Chapter 7

Exercise on Graphing and Least Squares Fitting in Excel

7.1 Purpose

The purpose of this experiment is to become familiar with using Excel to produce graphs and analyze graphical data.

7.2 Introduction

While Excel has a lot of useful features for data analysis, etc. it was designed primarily for use by people in business rather than for scientists, and so there are some things which scientists wish to do which require a little effort in Excel. Graphical analysis is one of the areas where this is true, as you shall see. (Even with the extra work, it's a lot more convenient than doing it all by hand.)

7.3 Theory

7.3.1 Graphing

Graph Type

The graph type mainly used by scientists is an **x-y** graph[1]. *Do not* choose a line graph!

Colour

The default grey background in many spreadsheets just looks bad in graphs; it obscures the data and serves no purpose. *Turn the background colour off!*

Gridlines

Grid lines should be either removed or in both dimensions. Gridlines in one direction only look odd on an "xy" graph. *Turn the gridlines off!*

Text

The main text of a graph consists of $x-$ and $y-$ titles, a *main title* and perhaps a *sub–title*. All of these may be set in Excel.

Series

Excel allows you to plot several different "series" of (x, y) data. Each series can be *customized*, with choices for many things, including the following:

- Patterns

  Each series can be plotted with lines, symbols, or both.

Do not connect the points like a dot-to-dot drawing!

[1] As long as there is some *mathematical relationship* between the variables, then an x-y graph illustrates the relationship. However, if the independent variable does not have a numerical value, then this doesn't apply. For instance, if you were graphing reaction time for men and women, then a bar graph would be the logical choice, since there's no *numerical* relationship between "men" and "women".

Do not use an arbitrary function just because it goes through all the data points!

- Markers

  There are many possible symbols which can be used for each series.

- Lines

  There are several line types available for each series.One important fact about how lines are used to connect points in a series; *all points in a series are joined by lines*, unless the line for that series is turned off.

In science, it is almost always wrong to have a dot-to-dot drawing. It is also wrong to have a curve which has no mathematical significance. For this reason, data points should not be connected by either line segments or a curve like a polynomial which is made to pass through each data point. The only line or curve which should be shown is the result of a fit which is based on some theoretical mathematical relationship.

Matching up x and y Values

When you create an xy graph is created in Excel, you *don't* input data values as (x, y) pairs. Instead you select series for each of x and y. The way the individual x and y values are associated is by where they occur in their respective series.

In Figure 7.1 you can see that the 5^{th} point in each series is highlighted. Even though the series all start in different rows and columns, since the number of cells in each is the same, corresponding values can be considered to be related. (If a cell is blank, then the corresponding point or error bar will not be plotted.)

Error Bars

In Excel, when you choose *custom* error bars, you can choose series for both x and y, and even potentially different series for the $+$ and $-$ directions.

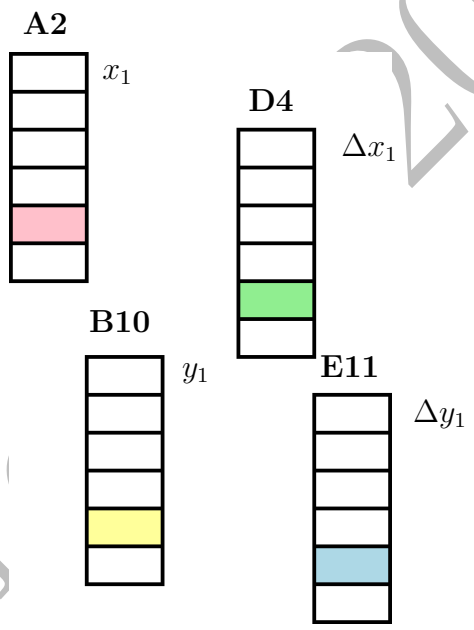

Figure 7.1: Spreadsheet layout with series for x, y, and error bars

You don't necessarily have to put markers on the ends of the error bars; the value in doing so is to make it clear that you're not just using "+" symbols for plotting the data points. Also, if you are including a grid on your graph, error bars without markers on the end may be hard to distinguish. However if the error bars will be clearly identifiable without markers, you don't need to use them.

7.3.2 Least Squares Fitting

The point of plotting a graph in an experiment is usually to extract information from the graph; often the data is plotted in such a manner that the model being tested suggests that the data should fit a straight line. If it does, then getting the slope and y intercept of the line of best fit along with their associated uncertainties is necessary. One of the two usual ways to determine the uncertainty in a graphical quantity is to calculate the **standard error**. (The other involves finding lines of maximum and minimum slope.) The following sections discuss using Excel to do least squares fitting and to calculate standard errors.

Determining the equation of the line by formulas

In Chapter 4, *"Graphs and Graphical Analysis"*, the lab manual explains how to calculate a least squares fit to a set of data. This can be done in Excel by creating additional cells corresponding to each data point which contain, respectively, x^2, y^2 and xy. At the end of the data, these quantities can be totaled to give the sums necessary to do the least squares fit.[2]

[2]You may notice that a particular quantity comes up a lot. It is

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 \tag{7.1}$$

It only takes a couple of lines of algebra to show that this equals

$$N\left(N-1\right){\sigma_x}^2 \tag{7.2}$$

where ${\sigma_x}^2$ is the sample standard deviation of the x values.

Determining the equation of the line using LINEST()

Excel contains a function to do least squares fitting. Unfortunately if produces a bunch of numbers without indicating what is what. It also has to be configured to do things the way we want. Make sure to configure it to give extra statistics.

*When using **LINEST()** to calculate the least squares fit, always set it to calculate the y-intercept, even when you expect it to be zero! This gives you important information about the data.*

Comparing the result given by the least squares fit using your formulas with your regression output should indicate what several of the quantities are.

(If you use **LINEST()** to do least squares fitting for a lab report, quote the quantities given with the names used in the lab manual. The unidentified block of cells given by Excel is not very meaningful.)

Determining uncertainties in the slope and y-intercept

Case I: Maximum and minimum slopes If the error bars are large enough, then the line of best fit should go through all of the error bars. In this case, there will be two data points which determine coordinates for a line of maximum slope which crosses all of the error bars. Consider the case for positive slope:

If we label two points x_1 and x_2, where $x_1 < x_2$, then we can see from Figure 7.2 that the steepest line which touches the error bars for both x_1 and x_2 is the line between $(x_1 + \Delta x_1, y_1 - \Delta y_1)$ and $(x_2 - \Delta x_2, y_2 + \Delta y_2)$. The slope of this line will then be

$$m_{max} = \frac{(y_2 + \Delta y_2) - (y_1 - \Delta y_1)}{(x_2 - \Delta x_2) - (x_1 + \Delta x_1)}$$

and then the y-intercept is given by

$$b_{min} = (y_1 - \Delta y_1) - m_{max}(x_1 + \Delta x_1) = (y_2 + \Delta y_2) - m_{max}(x_2 - \Delta x_2)$$

Similarly the line with the least slope which touches the error bars for both x_1 and x_2 is the line between $(x_1 - \Delta x_1, y_1 + \Delta y_1)$ and $(x_2 + \Delta x_2, y_2 - \Delta y_2)$. The slope of this line will then be

$$m_{min} = \frac{(y_2 - \Delta y_2) - (y_1 + \Delta y_1)}{(x_2 + \Delta x_2) - (x_1 - \Delta x_1)}$$

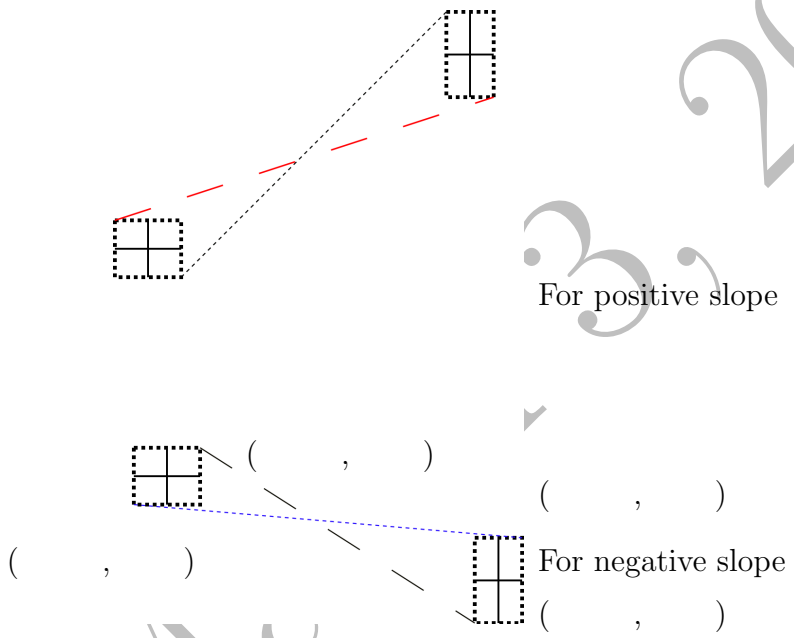

Figure 7.2: Maximum and Minimum Slope Coordinates from a Point

and then the y-intercept is given by

$$b_{max} = (y_1 + \Delta y_1) - m_{min}(x_1 - \Delta x_1) = (y_2 - \Delta y_2) - m_{min}(x_2 + \Delta x_2)$$

The case for a negative slope is shown in Figure 7.2; the analysis is left to the student.

The points for the maximum and minimum slope will not always be the endpoints on the graph.

Index function in Excel To calculate the slope and y-intercept in Excel from a block of data, we can use the *index* function. Its syntax is as follows:

- index(reference, row number, column number)
- reference is the cell range to look in
- row number (starts at one)
- column number (starts at one)

So if we have a data set of 6 values where the x values start in $A2$, and the Δx values start in $D4$, then we can get

$$x_2 + \Delta x_2$$

by the formula

$$= \mathbf{INDEX}(\mathbf{A2} : \mathbf{A7}, \mathbf{2}, \mathbf{1}) + \mathbf{INDEX}(\mathbf{D4} : \mathbf{D9}, \mathbf{2}, \mathbf{1})$$

(Note that the only difference is in which block of data to use.) You'd probably write the formula as

$$= \mathbf{INDEX}(\$\mathbf{A}\$\mathbf{2} : \$\mathbf{A}\$\mathbf{7}, \mathbf{2}, \mathbf{1}) + \mathbf{INDEX}(\$\mathbf{D}\$\mathbf{4} : \$\mathbf{D}\$\mathbf{9}, \mathbf{2}, \mathbf{1})$$

so that you could copy it and still refer to the same blocks of data.

Case II: standard errors If the error bars are small enough, then the points will be scattered in such a way that no line can be drawn which crosses all of the error bars. In this case, the uncertainties in the slope and y-intercept reflect the scatter of the points. In this case, the uncertainty in the slope and y-intercept will be calculated using the **standard errors** in the slope and y-intercept, in much the same way that the uncertainty for an average value is calculated using the *standard error of the mean.*

7.3.3 Displaying Lines

Unless you are going to give the equation of a line or curve, do not *show it on a graph!*

Plotting arbitrary lines

To display a line on the graph, such as a best fit line, one can use a series which has not yet been used. When one knows the *equation* of a line, all one needs is two endpoints so that a line can be drawn between them. To allow this, include 2 values at the end of your x series, x_{min} and x_{max} which are the minimum and maximum values from the x data, respectively. Placing the y values calculated from the line equation in the corresponding cells of another series will allow a line to be plotted between those points. (Set the format for that series to lines only.)

Using "trendline"

There is a built-in feature called *trendline* which allows you to display various fits to data. A *linear* trendline is, in fact, a least squares fit. Unfortunately, this feature does not automatically display the parameters for the fit, so it's not as much use as it could be.

7.4 Procedure

7.4.1 Preparation

You are welcome to go ahead and do as much of the exercise on your own as you wish; you can just bring your spreadsheet with you to the lab and demonstrate the points indicated. If you get it all done in advance, that's great.

If you do it on your own, then print a copy of the spreadsheet showing *formulas* along with the ones indicated in the post-lab questions. (You only need to show the formulas for rows mentioned in the instructions.) Print the graphs as well.

Pre-lab Questions

PQ1: Rewrite the Equations 4.7, 4.8, 4.9, 4.10 for m_{max}, b_{min}, m_{min}, b_{max} for a line with *negative* slope.

Pre-lab Tasks

PT1: Read over Chapter 4, *"Graphs and Graphical Analysis"*, and copy the Equations 4.3 to 4.14 to the appropriate places below. (For instance, copy Equation 4.4 to complete Equation 7.4.) *If you don't want to print the pages of the manual, copy Equations 4.3 to 4.14 to a piece of paper and bring it with you.*

PT2: Fill in the co-ordinates from **PQ1** in Figure 7.2.

PT3: Open the spreadsheet for this exercise from the web page. On the last three tabbed pages, insert rows as needed and type in your linearized data for each of the 3 linearizations. Save the spreadsheet on a floppy, a USB memory stick, or in your Novell account to bring to the lab.

7.4.2 Investigation

In-lab Tasks

In this exercise the in-lab tasks appear throughout the section.

Part 1: Plotting a Graph

Setting up the spreadsheet *Use the data which is already in the spreadsheet. Only change to your own linearized data after the formulas have all been set up correctly.*

1. Load the graph from the lab page.
2. Insert a chart in the box on the first tabbed page.
    - Make sure it is an xy (scatter) graph.

- The x series should be **C6** to **C9**.
- The y series should be **E6** to **E9**.

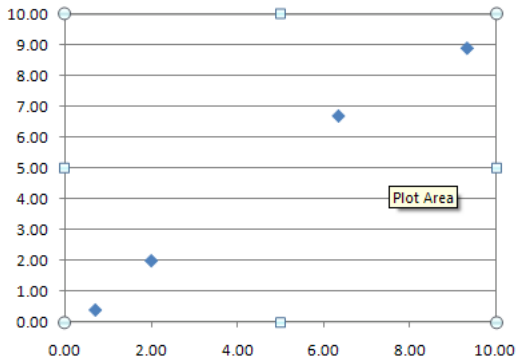

Figure 7.3: Basic graph without error bars

3. Click on the *"Layout"* tab. You should now see the options for error bars.

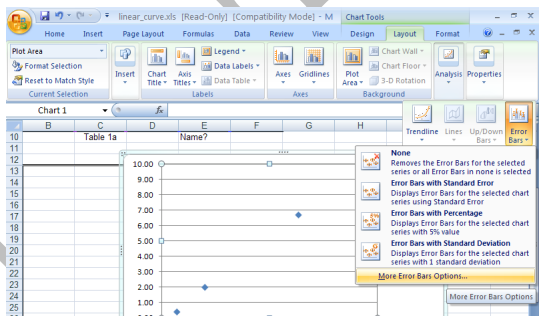

Figure 7.4: Adding vertical error bars

- Note options for x and y error bars.
- Select y error bars, and pick *Custom*.
- Select series **F6** to **F9** for both + and **-**.

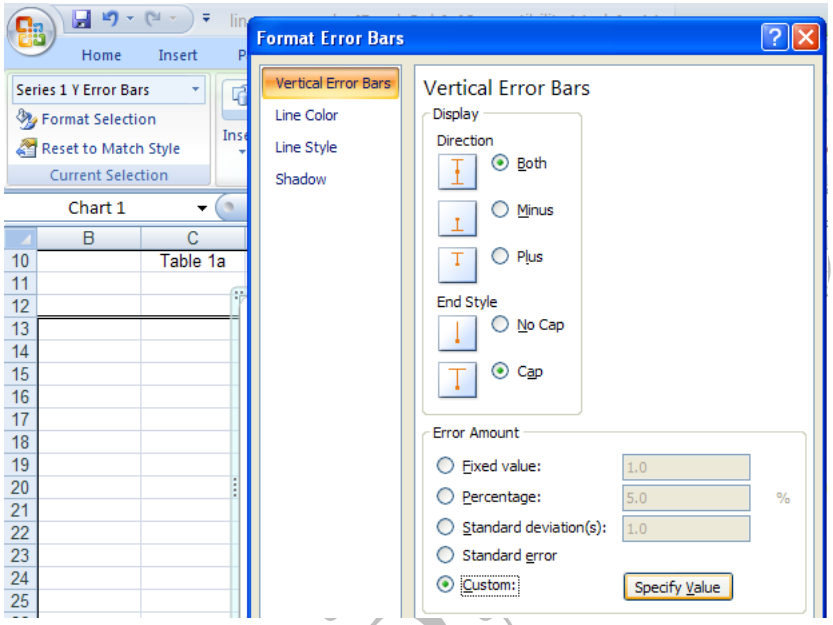

Figure 7.5: Custom vertical error bars

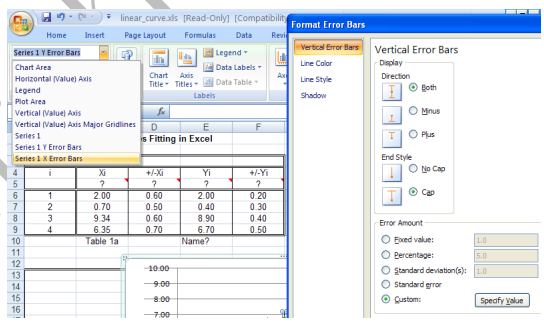

Figure 7.6: Adding horizontal error bars

- Repeat for x error bars, using **D6** to **D9**.

At this point there should *not* be a line connecting the data points. If there is, turn it off.

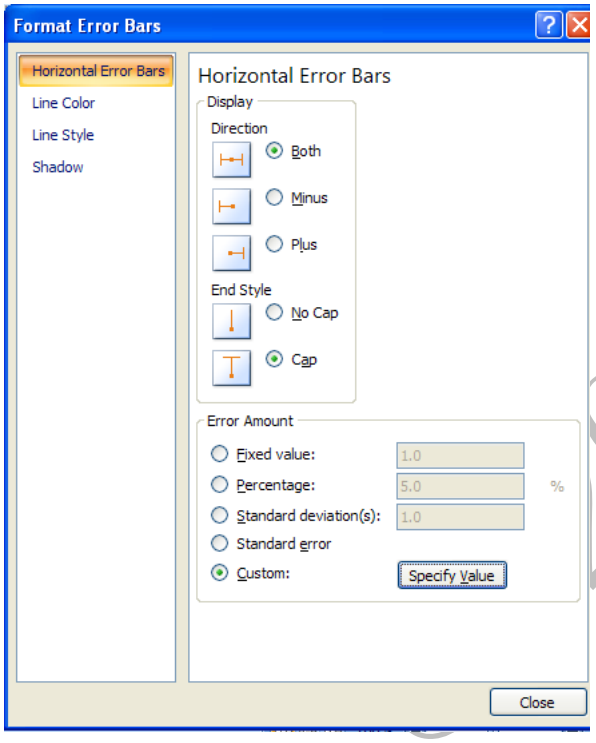

Figure 7.7: Custom horizontal error bars

IT1: Demonstrate graph as is.

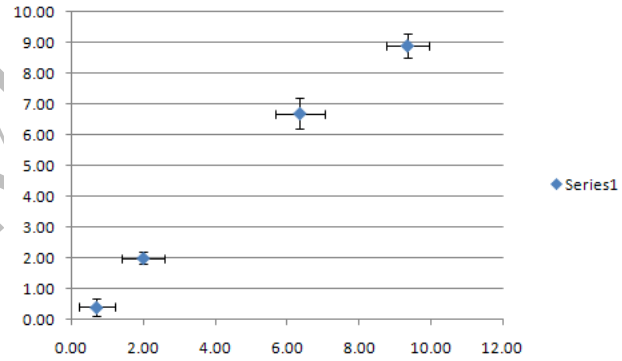

Figure 7.8: Graph with error bars

Part 2: Performing a Least Squares Fit Using Formulas

In this part, you're going to use formulas to do a least squares fit.

Remember that the least squares fit gives values for b, the y-intercept, and m, the slope, as follows:

$$b = \tag{7.3}$$

and

$$m = \tag{7.4}$$

For the linear case, S can be shown to have a value of

$$S = \tag{7.5}$$

$$\sigma = \tag{7.6}$$

where

$$\nu = \tag{7.7}$$

is the number of **degrees of freedom** mentioned earlier. (Often the symbol ν is used for degrees of freedom.)
Also, the **standard error** in the intercept is

$$\sigma_b = \tag{7.8}$$

and the standard error in the slope is

$$\sigma_m = \tag{7.9}$$

R^2 is defined as follows:

$$R^2 = \frac{\left(N\sum x_i y_i - (\sum x_i)(\sum y_i)\right)^2}{(N\sum {x_i}^2 - (\sum x_i)^2)\left(N\sum {y_i}^2 - (\sum y_i)^2\right)}$$

1. Go to the second tabbed page, and replace number cells for everything other than data values with formulas. (Make sure to use a function for N.)

2. Use the values of x_{min} and x_{max} in **C31** and **C32** and the equation of the line of best fit to calculate corresponding y values in **F31** and **F32** Add this series to your graph to show the line of best fit. Right click on one of the points, and format the data series. You want to get a line with no endpoints.

	A	B	C	D	E	F	G	H	I
16								190.0003	
17									
18		Least Squares calculations (by formulae)							
19		Quantity				Value			
20		Slope				0.9936			
21		Y-intercept				-0.0681			
22		Sum of Squares Error (S)				0.36613			
23		Standard Deviation				0.42786			
24		Standard Error in Slope				0.06208			
25		Standard Error in Y-intercept				0.35669			
26		R^2				0.99225			
27			Table 2a	Name?					
28									

Figure 7.9: Least Squares Fit Using Formulas

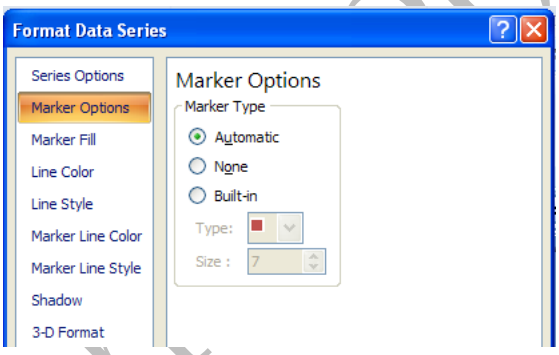

Figure 7.10: Marker options

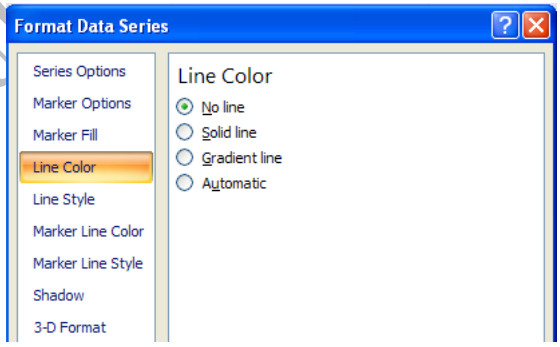

Figure 7.11: Line options

IT2: Demonstrate graph as is.

Part 3: Performing a Least Squares Fit Using Built-in Features

Here you're going to identify the values produced by Excel's **LINEST()** function by comparing them to the values you got from the formulas.

1. Go to the third tabbed page, and use the **LINEST()** function to do a least squares fit.
   - Put the function in **C18**, with *constant*=1 and *stats*=1.
   - Highlight **C18** to **D22**.
   - Press ⟨*F2*⟩ *followed by* ⟨CTRL⟩⟨SHIFT⟩⟨ENTER⟩.
   - Fill in **B18** to **B22** and **E18** to **E22** with names from the previous page. (You don't have to fill in the ones that weren't previously calculated.)[3]
2. On the graph, select data series as before, and choose *linear trendline*. See that it fits on top of best fit line from before, proving it is a least squares fit.

IT3: Demonstrate graph as is. Show that the linear trendline is actually a least squares fit.

Part 4: Finding Maximum and Minimum Slopes

[3]One of the quantities is the number of **degrees of freedom** mentioned earlier. It should be easy to identify. The two "extra" quantities produced are the Regression Sum of Squares given by

$$SSR = \left(\sum (y_i - \bar{y})^2\right) - S$$

which can be shown to be given by

$$SSR = \left(\sum {y_i}^2\right) - N\bar{y}^2 - S$$

and F given by

$$F = \frac{\frac{SSR}{\nu_R}}{\frac{S}{\nu}}$$

which you may find out about in a statistics class when discussing *Analysis of Variance*.

	A	B	C	D	E	F
13						
14						
15			Output from LINEST			
16						
17						
18		?	0.99359843	-0.06806879	?	
19		?	0.06208041	0.35668978	?	
20		?	0.99225289	0.42786005	?	
21		?	256.160768	2	?	
22		?	46.8938716	0.36612844	?	
23						
24						
25			Table 2b	Name?		
26						

Figure 7.12: Least Squares Fit Using LINEST

*Note that on the "Small Scatter" tab of the spreadsheet, the values in cells **D20** to **D23** are **point numbers**, (i.e. point 1 is the first data point, etc.), and so the values in each of those cells must be integers, and they must be between 1 and N, since each one refers to a **data point number**.*

1. Go to fourth tabbed page, and put in reference to third page (ie. page using **LINEST()**) to create meaningful tables.

2. Go to fifth tabbed page and add series for lines of maximum and minimum slope using **E20** to **F21** and **E22** to **F23**. Format the series like the line of best fit to have no endpoints.

3. Change points referenced in **D20** to **D23** to produce lines of maximum and minimum slope which cross all error bars.

4. Look at the cells calculating uncertainty in slope and y intercept and understand how they are calculated using the **index()** function.

5. Put in formulas in **C38**, **C39**, **E38**, **E39**, **F38**, **F39** to get endpoints of lines of maximum and minimum slope which go the full width of the graph. Replace these values for the series above. (**E20** to **F21** and **E22** to **F23**).

IT4: Demonstrate graph as is. Explain the determination of uncertainties in slope and y-intercept this way.

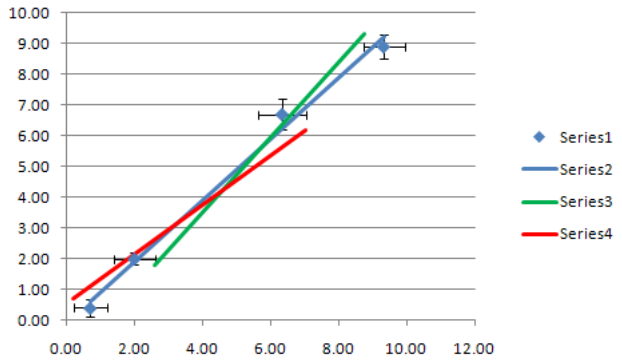

Figure 7.13: Adding lines of maximum and minimum slope

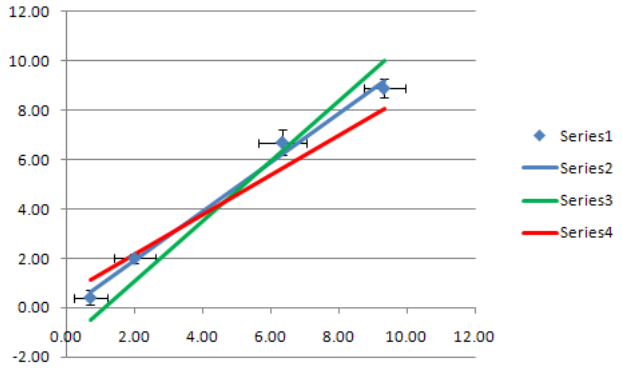

Figure 7.14: Full length lines of maximum and minimum slope

Once this spreadsheet is set up, you would only usually include the information from either the "small scatter" *or the* "large scatter" *tab in a report; all of the other pages are just in order to make the exercise more organized. (Don't just print the spreadsheet page, but use it as a guide to figure out what information to include and how it should be organized.)*

7.4.3 Analysis

1. Insert rows as needed in each page of the spreadsheet after row 6 but before row 9 to allow as many data points as you have in your data

from *"Standing Waves on a String"* . Modify formulas on any page where necessary to work correctly with this change.

2. For each linearization from *"Standing Waves on a String"*, replace the first page data with links to your linearization data, and print the graph and least squares fit results, including uncertainties.

3. Using the equations for frequency and its uncertainty from the linearizing exercise, determine frequency and its uncertainty from each linearization.

4. Discuss the similarities and differences in results of the three linearizations, and decide which one you will use for your lab report. *If there are big differences between the values of R^2 for the different linearizations, you should see a difference in how linear each graph appears to be. Different types of errors in the original data will affect the different linearizations differently, and so if there's a big difference you will have a hint at what errors may exist in your data.*

Post-lab Discussion Questions

Q1: For each of the three linearizations of your data from *"Standing Waves on a String"* , plot the graph, perform the least squares fit, and calculate the frequency and its uncertainty from the results. *Hint: If you save this spreadsheet and make a copy for each linearization you will not have a lot of new calculations to do.*

Q2: Did the frequencies given by the different linearizations agree within experimental uncertainty? Discuss the similarities and differences in results of the three linearizations, and explain which one you will use for your lab report and why.

Q3: Were the standard errors smaller or larger than the ranges given by the lines of maximum and minimum slope? Does that make sense by the definition of "large" and "small" scatter?

7.5 Recap

By the end of this exercise, you should understand the following terms:

- linear graph
- error bars
- least squares fit
- correlation coefficient
- large scatter of data points
- small scatter of data points

In addition, you should, using Excel, be able to:

- plot a linear graph
- add error bars
- perform a least squares fit
- show the least squares fit line on the graph with the data

You should also be able to

- determine whether the points on a graph classify as either "small" or "large" scatter, and calculate graphical uncertainties appropriately in either case;
- compare different linearizations of the same function and to explain why one may be preferred over others.

7.6 Summary

Item	Number	Received	weight (%)
Pre-lab Questions	1	_____	10
In-lab Questions	0	_____	0
Post-lab Questions	3	_____	50
Pre-lab Tasks	3	_____	10
In-lab Tasks	4	_____	30
Post-lab Tasks	0	_____	0