

**371 Lab Rybolt Linear Regression, Excel Directions, and Sample Figure****Linear regression and formulas used**

A line is given by a formula  $y = mx + b$  where  $m$  is the slope and  $b$  is the intercept.

Given real experimental data that may be expected to follow this formula, it is likely that the fit will not be perfect. Therefore, the slope and intercept must be chosen that give the best fit between the experimental data and the equation  $y = mx + b$ . The method of least squares is a type of linear regression that tries to find a line that comes closest to the experimental points. The correlation coefficient,  $r$ , can also be determined and it is a measure of how close the line comes to the data.

The value of  $r$  varies from 0 to 1 with 1 indicating a perfect fit. Often reported as  $r^2$  this can be thought of the percent of  $y$  value variation that is accounted for by  $x$  value variation. For example, if a plot of the boiling point of 10 hydrocarbons versus the number of carbon atoms gave a plot of  $r^2=0.85$  then one could think of the number of carbons as accounting for 85% of the boiling point variation. Notice that this statement is only valid for this specific correlation. A modified data set such as introducing functional groups such as  $-OH$  into the data set could completely change the correlation.

The equations used are:

$$m = \{ \sum xy - (\sum x \sum y)/n \} / \{ \sum x^2 - (\sum x)^2/n \}$$

$$b = \langle y \rangle - m \langle x \rangle$$

$$r^2 = \frac{\{ \sum xy - (\sum x \sum y)/n \}^2}{\{ \sum x^2 - (\sum x)^2/n \} \{ \sum y^2 - (\sum y)^2/n \}}$$

where

$$\langle y \rangle = \sum y / n \quad \text{average of all } y \text{ values}$$

$$\langle x \rangle = \sum x / n \quad \text{average of all } x \text{ values}$$

$n$  = number of pairs of  $x, y$  data

$\sum$  indicates to take a summation or add values together

If doing a linear regression by hand it will be easiest to first determine each of the parts and then put them together. First determine  $\sum x$   $\sum y$   $\langle x \rangle$   $\langle y \rangle$   $\sum xy$   $\sum x^2$   $(\sum x)^2$   $\sum y^2$   $(\sum y)^2$  and then substitute these values in the formulas given above.

For convenience and speed we use programs such as Excel that do linear regression calculations for  $(x,y)$  data sets of any size.

## Older Excel (*Excel 97-2003*) instructions for data plotting and linear regression

### Order of Data

Assume you have two columns of numbers with the x values in the column on the left side and the corresponding y values in a column on the right. This arrangement is the default assumption of Excel—it assumes points are in the order (x,y). That is column of x data to left of y data. However, whenever we refer to a plot of data, we always say “y versus x.” For example, if some instructions call for you to plot the pressure versus the volume then the P values are plotted on the y axis and the V values on the axis. But Excel would assume the spreadsheet containing data points are lined up in columns where V is before P. Confusion on this point causes students lots of troubles that are avoided by just remembering:

**data is recorded as (x,y) points but we say “y versus x”**

### Data in Cells

To practice set up cells in Excel document as indicated below, carry out instructions below and then compare to the sample figure given at end of these directions.

	A	B
1	Time(s)	Temp(K)
2		
3	10.1	8.02
4	22.0	17.23
5	36.2	27.56
6	60.3	46.15
7	70.1	54.90

Format–Cells Category Number and set each column to the appropriate number of decimal places. We always want the last 0 (if it was part of the measured number) to be shown and not left off. You will need to fix number format to indicate the number as it was actually read (that is no fewer and no more significant figures).

You should highlight cells containing data and headings A-1 to A-7 and B-1 to B-7. Select Format–Cells–Alignment–Horizontal–Center. And when data is printed (select File–Print) you should see that column headings and data line up together.

### Plotting Data

Highlight the data cells (but not the headings) to be plotted and then click on the Chart Wizard icon along the top toolbar. New window will appear. Click XY (scatter) and select Chart subtype that shows just the points (no lines).

Click Next and should observe initial plot of data. Click Next.

Enter chart title and make sure you are descriptive and specific (such as “Temperature versus Time for Helium Gas Warming”). Enter Value (X) Axis (such as Time(s)). Enter Value (Y)

axis (such as Temp (K) ). You should try to enter descriptive label of axis and show units in parentheses.

Click Next and Place chart window opens. Select As new sheet. Click Finish and you should observe full graph with title and axes labels.

Click anywhere on background to get color palette. Select Custom Color and then change the background from gray to white. Click OK.

Click on any number along the Y axis and a Format Axis window should open. Change tic marks to inside for major and minor and leave tic mark labels next to axis change tic marks as desired (usually good to have major and minor tic marks). As needed, look at other options for Format Axis and change Scale, Font, Number, and Alignment. Be sure to select the correct number of decimal places under number option. This should always be something you select and not just a default that Excel decides. When done click OK.

Click on any number along x axis and repeat the above procedure.

Observe points on display. Click on data point in box to right of graph and change to the desired style (a circle, for example). Click OK.

### **Do Linear Regression**

At this point there should be data points but no line through data points.

Go to top menu and select Chart–Add Trendline. Under Type, pick linear. Under Options Click display equation and display R-squared value on chart. Click OK.

Observe the linear regression equation placed on graph. Click on this equation and a window opens. Select Number and select an appropriate number of decimal places. You want to avoid having the slope or intercept rounded off in such a way that you lose significant figures. May be easiest to select Scientific and then select number of decimal places to give data with same number of sig fig as your inputted data for graph. In other words, if the data you plotted had 3 sig fig then your slope and intercept are limited to 3 sig fig. If you have an extra sig fig, you can round off later but you do not want to have less sig fig in slope than were in the data that were plotted. Click OK.

### **Saving and Printing Graph**

Select File–Save As to save a copy of graph. Select View–Sized with Window. Review to see how graph fits on the page. Adjust position and desired size. Select File–Print and then print.

### **For Multiple Y groups on same plot:**

Prepare data sheet with numbers for X, Y1, Y2, etc. in columns in this order:

Highlight data.

Click on Chart Wizard Icon.

Pick type of graph (chart type).

It will assume first column is X, next is Y1, next is Y2, etc.

Click on Next and continue as in regular X, Y plot.

## Newer Excel (*Excel 2007*) instructions for data plotting and linear regression

### Order of Data

Assume you have two columns of numbers with the x values in the column on the left side and the corresponding y values in a column on the right. This arrangement is the default assumption of Excel—it assumes points are in the order (x,y). That is column of x data to left of y data. However, whenever we refer to a plot of data, we always say “y versus x.” For example, if some instructions call for you to plot the pressure versus the volume then the P values are plotted on the y axis and the V values on the axis. But Excel would assume the spreadsheet containing data points are lined up in columns where V is before P. Confusion on this point causes students lots of troubles that are avoided by just remembering:

**data is recorded as (x,y) points but we say “y versus x”**

### Data in Cells

To practice set up cells in Excel document as indicated below, carry out instructions below and then compare to the sample figure given at end of these directions.

	A	B
1	Time(s)	Temp(K)
2		
3	10.1	8.02
4	22.0	17.23
5	36.2	27.56
6	60.3	46.15
7	70.1	54.90

Format–Cells Category Number and set each column to the appropriate number of decimal places. We always want the last 0 (if it was part of the measured number) to be shown and not left off. You will need to fix number format to indicate the number as it was actually read (that is no fewer and no more significant figures).

You should highlight cells containing data and headings A-1 to A-7 and B-1 to B-7. Select Format–Cells–Alignment–Horizontal–Center. And when data is printed (select File–Print) you should see that column headings and data line up together.

### Plotting Data

Highlight the data cells (but not the headings) to be plotted, select Insert option (along top toolbar). New options will appear at top of window. Click Scatter from among chart options and then select the Chart icon that shows just points (no lines).

You should now observe an initial plot of the data. Select Design option (top menu bar). Select from among Chart Layout icons the option (may be first among choices) that includes chart title and axes labels. Click on “chart title” and type a descriptive and specific title (such as “Temperature versus Time for Helium Gas Warming”). Click on X “axis title” and enter label

(such as Time(s)). Click on Y “axis title” and enter label (such as Temp (K) ). You should always use a descriptive or symbolic label of axis and show units in parentheses.

Select Design (top menu bar) and then select Move Chart and select New Sheet option and then click OK. Select Page Layout and select orientation in Portrait form and then rescale so chart fills about  $\frac{1}{2}$  to  $\frac{2}{3}$  of the page. We want each graph on a separate page by itself. You should see a page with graph with title and axes labels. Background of plot should be white.

Select Layout then select Axes and Primary Horizontal. Next pick More primary Horizontal Axis Options. Notice you now have control over maximum, minimum, major units, tic marks, etc. among Axis Options. You may want to modify range of axis or add tic marks which help to see where values are along axes.

### **Do Linear Regression**

At this point there should be data points and suitable figure but no line through data points. Go to top menu and select Layout and then select Trendline and More Trendline Options. Under Options click Trend type, click Linear, and then click Display Equation and Display R-squared. Click close.

Click on line equation and  $r^2$  value on the graph and box should appear around values. Then select from top menu Format then Number (from new window) and then set to appropriate number of decimal places. Then select Close. You want to avoid having the slope or intercept rounded off in such a way that you lose significant figures. May be easiest to select Scientific and then select number of decimal places to give data with same number of sig fig as your inputted data for graph. In other words, if the data you plotted had 3 sig fig then your slope and intercept are limited to 3 sig fig. If you have an extra sig fig, you can round off later but do not want to have less sig fig in slope than were in the data that were plotted.

### **Saving and Printing Graph**

Select Office button on upper left and select “Save As” May want to save as Excel 97-2003 workbook file. This choice is compatible with earlier xls files. Otherwise if you use an older version of Excel on another computer you will not be able to open new.xlsx Excel file. Newer Excel software can open older file format (.xls) but older Excel software cannot open newer (.xlsx) files. Select Office Button on upper left and select “Print.”

The Mac2008 version of Excel is similar to above but layout appearance is slightly different. Need to have highlighted points by clicking on a data point on plot to be able to generate linear or other regression lines. Need to open toolbox icon to change axes and titles

### **For Multiple Y groups on same plot:**

Prepare data sheet with numbers for X, Y1, Y2, etc. in columns in this order:

Highlight data.

Click on Insert.

Pick type of graph from icons: Scatter (points only).

Plot will appear and it will assume first column is X, next is Y1, next is Y2, etc.

Continue and modify as before in regular X, Y plot.

