, these results are not consistent with the answers to (a) and (b): you should think about this to make sure you understand the reasons for the apparent discrepancies. The 229 figure is a prediction of the extra cost for a town centre house compared with a *similar* house in the suburbs. The answer to (a), of course, is comparing larger houses in the suburbs with smaller houses in the town centre. The *p* values and confidence intervals indicate that, other things being equal, bigger houses and houses in the town centre really do cost more. On the other hand, the *p* for the negative garden size slope suggests that you can't be sure if this is positive or negative – with another sample it may well turn out to be positive.

(f) This is not really a sensible question! The regression coefficient for the Garden size is much bigger, but this is measured in acres (an acre is 43560 square feet) instead of square feet, so this is not surprising!

(g) With Location as the only independent variable, the slope is 165 (or -165 if you coded the town centre as 0) and the p value is 0.0003 – so this is statistically significant. With house size as independent variable, the slope is negative but the p value is 0.15 which is not significant. These two answers ignore the other variables – they are similar to your answers for (a) and (b).

(h) There are two obvious things she could do. She could use the model to predict the price of a house – if it's below the predicted price it may be good value. And she could look at the slopes and ask herself if, for example, an extra square foot of house size is worth $50 to her, and if it is worth an extra $229,000 to live in the town centre.

Endnotes – these include more detail which can be ignored for a rough understanding

[1] This illustrates how you can work out whether a result is significant at the 5% significance level from the 95% confidence interval.

[2] For another approach to deriving confidence intervals for regression slopes, and also a direct illustration of what conclusions might be derived from different samples from the same source, see http://woodm.myweb.port.ac.uk/BRLS.xls. and http://woodm.myweb.port.ac.uk/BRL.xls.

[3] The standard errors in Table 2 are perhaps worth mentioning. In rough terms, they are standard deviations which can be interpreted in terms of the normal distribution as explained in Part 1 (http://woodm.myweb.port.ac.uk/stats/StatNotes1.pdf). For example, if the standard error of a prediction is 6 units, this suggests that there is a 68% chance that the prediction will be within 6 units of the true value, and a 95% chance that it will be within 12 units. For large samples, this is a good approximation, but for small samples like the six in Table 1, it is a very rough approximation. You should also note that the way that *p* values and confidence intervals are calculated depends on certain assumptions which may not be met in practice – see, for example SPSS user guides such as Norusis, 1993 or later editions.

[4] Dummy variables can only be used for two categories like yes/no or male/female. If, say, you had three makes of car – Ford, Toyota and Nissan – and you coded them 0, 1 and 2 – you would be assuming that Toyota is numerically between Ford and Nissan which obviously makes little sense. To deal with this with regression you need three dummy variables: Ford (0/1), Toyota (0/1) and Nissan (0/1).

[5] A good source for more detail on all this are the SPSS user guides such as Norusis, 1993 or later editions.

[6] For example, there may be problems if two or more of the independent variables are strongly correlated (say, more than 0.9 or less than – 0.9), or if you incorporate too many variables (the statistics package SPSS has a procedure called stepwise to help users decide which variables should be included). More details should be in one of the SPSS user manuals (e.g. Norusis, 1993 or later editions).

[7] You should be able to work out what the MSE is from the value of R squared (0.93) and the standard deviation of the Sales figures in Table 1 (3760). If in doubt, reread the explanation of R squared as proportional reduction in error in Part 2. You can use predmvar.xls (at http://woodm.myweb.port.ac.uk/nms/predmvar.xls) to check your answer.