Drawing Regression Lines on Excel Graphs

file: d:\B173-2013\regression.wpd
date: September 15, 2013

Introduction

Previously, everyone constructed a simple XY graph. Be sure that you can do this; it is essential for what we do from now on.

The next step to using Excel to help us understand fisheries science is to be able to plot a relationship on the graph. For example, we might wish to plot TL versus SL, and on the same graph, we might wish to draw the line of best fit. Such a line is called a regression line. In order to do this, we need to know a few more things about how to make spreadsheets work for us.

Regressions

Recall that the equation of a line can be written as

$$y = mx + b$$ (and you thought you'd never use that!)

This means that if you know the slope (m) and the y-intercept (b) of a line, then given any value of x, you can calculate the appropriate y. We will use this A LOT in this course so get good at it now.

This can be really useful. For example, looking at the data in the Exercise from last week, it certainly appears like there is a positive relationship between the two variables. In other words, as the SL increases, so too does the TL. But by exactly how much does the TL increase for every increment of SL? Prior to computers, you might be inclined to "eyeball it", i.e., take a ruler and draw a line that seems to best go through the points.

Eyeballing it can be remarkably accurate but it is not repeatable: the line I might draw would be different than the line you might draw. There is a method that takes the chance out of the process. It is called generating a least-squares regression. We needn't be concerned at this point exactly why it works, but suffice it to say that a least-squares regression tells you the line that minimizes the distance (actually the square of the distance) between the line and all of the data points. This is said to be the line of "best fit".

So, how do you find the slope and intercept for this "best fit" line for a given set of data? As you might imagine, spreadsheets are quite good at exactly this sort of thing.

Once we know what the line is, how do we plot this line on the actual graph?

There are automatic techniques built into Excel for doing this, called TREND or TRENDLINE. **DO NOT USE THE TREND or TRENDLINE procedures -- they often give incorrect answers in terms of what you expect that they do. Using the TRENDLINE procedure on an exam will result in you failing the entire lab portion of the course. YOU HAVE BEEN WARNED.** Why? Because sometimes they give the "right" answer and sometimes they give a different answer for reasons that are not likely obvious to you. In addition, understanding how to get a line plotted on a graph in a

spreadsheet will prove very useful later on when things get much more complex and we are drawing curves, not straight lines.

How we plot a line on a graph is simple, though a little unconventional to most people's way of thinking, i.e., we do not simply say: here is the formula for a line, plot it. Rather, if you know the formula describing a line (or any other shape) then all you need to do is enter a bunch of x values, compute the appropriate y values using the formula, plot those values and connect the dots. Why does this work? [Because all the y values will be on the line, right?]

So, for example, if the formula is $y = (3 * x) + 10$, or more correctly,

$$TL = (3 * SL) + 10$$

then you enter a column of SL values for the range that you are interested in, such as 100, 200, 300, 400 and 500 and then to the right of each SL value, you enter the formula that will give you the correct TL value, i.e., = (3*SL)+10 where SL is the address of the corresponding SL value:

It will look like this:

	A	B	C	D...
1	SL	TL		
2				
3	100	=(3*A3)+10		
4	200	=(3*A4)+10		
5	300	=(3*A5)+10		
6	400	=(3*A6)+10		
7	500	=(3*A7)+10		
...				

But, surely this is way too tedious to do. There must be a faster way than all that typing. Recall that copying is relative, i.e., if I have a formula in D1 that says (A1+B1), when I copy that to D2 it will now say (A2+B2) because D2 is one cell down from D1. [If you do not understand this, **STOP RIGHT HERE AND ASK ME BEFORE CONTINUING!!!!**]

So, we could enter the formula in B3 and then copy it down from B3 to B7. That would be much less typing.

Absolute versus relative values

But, what if we didn't know the value for the slope, i.e., the '3' in the (3*A3)+10, but rather we were going to calculate it somehow in the spreadsheet? Let's say it ended up in cell D1. No problem, we just change the formula in B3 to read (D1*A3)+10. Now copy it down.... But wait.... Things don't look right. Look carefully at B7. Does it read the way you think it should?

No, it doesn't. This is because when you copied the formula, not only did it change the A3 relatively but it also changed the D1 relatively, which is not what you wanted. You wanted the D1 to always stay D1, not to change.

To do this, you have to make it explicit in the original formula in B3 that the D1 is to always refer to exactly and only that cell. You do this by putting a little '$' in front of the thing you want to make absolute, i.e., not relative. Note that either the D or the 1 or both could be absolute or relative. In this case, we want both to be absolute, so we change the formula in B3 to read ($D$1*A3)+10. Now copy this down from B4 to B7 and see that it works correctly.

**Think about what we have done. Putting the $ sign in a formula makes no difference to the value calculated in that particular cell; it only affects what will happen when you copy that cell to another location. Be sure that you understand this.

Similarly, you might want to calculate the intercept of the line, i.e. the '10'. Assume that it is stored in D2. Put “10" in D2 and adjust the formula in B3. The (final!) formula becomes ($D$1*A3)+$D$2. Copy this down and see that it works correctly.

Do you see that if you were to change either the value of the slope (D1) or the value of the intercept (D2), the rest of the formulas will automatically incorporate the change and recalculate appropriate values?

You absolutely must understand all of this completely to proceed.

Calculating a regression for data

Let's say we want to draw the regression line for the data we plotted last class. To do so, we need the formula of the regression line and that means that we need to find the slope and y-intercept of the line. Excel's SLOPE and INTERCEPT function can find these values for us. The SLOPE function takes two ranges as its input (the y's, then the x's), and returns the slope. The INTERCEPT function operates similarly, but returns the intercept.

Enter the data from last class, putting your name in cell A1, "Fish" in cell A6 and so on, i.e., the 340 ends up in cell C17 (as shown below).

In cell A19 type "Slope" and in cell A20 type "Intercept"

	A	B	C
1	Fisheries Regressions		
2	Name:	Ron Coleman	
3	Date:	September 15 , 2013	
4	File:	regression.xlsx	
5			
6	Fish	SL (mm)	TL (mm)
7			
8	1	135	150
9	2	182	200
10	3	203	220
11	4	220	245
12	5	234	247
13	6	256	280

14	7	266	279
15	8	277	290
16	9	287	310
17	10	300	340
18			
19 Slope			
20 Intercept			

Now, calculate the slope and intercept of the regression of TL on SL as follows:.

In cell B19 select the Formula tab. In the drop-down box, choose “More Functions” then "Statistical", then "SLOPE"

Click on the mini-spreadsheet icon for the y's, highlight C8 to C17 and hit Enter (recall the y’s are TL). Click on the mini-spreadsheet icon for the x's, and select B8 to B17 and hit Enter (recall the x’s are SL).

Then hit "OK".

The value that you see, i.e., 1.07 is the slope. You should format the cell to show only 2 decimals.

Use the INTERCEPT function to put the intercept in B20. You should get 3.85.

[There are other ways of using Excel to get these values, e.g., you can use the LINEST function, which is a little more complicated, but has its place, or you can use the Data Analysis Toolpak. For this course, you will find it useful to use the SLOPE and INTERCEPT functions]

Now we can complete the data table. Put a title in D6 that says “Regression”

The line we want to plot is TL = (1.07 * SL) + 3.85 or, more precisely, TL = ($B$19*SL) + $B$20

To get these values into the D column, enter ‘=($B$19*B8)+$B$20" into cell D8 and copy it down through D17. [If you don’t know this little trick, try grabbing the bottom right corner of cell D8 and drag it down to copy D8 to the cells below – this can be a real timesaver]. Notice the over-abundance of decimals. Select. Cells D8 through D17, then right-click, select Format Cells, then choose the Number tab and in it the Number property and set the Decimals to 1.

Making the Graph

Plot the same graph you did last class, i.e., TL vs SL. Be sure to get rid of the background shading, grid lines and legend.

Be sure to ADD appropriate titles to the x and y axes.

When you are done, right click on the graph and choose "Select data". Then Add a new series, call it "Regression" and choose B8 to B17 for the x data and D8 to D17 for the y

data.

In general, in science, theoretical relationships (such as regressions) are plotted as continuous lines with no data markers, whereas actual data are plotted as discrete data points (with no connecting lines).

To achieve this, when you are looking at the graph, position the mouse exactly over one of the points on the regression line and right click. Choose “Format Data Series”, and choose the “Line Color” to be “Automatic” and the “Marker Options” to “none”.

Now you should see your data plotted as distinct points (which is correct) and the theoretical relationship plotted as a line with no markers (also correct).

Exercise 1:

1. Do the above to generate the graph showing the data and the regression line. Print the graph out full-page size and write your name on it.

2. Also print out the spreadsheet showing the data and regression calculations. I strongly suggest you use "Print Preview" to be sure what will print. Turn in both the graph and the spreadsheet printout.

-- END