# The method of Least Squares – using the Excel Solver

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

*22 October 2012*

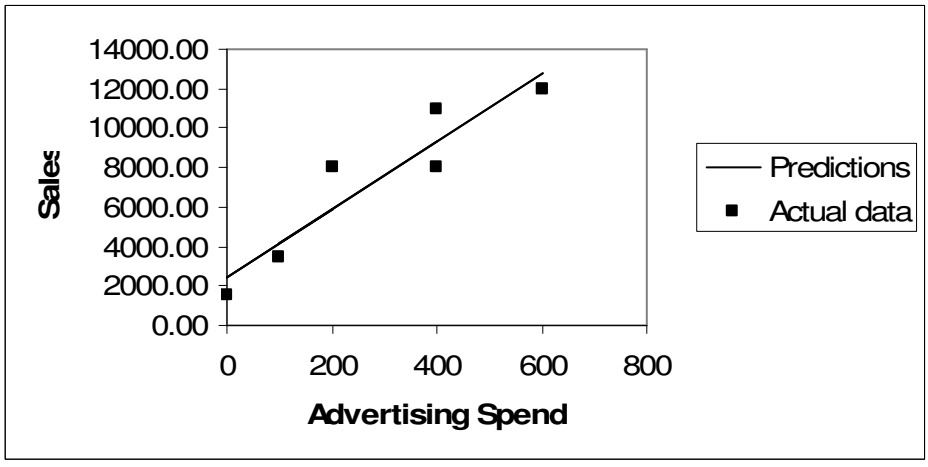
(Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. The data in the tables, and the figures, are in the spreadsheet at http://woodm.myweb.port.ac.uk/stats/StatNotes.xls and there is a video at http://youtu.be/RIAcq0NMGtA.)

My aim here is to explain the least squares method used in regression. I will use the Excel spreadsheet at http://woodm.myweb.port.ac.uk/nms/pred1var.xls. (There is a similar spreadsheet at http://woodm.myweb.port.ac.uk/nms/predmvar.xls which shows multiple regression works.)

We'll start with the following table of data:

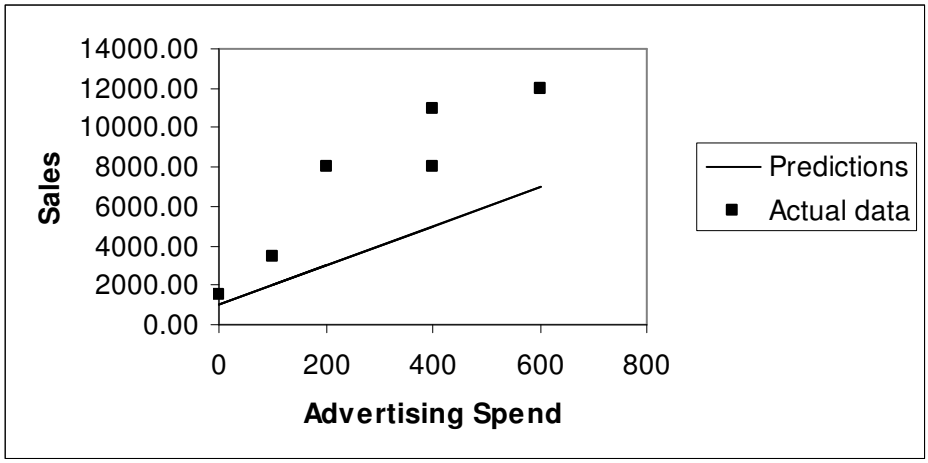
| Advertising Spend (£) | Sales (£) |
|---|---|
| 200 | 8000 |
| 100 | 3500 |
| 400 | 11000 |
| 600 | 12000 |
| 0 | 1500 |
| 400 | 8000 |

We want to produce a regression line to enable us to predict the Sales figures from the Advertising spend. The answer looks like this:




How is the line worked out?

Let's start with the wrong line. Then we can see how to improve it. The next figure shows a line which does not fit in any sense.




We now look at the error in making a prediction for each point. Let's take the first point in the table above for which Advertising Spend is 200. The actual Sales figure corresponding to this is 8000, but the prediction for this in Figure 3 is 3000 (= 10 x 200 + 1000). The error here is obviously £5000 (= 8,000 – 3,000). (Or −£5000 if you do the subtraction the other way. Don't worry which is right – it doesn't make any difference.) This is obviously a lot more than the corresponding error in the first figure. (The error is the vertical distance between the point representing the actual data and the line.) The reason the line in Figure 2 makes better predictions than the line in Figure 3 is that the errors made by the first line are smaller.

The spreadsheet pred1var (at http://woodm.myweb.port.ac.uk/nms/pred1var.xls ) is designed to show this process in action. I will include tables and diagrams from this spreadsheet below, but I would recommend that you open the spreadsheet and experiment as we go along.

Start by putting the data into the spreadsheet. (To put the data into pred1var, click on the Data worksheet tab at the bottom. The green cells are intended for your entries. In the example above, Sales is the *dependent* variable, and Advertising Spend is the *independent* variable. To enter the data you just have to put the Advertising Spend figures in cells B8 to B13, and the Sales figures in cells C8 to C13. You can either do this by typing them in, or go to http://woodm.myweb.port.ac.uk/stats/Pred1varTable1v2.xls for a version of the spreadsheet with the data entered.**)**

In practice, for good reasons (see Wood, 2003, p. 166), the measure of error used is the square of this error: 25,000,000 (= 5000 x 5000). To find the "overall" error for the line above, we just work out the average of these square errors, known as the ***mean square error*** **(MSE)**:

*Calculation of MSE (from the Model sheet of pred1var)*

| Refno | Actual Sales | Sales Prediction | Error | Square error |
|---|---|---|---|---|
| 1 | 8000 | 3000 | -5000 | 25,000,000 |
| 2 | 3500 | 2000 | -1500 | 2,250,000 |
| 3 | 11000 | 5000 | -6000 | 36,000,000 |
| 4 | 12000 | 7000 | -5000 | 25,000,000 |
| 5 | 1500 | 1000 | -500 | 250,000 |
| 6 | 8000 | 5000 | -3000 | 9,000,000 |
| | | | MSE | 16,250,000 |

This table shows the working. All we do is work out the square error for each of the six data points, and then find their average – the MSE. The mean square error for this line is just over 16 million! The square root of this (root mean square or RMSE) is 4031. This gives us a figure (£4031) for the average error we are making if we use this line to make our predictions. (It's a slightly unusual sort of average, like the standard deviation, but still an average.) The corresponding figures for the line which fits better in Figure 2 are MSE = 1,825,503 and RMSE = £1351. These are much lower, reflecting the obviously better fit of this line.

How can we find this better fit line? One way would be trial and error. Just experiment with different slopes and constants until you find the values which make MSE as low as possible.

*Quick question 1. What's the best fit (as measured by MSE) you can get by adjusting the slope and constant?*

However, help is available! Excel has an add-in called **Solver** which will automate this trial and error process. To run this, go to the Model sheet (click on the tab at the bottom), and click Tools (or Data in Excel 2007). With luck, Solver will be on this menu. If not, you will need to install it by clicking Add-Ins and ticking the Solver 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.) The relevant entries should appear automatically – check them to make sure they make sense. Then click Solve and you should have the results below:

***Calculation of MSE for best fit line (from the Model sheet: values for the constant and slope have been found by Solver)***

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | Constant | Slope | | MSE | 1,825,503 |
| | | | 2446.313 | 17.24831156 | | RMSE | 1,351 |
| | | | | | | PRE | 0.87 |
| | ACTUAL DATA | | MODEL | | | | |
| Refno | Advert'g | Sales | Constant | Extra Sales for Advert'g | Sales Prediction | Error | Square error |
| 1 | 200 | 8000 | 2446.31 | 3449.66 | 5895.98 | -2104.02 | 4,426,920 |
| 2 | 100 | 3500 | 2446.31 | 1724.83 | 4171.14 | 671.14 | 450,434 |
| 3 | 400 | 11000 | 2446.31 | 6899.32 | 9345.64 | -1654.36 | 2,736,915 |
| 4 | 600 | 12000 | 2446.31 | 10348.99 | 12795.30 | 795.30 | 632,502 |
| 5 | 0 | 1500 | 2446.31 | 0.00 | 2446.31 | 946.31 | 895,508 |
| 6 | 400 | 8000 | 2446.31 | 6899.32 | 9345.64 | 1345.64 | 1,810,740 |

This (just over 1.8 million) is the lowest you can make the MSE with this data. The constant is just over 2446 and the slope is just over 17. (Solver has found these numbers and put them in the worksheet.) Any other values of the constant and slope will give higher values of MSE. This line is the best fit possible! **This method is known as the method of *least squares* because the idea is to make the *squares* of the errors as small as possible. It is a method very widely used in statistics**.

There is another essential bit of information provided by the least squares method. This is **PRE which is 0.87 or 87%**. This stands for "proportional reduction in error" (not a standard and widely used phrase, unlike MSE and RMSE). This is a measure of how well the data fits the prediction. We'll look at the detailed rationale behind it below, but first I'll mention two commoner names for PRE.

Provided you have used the Solver to find the best fit model, PRE is also known as the *coefficient of determination* or as **R squared**. This **measures how well the model fits the data – R squared 0 representing a useless prediction where the independent variable is of no help, and 1 representing a perfect prediction with no errors.** The reason for the name R squared is that it is equal to the square of the correlation coefficient, which usually has the symbol, *r*.

To see where the idea of PRE and R squared comes from, look again at the table above. The difficulty with MSE as a measure of how well the model fits the data is that it depends on the size of the numbers involved. The variable we are trying to predict is Sales which is measured in thousands of pounds. This is the reason why the MSE figures look rather large.

PRE gets round this problem by using MSE to define another scale, going from 0 representing the case where the independent variable is of no help in making the predictions, to 1 representing the case where the independent variable enables us to make perfect predictions.

We can see what will happen if the Advertising Spend is of no help by ignoring it. The obvious prediction then it to predict that the Sales will be equal to the mean of the sales, regardless of what is spent on

advertising. The mean of the Sales is £7333.33 so this is our prediction. The MSE now is about 14 million – 14,138,888 to be exact[i]. **(**You can get this on the model sheet by putting £7333.33 in for the constant, and 0 for the slope – if you are not sure why, try it and look at the Graph sheet.**)**

For a perfect prediction, on the hand, there would be no errors, and the MSE would be 0. Obviously this would only be possible if all the data points lie in a straight line, which they do not in this case. In practice the best fit line here (Figure 2) has MSE=1,825,503 (Table 5).

*Quick question 2. What % reduction in MSE from the worst case of 14,138,888 does this best fit line represent?*

The least squares method is very widely used in statistics. There is a similar spreadsheet at http://woodm.myweb.port.ac.uk/nms/predmvar.xls which shows how it works in multiple regression.

## Exercise

**1** Use the data in *Table 2* in the Part2Data sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls and the spreadsheet at http://woodm.myweb.port.ac.uk/nms/pred1var.xls to produce a regression model. **(C**lick on the Data sheet in the second spreadsheet. Enter the six Advertising Spend figures in green cells at the top of the Independent variable column [cells B8 to B13], and the Sales figures in the next column. Now click the Model sheet, and use Solver. This should be under Data (or Tools in Excel 2003). If not, you will need to install it: in Excel 2007 you will need to click the Office Button at the top left and then Excel Options to find the Add-ins.)

(a) Check you can see how the model works.

(b) Use this regression model to predict the sales that would result from an advertising spend of £500, and £2000. Are the answers sensible?

(c) Write down the MSE, slope and the coefficient of determination (R squared), and make sure you understand what they mean. What does the slope tell you about the impact of advertising on sales?

(d) Check that R squared is the square of the correlation coefficient.

## Notes on answers to Quick questions

*Quick question 1.* The best answers are in the table above (constant = 2446, slope = 17.245, MSE = 1825503), but you are unlikely to get these without computer help!

*Quick question 2.* The reduction is 12,313,385, which is 87% of 14,138,888. The PRE, or R squared, is 87%.

[i] 14,138,888 is the *variance* of the original data. This is the square of the standard deviation, or the answer you get before the final step of taking the square root (see Part 1 of these notes). If you compare what we are doing here, with the method for calculating the sd, you should see the similarities.