Data analysis 5
Week 11 practical

This practical assesses how to perform correlation and linear regression using Excel or MATLAB.
You may open the Blackboard assessment and work through questions 1-7. Following which you may wish to look at the explanations below (either Excel or MATLAB).

5.1 Fitting a power law

5.1.1 Excel method

When fitting a curve to data in Excel the trickiest thing can be pre-processing your data so you remove the bad values. In this example you will fit a power law of the form $y = ax^b$, where a and b are constants.

1. Download the `AustralianCirrusClouds.csv` file from Blackboard and save somewhere on your computer. Open it up in Excel by clicking on it.
2. You'll see that the file contains 3 columns of data: time in decimal days; Extinction from the lidar; and Ice water content from the *in-situ* probe.
3. Next Insert a column between columns `A` and `B` and call it time. Excel time format is the number of decimal days from the year 1900, whereas the spreadsheet has number of decimal days from 0 A.D. Set column `B` equal to the first row minus `693960` (which is the number of days between 0 AD and the year 1900). Repeat this for all rows.
4. Set the formatting to date and time format—highlight all cells in column `B` and right click on them, and select `Format cells`. Select 'custom' then in the box on the right scroll down to select `dd/mm/yyyy hh:mm:ss`. Press OK.
5. Plot both datasets on the same time vs y scatter graph (i.e. time on the x-axis and both extinction and ice water content on the y-axis) to inspect for any possible correlations.
6. As discussed we will be fitting a power law to this data. This means a straight line of log Ext (log of extinction) against log IWC (log of the ice water content). In two adjacent columns calculate the log of the *Extinction* and the *Ice water content*. You'll notice that there are some errors as you will have taken the log of zero, which is undefined. These will need to be removed from the data, before fitting a curve.
7. Remove the bad data by setting every value that is *not-a-number* to something that can be filtered following the proceedure below:
   (a) In an adjacent column, e.g. column I, type:
   `=IF(NOT(ISNUMBER(F2)),"",F2)`

can you see what this does?. In this example the cell in `F2` should be the log of the Extinction and the cell in `G2` should be the log of the Ice water content. Copy this down for all Extinction and Ice water contents (using relative references).

(b) It is desireable to get rid of the blank gaps in the list of numbers. Do this by first selecting both of the two columns just created right clicking on the cells and selecting `copy` and then right click on another cell and select `paste special->values`. Secondly select the new columns of data (which also contain blanks) and press `F5`. A pop-up box will open; click on `Special` then select the `constants` radiobutton from the list and un-check all of the check boxes below the radio button, except `Text`. Press OK, the blanks should be selected. Right click on the selection and delete all of the selected cells. The blank cells are now deleted!

8. Calculate the correlation coefficient, r, using `=CORREL(array1,array2)` where `array1` is the x-values and `array2` is the y-values.
9. Now we wish to fit a straight line to the two columns of data. You can either do this using the `LINEST` function or the `Regression` tool in the *data analysis toolbox*. The latter is straight forward—just go to `data analysis` and select `regression`, while the former is a little trickier as it is an array function.
   (a) To use `LINEST` select two empty adjacent cells (in the horizontal direction) and type `=LINEST(Array1,Array2)`, where `Array1` are your y-values in the list and `Array2` are your x-values in the list.
   (b) Press `ctrl+shift+enter` and you should get two numbers in the cells, which are the fit coefficients for a straight line—e.g. $y = b_1 x + b_0$.
   (c) You are fitting $y = ax^b$ or rather $\log y = \log a + b \log x$. Therefore in order to calculate a you should take the anti-log (i.e. 10^{b_0}) of the b_0 fit coefficient (do you see why?). b is just equal to b_1 in this example.
10. Once you have the fit coefficients try plotting the *actual* Extinction against Ice water content and then plot Extinction against *parameterised* ice water content (i.e. against $aExt^b$). If you've got this correct you should find that the parameterised form agrees well with the data!

5.1.2 MATLAB method

When fitting a curve in MATLAB the trickiest thing is probably getting your data into MATLAB.

1. Download the `AustralianCirrusClouds.csv` file from Blackboard and save somewhere on your computer. Read into MATLAB by changing the MATLAB directory to the location of the file on your computer and reading in using:

```
1 dat=csvread('AustralianCirrusClouds.csv',1,0);
```

The 1 means to skip the text in the first row and the 0 means do not skip any columns. You'll notice that `dat` is a table / matrix of size 1257 by 3 (e.g. `size(dat);`)

2. Calculate the log of the two columns (Extinction and Ice Water content) by doing:

```
dat2=log10(dat(:,2:3));
```

3. Remove the bad values (because you've taken log of a zero) by searching for the indices of the good values and referencing the `dat2` matrix with those indices. Do this as follows:

```
index=find((~isinf(dat2(:,1)) & ~isinf(dat2(:,2))) ...
      & (~isnan(dat2(:,1)) & ~isnan(dat2(:,2))));
```

The fragment of code above means find all of the elements in `dat2(:,1)` and `dat2(:,2)` that are not infinity or *not-a-number* (i.e. all the real numbers!). The MATLAB code $\sim$ means *not*. If you were to exclude it, index would contain all of the values that are either infinity or *not-a-number*. The `&` means *and* in the expression above.

4. Secondly, once you've found the indices that you want to keep, create a new array / table by using `index` to reference `dat2`:

```
dat3=dat2(index,:);
```

`dat3` should now contain all the data, with blanks removed.

5. Now calculate the correlation coefficient using:

```
corrcoef(dat3(:,1),dat3(:,2))
```

The elements that aren't unity are the correlation coefficient. You could also use:

```
corr2(dat3(:,1),dat3(:,2))
```

6. Now fit a straight line of the form $y = b_0x + b_1$ by using:

```
P=polyfit(dat3(:,1),dat3(:,2),1);
```

The output, `P` contains both the fit coefficients. You may need to type `format long` to print the required significant figures.

7. In order to calculate a and b in $y = ax^b$ take the anti-log of `P(2)` (ie. `10^P(2)`).
8. Try plotting actual extinction vs ice water content as a scatter plot:

```
plot(dat(:,2),dat(:,3),'.'))
```

then hold this plot and plot the extinction vs parameterised ice water content and set the axes to display on log scale:

```
hold on;
plot(dat(:,2),10^P(2).*dat(:,2).^P(1),'k-');
set(gca,'yscale','log','xscale','log')
```

5.2 Using your fit to tell you something about Mars

Download the `martian_clouds.csv` file from Blackboard. Either click on it to open it in Excel or type:

```
datm=csvread('martian_clouds.csv',2,0);
```

to read into MATLAB.

1. It contains an extinction vertical profile from Mars for both dust and clouds (look at the headers: height; dust extinction; cloud extinction). Use the cloud extinction and the parameterised equation to derive the vertical profile of ice water content.
2. What is the peak ice water content? (i.e. calculate another column using your power law).
3. What is the average ice water content between the levels of *height* = 2700 and *height* = 3900 metres inclusive? E.g. calculate the average ice water content between these heights:

```
ind=find(datm(:,1)>=2700 & datm(:,1)<=3900);
xbar=mean(iwc(ind));
```

   where `iwc` is the *ice water content* calculated from your power law.

At the end, save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`).