Appendix 3 - Using A Spreadsheet for Data Analysis

Linear Regression - an Overview

Scientists often choose to seek linear relationships, because they are easiest to understand and to analyze. But, even when data is supposed to be linear, data points seldom form a perfect straight line, so linear regression analysis is used to determine the best straight line that fits your data, i.e. the computer calculates the formula of the straight line that best fits your data. The general formula for a straight line is:

\[ y = ax + b \]

where

- \( a \) = the slope of the line
- \( b \) = the y intercept, where the line crosses the y-axis

The x values are called the independent variables and the y values are called the dependent variables. The value for x is somewhat arbitrarily chosen (independent), and the value for y then "depends" on that value. Once \( a \) and \( b \) have been determined, the value of y can be predicted for any chosen value of x.

When a theoretical model is in the form of a straight line, the slope and intercept provide theoretical information about your data. For example, the formula for density is \( D = m/V \), this rearranges to \( m = DV \) which is in the standard form for a straight line, \( y = mx + b \). When mass is plotted on the y axis (Y=mass) and volume on the x axis (x=volume), the formula says the slope should be the Density (\( a=D \)) and the intercept should be zero (\( b=0 \), no term in the density equation).

- The slope, \( a \), is the change in y divided by the change in x (rise/run), so its units are the units of y divided by the units of x.
- The intercept, \( b \), is the value for y when x equals zero, so \( b \) has the same units as y.
- For example, if you were to plot the masses of some set of objects on the y-axis and their corresponding volumes on the x-axis, the result would be a straight line with a positive slope. What is the unit/label on mass (y)? What is the unit/label on volume (x)? The unit/label on \( a \) will be mass divided by volume, what do you get? Note that this is the unit/label for density. Therefore, the slope of the straight line obtained by plotting the mass of a sample vs its volume is the density. The y intercept, \( b \), has the same unit/label as y, so what is the unit/label on \( b \)?
- Remember, not all data have a linear relationship, so linear regression is not appropriate for all data sets. Look at the graph first to see if the data is truly linear before you use linear regression. Data points "out in left field" should be double checked for accuracy (data entry, measurements) or omitted from regression analysis. Above all, don't force a linear relationship on data that is obviously not linear (i.e. curved).

Spreadsheet Instructions

This section leads you step-by-step through the process of entering data, creating and copying formulas, creating a graph, inserting a trendline line (obtained by linear regression), and printing the spreadsheet and graph.

When you create a spreadsheet for a lab report or assignment, make sure you document it properly. It should contain enough information so someone could identify it and figure out how you did your calculations from a printed copy of the spreadsheet. In addition it should conform to the format expected for presentation of scientific data. So, make sure the following information is provided in every spreadsheet you turn in:

1. Your name and partner’s names.
2. Date.
3. Title for the experiment or assignment.
4. Label all data columns and other quantities calculated.
5. Show sample formulas for all calculations above the appropriate column or in a neighboring cell.
6. Include units/labels and display the proper number of significant figures on the “target” values that you will report in your conclusion, i.e. final results only, not intermediate calculations or data.
7. If you know the "true" value, determine the percent error.
8. Each graph must have a title, axis titles with units on both axes (the units are often enclosed in parentheses or separated from the title by a slash “/”), and a legend if two or more series of data are plotted.
9. The scale of each axis should be marked with ticks and values (default setting in Excel).
10. Always follow the graphing convention for scientific data: data points as **markers only** and regression line as a **line without markers**.

**Initial Data Entry**

- Enter partner names, date, and an experiment title.
- Enter your data in columns including labels at the top of each column. Excel expects the x values to be in the first column and y values in the second column. To adjust the column widths, move your cursor to the grey bar with the column letters and drag the right edge of the column until it fits your label. Note that spreadsheets ignore significant figures. For this class we will only worry about displaying significant figures correctly on the final values calculated.
- Save files to your J: drive or personal drive (floppy, flash, etc.). Don’t save to the local hard drive in the computer labs, those drives get cleaned often so your files may just disappear. Make sure to save often when constructing a spreadsheet, so you can recover easily from crashes or wrong mouse clicks.

**Entering Formulas For Calculations**

- Every cell formula must begin with a mathematical symbol (=, +, -, /, *, $, @, etc.) for Excel to interpret it as a formula instead of text.
- A cell address, the column letter followed by the row number (like xy coordinates), is used like a variable in formulas.
- When a formula is copied to another cell, the cell addresses are adjusted **relative to the new location**. For example if the formula +A2 is typed in cell B2 and then copied to cell B3, the formula in B3 will be +A3.
- To keep the cell address the **same** (not adjusted) when it is copied, use $’s in the cell designations, like $A$2. For example, if the formula $A$2 is typed in cell B2 and then copied to cell B3, the formula in B3 will be $A$2, it won’t change. Rows and columns are treated independently, so both $’s are needed to fix the address to a given cell. In contrast, when $A$2 is copied, the column designation (A) will be locked, but the row designation (2) will be adjusted relative to the new location.
- There are also a large number of mathematical functions available. These formulas may be used alone or embedded in a formula. Excel’s defined functions may be typed directly into the cell (if you know what it is) or can be inserted into the cell via the Insert/Functions menus.

**Graphing The Data**

- Highlight your data columns (numbers only, not the labels) and click on the **Chart Wizard** icon on the toolbar or choose Insert/Chart from the menus. The **Chart Wizard** will lead you through four steps to create a graph of your data. This tutorial will show you all the options, but when you create graphs for future assignments, you need to select only those options you wish to change and skip the others (accept the defaults).
Step 1
Select:
Chart Type = XY Scatter
Chart Subtype = markers only (the default choice).
Click Next.

Step 2 - Series tab
Make sure the x values are the diameters and the y values are the circumferences.
In the Name box type Data (this will become the label in the legend box)
Click Next.

Step 3 - Titles tab
Enter a title for the graph
For x axis enter Diameter (cm)
For y axis enter Circumference (cm)

Step 3 - Axes tab
Experiment with the options and note the result on the adjacent graph.
Reselect the defaults (shown above) which display numbers and tics on each axis.

Step 3 - Grid Lines tab
Select your preference for display of gridlines on the graph.
Defaults are OK.

Step 3 - Legend tab
Make sure Show Legend is selected (default).
Select desired position for the legend.
(We will add a trendline in a bit so a legend will be necessary).

Step 3 - Data Labels tab
Experiment with the options and note the result on the adjacent graph.
Make sure None is selected (default).
Click Next.

Step 4
Select Graph as Object and select the sheet where your data resides as the destination.
Click Finish.
Adjusting the Layout for Printing

- You should now see a graph displayed in the spreadsheet similar to that shown in Fig. 2. Adjust the size and location so the spreadsheet and graph fill the page when printed.
  - Make sure you have deselected the graph (click outside the graph), and then select View/Page Break View. (The Page Break View and the Normal View options will not show up in the View menu if the graph is selected.) In this view you see the whole page, but you can still work in the spreadsheet. Check to see that your spreadsheet info lie completely on the page or fit neatly on the pages they cover. If not, cut and paste until they are.
  - Next click on the graph to select it, and drag it to an open space below your data. Then stretch and position it to fill the rest of the page (similar to Fig. 3). Deselect the graph and switch back to Normal View (in the View menu). Notice in Normal view there is now a dotted line displayed along the left edge which will be the edge of the page when printed. Save your spreadsheet.

Add a Trendline

If your graph shows that the data are in a roughly linear relationship; you can use linear regression to calculate the equation of the straight line that best fits your data.

So, the next step is to extract the regression equation and plot the regression line (Excel calls it a trendline).

- Right click on one of the data points in the graph and select Add Trendline ...

<table>
<thead>
<tr>
<th>Trendline - Type</th>
<th>Trendline - Options</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Trendline - Type" /></td>
<td><img src="image2" alt="Trendline - Options" /></td>
</tr>
</tbody>
</table>

Select linear (default). Choose the data set to apply the trendline to (only one choice here).

Name is the title in legend box. Forecast extends the line in the graph forward and backward beyond the data.

Do not select Set intercept = 0

Select Display equation on chart. Display R² on chart - your option. R² is a “goodness of fit” measure, R²=1 is “exact fit” and R²=0 is “no fit.”

Click OK.

Page Break View of the graph as it should appear now.

Notice that the data is plotted as markers only, and the regression line is plotted as line only. This is standard format for scientific graphs; make sure all your graphs adhere to this convention.

- Edit the graph by right clicking on a component of the graph (the grid, a line, an axis, an axis label, the title, the legend box) and opening the Format menu. For the Chart Area and the Plot Area, notice that, in addition to the Format menu for the item chosen, you also have access to Chart Type, Source Data, Chart Options, and Location - the four steps of the Chart Wizard as well as cut and paste options.
Documenting Information from the Regression Equation

- Document in your spreadsheet what you expect to learn from the slope and intercept of the regression equation. For example, the formula for density is \( D = \frac{m}{V} \), this rearranges to \( m = DV \) which is in the standard form for a straight line, \( y = mx + b \). When mass is plotted on the y axis and volume on the x axis, the formula says the slope should be the Density and the intercept should be zero (no term in the equation). So, in your spreadsheet you would enter:

\[
\begin{align*}
\text{slope} &= \text{density} = \text{value} \\
\text{intercept} &= \text{zero} = \text{value}
\end{align*}
\]

- Significant figures for slope and intercept: Look at the average number of significant figures for the x and y data; the lowest average number of significant figures sets the significant figures for your slope and intercept. Adjust the cell format the values for slope and intercept so they display the correct significant figures.
  - Here we recognize slope and intercept were determined from our measurements for the x and y data and regression averages those values in the calculation.
  - Another way to determine significant figures is to do a more detailed regression analysis and look at the standard deviations for the slope and intercept, but we won’t go into that here; it will be covered in later chemistry and physics courses.

- Enter the units/label for the slope and intercept values.

- When you know the accepted or true value, calculate the percent error, and then label and document the calculation. The formula for percent error is:

\[
\text{percent error} = \frac{|\text{actual value} - \text{your value}|}{\text{actual value}} \times 100
\]

The vertical lines in the numerator indicate absolute value, i.e., the difference is always positive, which is accomplished with the ABS function.
  - Enter “% error=” into a cell. In the adjacent cell of the next column, enter the formula to calculate the percent error. Then add the unit/label “%” in the next column.
  - Change the cell format to display the proper number of significant figures.
  - Document the formula used. The easiest way to do this is to combine copy and past with manipulation of the way Excel displays cell contents. First move the cursor to the cell with the formula and then enter “edit mode” - either press F2 (Fig. 4) or click in the status line where the formula is displayed (Fig. 5). Then delete the = sign at the beginning of the formula and get out of edit mode (hit return or click outside the cell). Now the formula will be displayed as text. Copy this text version of the formula to cell just above or below. Then replace the = sign as the first character in the original cell. Note how you have documented your calculation, specified the number of significant digits for your answer and included the units for that answer. Make this a habit for all spreadsheet exercises you turn in. Save your spreadsheet.

- A completed example of the spreadsheet expected for this class is shown on the next page.

Printing the Spreadsheet and the Graph

- In File/Page Setup select Portrait or Landscape, set margins, and set the print scaling to get the best fit on the page.

- Use View/Page Break View to arrange things to print nicely.

- Use the File/Print Preview in the File menu to see how it will look.

- File/Print or Control p to print the spreadsheet and inserted graph. If you select the graph and then print, only the graph prints.
Final Report

- Each group turn in the printout of the spreadsheet and the graph.
- Make sure each member of the group signs the printout as an assertion that this was a group effort.
- Several labs this semester will require you to use Excel to analyze the data. At those times refer back to this exercise to refresh your memory about the procedures you need. Always remember to completely document your spreadsheet - identify the values and equations used.

Example of spreadsheet for this class

<table>
<thead>
<tr>
<th>Diameter (cm)</th>
<th>Circumference (cm)</th>
<th>DEW</th>
<th>06/30/2004</th>
<th>Detn of Pi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.46</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.31</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.94</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.55</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.01</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.36</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.83</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.9</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data

Linear (Data)

Detn of Pi

\[
y = 3.1832x - 0.3061\]

\[
R^2 = 0.9974\]