PHYSICS 331 NOTES ON MICROSOFT EXCEL "LINEST" 27 March 2005

A detailed description of how to calculate a least squares fit to a straight line is given in Chapter 6 of Bevington and Robinson, with a summary on p.114 for the case where the y values have different precisions (as in Fig. 6.2) and the x value are precise.

The programs described there are available along with two errata for the book (p.3 and p.12) at

www.mhhe.com/bevington

A much simpler algorithm can be used if all the y values have the same precision (as in Fig. 6.1) and is available through the function "linest" in Microsoft Excel. The following pages are from

www.colby.edu/chemistry/PChem/notes/**linest**.pdf

This is a document dated 20 August 2002, which explains how to get estimates of statistical uncertainties in slope and intercept along with results for least squares fitting using Excel on either Mac or PC.

It is interesting to note that versions of Excel before 2003 actually gave incorrect results when the curve was constrained to go through the origin and/or some data were collinear: "Excel 2002 and earlier versions always return results that are not correct when the intercept argument is set to FALSE…LINEST has been greatly improved for Excel 2003. If you use an earlier version of Excel, verify that predictor columns are not collinear before you use LINEST… Predictor columns (known_x's) are collinear if at least one column… can be expressed as a sum of multiples of others." This is all explained at

http://support.microsoft.com/kb/828533

A detailed description of how to do least squares fitting using explicit formulae in Excel as well as using "linest" is available (but it does not tell the trick for entering an array formula using a Mac) at

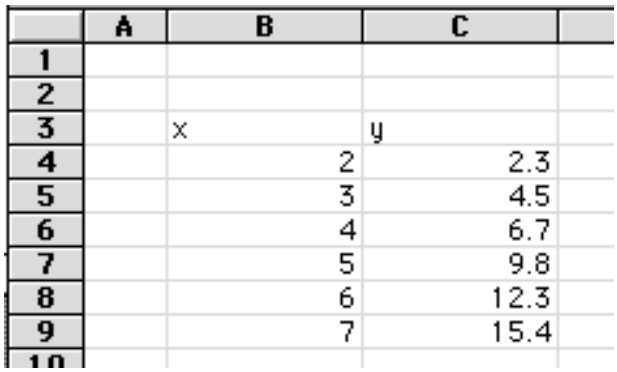
http://dept.physics.upenn.edu/~uglabs/Least-squares-fitting-with-Excel.pdf

LINEST in Excel

The Excel spreadsheet function "linest" is a complete linear least squares curve fitting routine that produces uncertainty estimates for the fit values. There are two ways to access the "linest" functionality; through the function directly and through the "analysis tools" set of macros. These instructions cover using "linest" as a spreadsheet function.

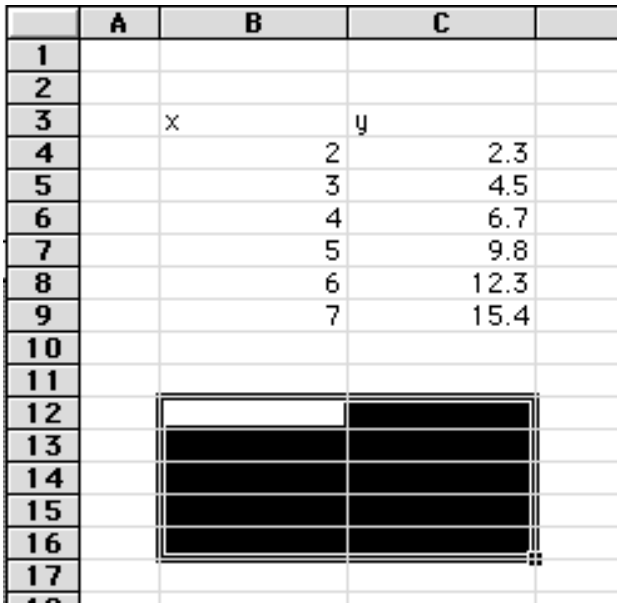
Using linest in your spreadsheets is very easy, after you master the concept of an array function. Array functions are functions that while entered into a single spreadsheet cell produce results that fill several cells. The steps outlined below take you set-by-step through the process of linear curve fitting.

Step 1. Type in your data in two columns, one for the x variables and one for the y. You can use any labels you would like. "x" and "y" are used in the example at right for convenience.



	A	B	C
1			
2			
3		x	y
4		2	2.3
5		3	4.5
6		4	6.7
7		5	9.8
8		6	12.3
9		7	15.4
10			

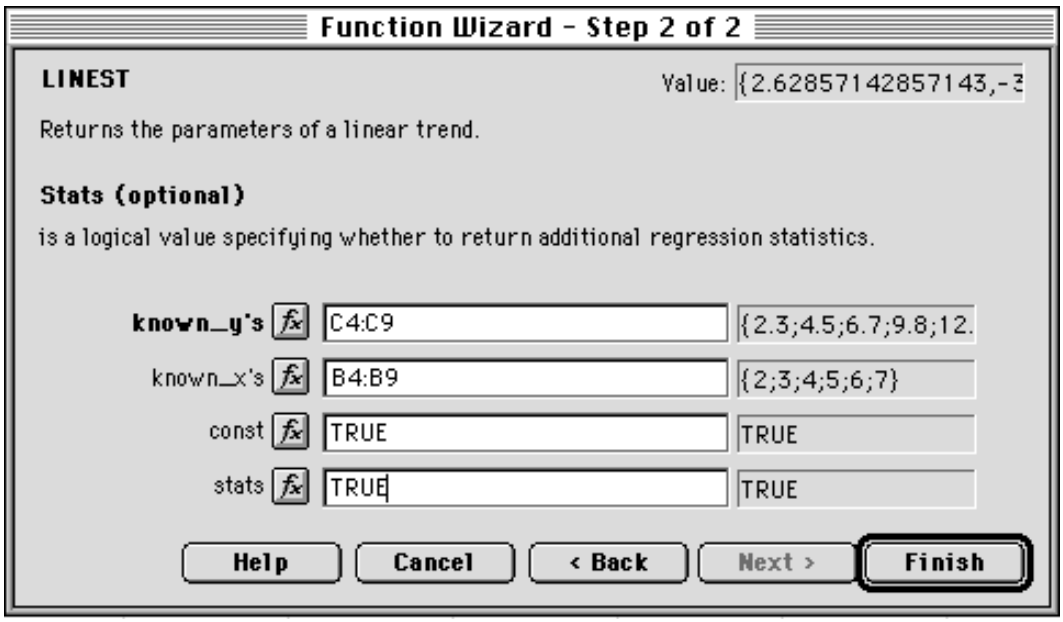
Step 2. Select the area that will hold the output of the array formula. For "linest" you should drag to form a 5 row by 2 column data array.



	A	B	C
1			
2			
3		x	y
4		2	2.3
5		3	4.5
6		4	6.7
7		5	9.8
8		6	12.3
9		7	15.4
10			
11			
12			
13			
14			
15			
16			
17			

Step 3. Click in the formula bar at the top of the screen. Now press the function wizard button. This button is in the formula bar and is labeled "fx". A two-part scroll box will appear; in the left scroll box click on "Statistical" and in the right click on "LINEST". Next click on "Next>." The window shown below will appear. On your spreadsheet select the cells containing the y values by

dragging in the original spreadsheet using the mouse. Click in the "known_x's" dialog box, and select the cells containg the x values. Type in "TRUE" in the last two dialog boxes. The first TRUE indicates that you wish the line to be in the form y=mx+b with a non-zero intercept. The second TRUE specifies that you wish the error estimates to be listed. The Function Wizard dialog box should then appear as below.



Step 4. Click on "Finish." The formula bar should then appear as below, although your y and x cell ranges may be different, of course. If the values are incorrect, you can edit them as you would normally.

Step 5. **Now here is the important step**. LINEST is an array function, which means that when you enter the formula in one cell, multiple cells will be used for the output of the function. To specify that LINEST is an array function do the following. Highlight the entire formula, including the "=" sign, as shown above. On the Macintosh, next, hold down the “apple” key and press "return." On the PC hold down the “Ctrl” and “Shift” keys and press “Enter.” Excel adds "{ }" brackets around the formula, to show that it is an array. Note that you cannot type in the "{ }" characters yourself; if you do Excel will treat the cell contents as characters and not a formula. Highlighting the full formula and typing the “apple” key or “Ctrl”+”Shift” and "return" is the only way to enter an array formula.

The least squares results should be printed as shown below. The labels in the first and last column aren't provided by the LINEST function. We've added them to show the meaning of each cell. For example, the slope is 2.629±0.085 and the intercept is -3.33 ± 0.41.

	x	y	
	2	2.3	
	3	4.5	
	4	6.7	
	5	9.8	
	6	12.3	
	7	15.4	
slope	2.62857143	-3.3285714	intercept
±	0.084997	0.40910554	±
r^2	0.995835	0.35556796	s(y)
F	956.384181	4	degrees of freedom
regression ss	120.914286	0.50571429	residual ss

Step 6: You should now evaluate the model that you have built. The r^2 value is often used for this purpose, but it is only a rough indicator of the goodness of fit. The r^2 value is calculated from the total sum of squares, which is the sum of the squared deviations of the original data from the mean:

$$\text{total ss} = \sum_{i=1}^{n}(y_i - y_{av})^2)$$

and the regression sum of squares, which is the sum of the squared deviations of the fit values from the mean:

$$\text{regression ss} = \sum_{i=1}^{n}(\hat{y}_i - y_{av})^2$$

Giving: $$r^2 = \frac{\text{regression ss}}{\text{total ss}} = \frac{\sum(\hat{y} - y_{av})^2}{\sum(y_i - y_{av})^2}$$

Values close to one are good. The uncertainties in the slope and intercept are much better for judging the quality of the fit. In the example the uncertainty in the slope is 0.085/2.629*100 = 3% and the uncertainty in the intercept is 12%, which is only about two significant figures in each. The uncertainties in the slope and intercept are not as good as the r^2 of 0.996 might have indicated! An even better statistical test of the goodness of fit is to use the Fisher F-statistic. The F-statistic is the ratio of the variance in the data explained by the linear model divided by the variance unexplained by the model. The F-statistic is calculated from the regression sum of squares and the residual sum of squares. The residual sum of squares is the sum of the squared residuals:

$$\text{residual ss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} r_i^2$$

Dividing by the degrees of freedom, gives the variance of the y values

$$s_y^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

The regression sum of squares, the residual sum of squares, and the standard deviation of the y values, s(y) are all listed in the linest output. The F-statistic is then the ratio of the variances:

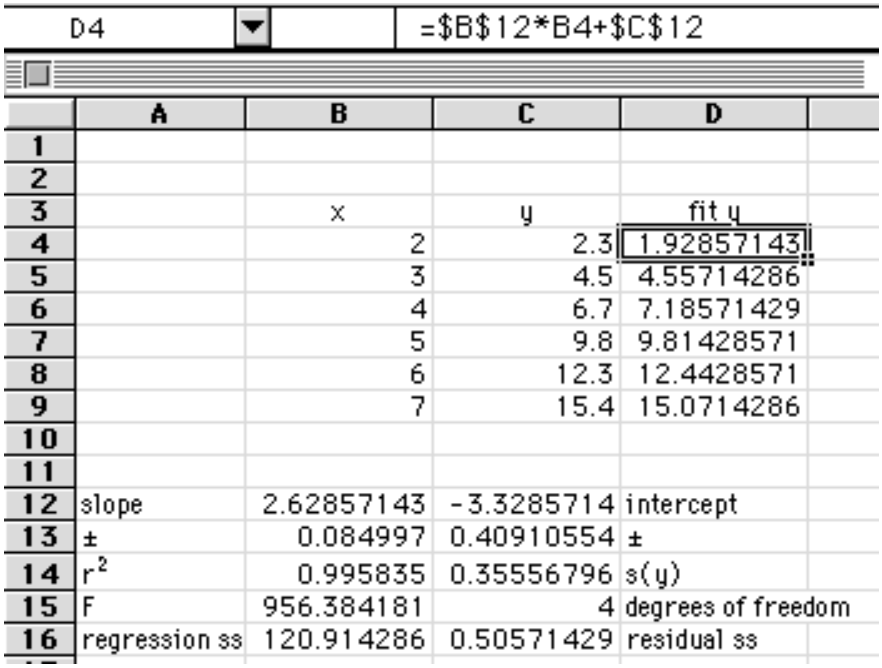
$$F = \frac{\text{variance explained}}{\text{variance unexplained}} = \frac{\text{regression ss}/v_1}{\text{residual ss}/v_2} = \frac{\left(\sum(\hat{y}_i - y_{av})^2\right)/v_1}{\left(\sum(y_i - \hat{y}_i)^2\right)/v_2}$$

You use the F-statistic under the null hypothesis that the data is a random scatter of points with zero slope. Critical values of the F statistic are listed in standard statistics texts, the CRC Handbook, and Quantitative Analysis texts. If the F-statistic is greater than the F-critical value, the null hypothesis fails and the linear model is significant. For the degrees of freedom, which are abbreviated in most tables as v_1 and v_2, use $v_1 = 1$ and $v_2 = n - k$, where k is the number of variables in the regression analysis including the intercept and n is the number of data points. The value for v_2 is listed as the degrees of freedom in the linest output. A small part of the F-table is shown at right for an α value of 0.05, that is, 95% confidence. For the example above, $v_1 = 1$ and $v_2 = 6 - 2 = 4$. The F-critical value is 7.71. The F-statistic for our example is 956.38, which is much greater than the F-critical value. You are 95% sure that your data is not a random scatter of points and that the regression is justified.

F-critical values at α=0.05

v_2	F (v_1=1)
1	161.5
2	18.51
3	10.13
4	7.71
5	6.61
6	5.99
7	5.59
8	5.32

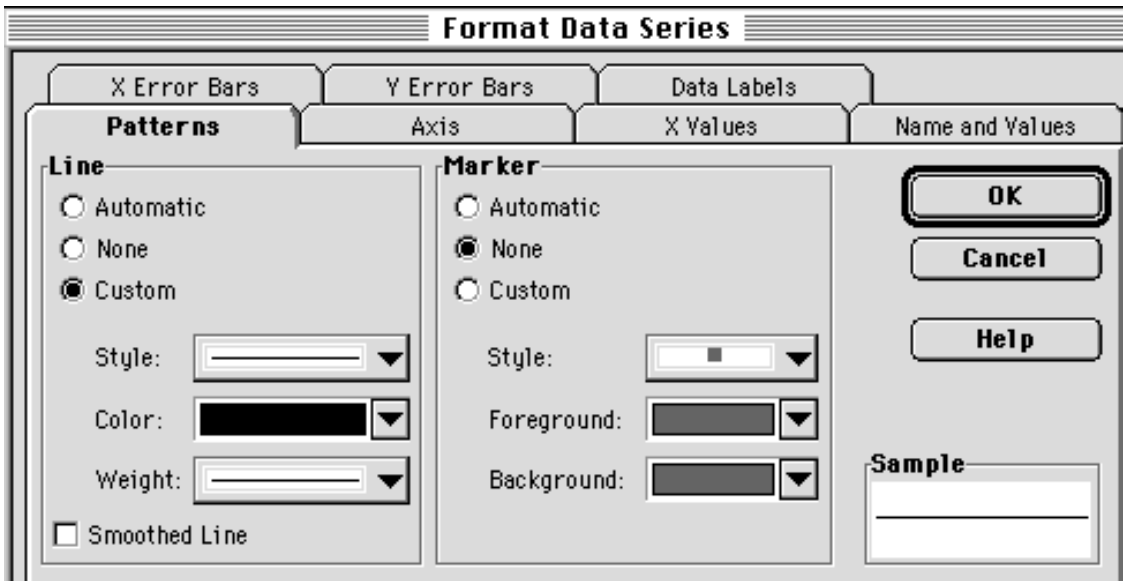
Step 7. You will now need to calculate the fit y values, $\hat{y}_i$, which are the values that lie on the line at the given x values. You can use the TREND array function for this, but it is just as easy to simply calculate the fit y values directly. Start a new column next to the y values. In this new column enter the formula that gives $\hat{y}_i = m x_i + b$, with the slope and intercept from the LINEST output:



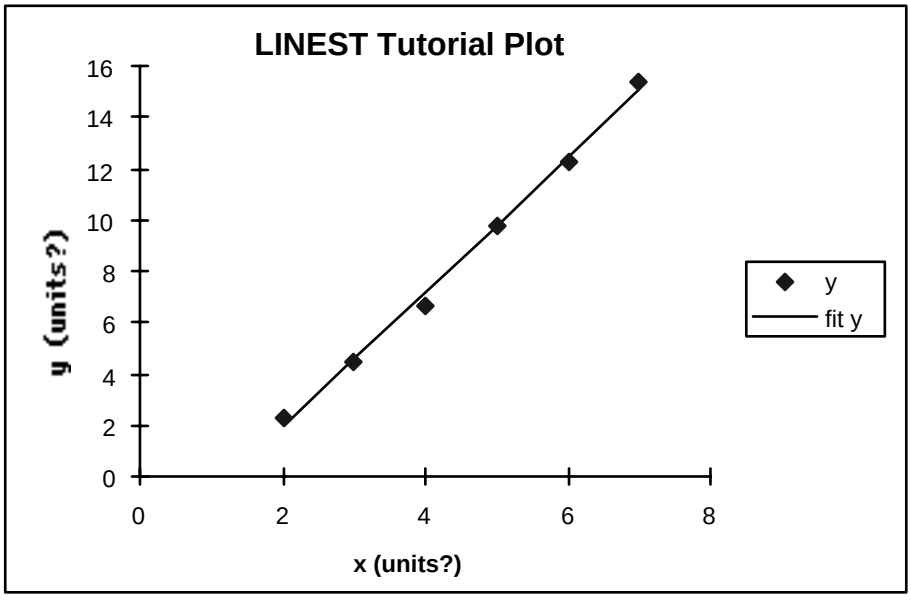
	A	B	C	D
1				
2				
3		x	y	fit y
4		2	2.3	1.92857143
5		3	4.5	4.55714286
6		4	6.7	7.18571429
7		5	9.8	9.81428571
8		6	12.3	12.4428571
9		7	15.4	15.0714286
10				
11				
12	slope	2.62857143	-3.3285714	intercept
13	±	0.084997	0.40910554	±
14	r^2	0.995835	0.35556796	s(y)
15	F	956.384181	4	degrees of freedom
16	regression ss	120.914286	0.50571429	residual ss

Step 8. You can now use the "Chart Wizard" to help graph the results: first select the three columns in your spreadsheet. Include the column labels. Click on the Chart Wizard icon: 
The cursor will change shape indicating that you are to drag on your spreadsheet where you want the plot to appear. Remember for lab reports that charts should be at least half a page.
The Wizard will then take you through setting up your graph. Do a scatter graph, and choose the format that has plot symbols, but <u>not connecting lines</u>.

Step 9. You now need to replace the plotting symbols for the fit y values points with connecting lines. Double click on one of the fit y value data points. The "Format Data Series" dialog box will appear. Change the default settings to no plot symbol (marker) and connecting lines as shown below:

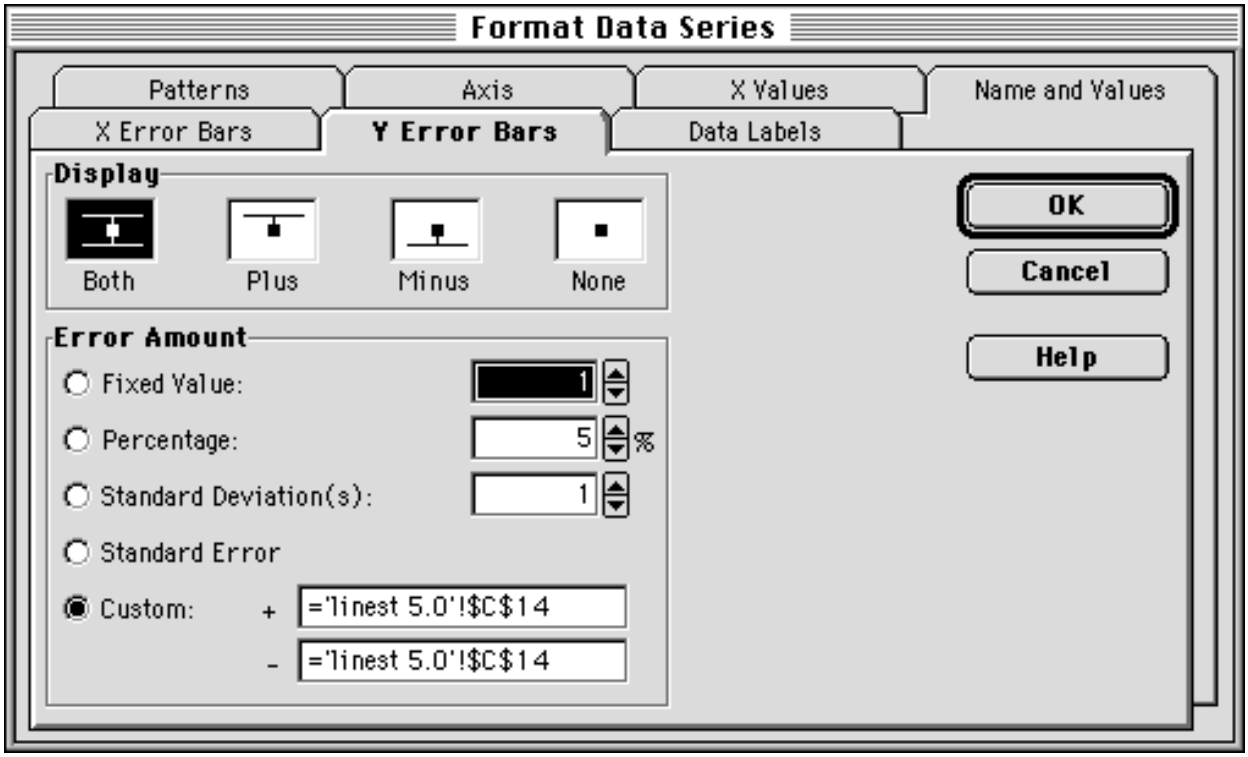


The plot should now appear as at right. Charts and spreadsheet cells can easily be copied and pasted into Word documents.



Adding Error Bars to Plots

After you have your graph displayed you can easily add error bars. Double click on one of the plotting symbols for your data. The dialog box shown below will appear. Click on the "Y Error Bars" tab. Click on the "Both" icon. Next click on the "Custom" button. Next click in the "+" box and then select the cell in your spreadsheet that contains the s(y) value. Repeat this last step in the "-" box. Click on "OK" and the error bars should appear on your plot.



The final chart, in all its glory looks like this: