MONT 102N – Environmental Mathematics
Lab on Linear Regression, Correlation, and Data Analysis
October 19, 2011

Background and Goals

In this lab, we want to use some of the ideas we have been discussing about regression, correlation, and so forth on the problem set for this week and try to understand the patterns that are there. We will use many of the statistical features of the Microsoft Excel spreadsheet program.

Regression and Correlation in Excel

To compute the least squares regression line and correlation coefficient for an x-y data set, as we saw in class on Monday, the following options are available in Excel:

- Normally, you would start by entering the x-values in one column and the y-values in another parallel column. So that things line up properly, put the x- and y- values from the same data point on the same row. Put a text header in the cell at the top of each column to help identify the two quantities.
- To generate the y-versus-x scatter plot, highlight the cells in those columns, then press the Insert tab on the top, look for the Scatter option under the Charts category and press that button. You can experiment with the different plotting styles. Each time you generate a plot, it is overlaid on the grid, but you can drag and drop the plots to rearrange them if you want.
- The coefficients m and b in the equation of the least-squares regression line can be computed and displayed in cells of the spreadsheet like this. For example say you have entered the x-values in cells $A3$ through $A25$ and the y-values in cells $B3$ through $B25$ (with column headings in $A2$ and $B2$).
  1. First go to an unoccupied pair of cells on the same row, and highlight them
  2. Enter `=LINEST(B3:B25,A3:A25,TRUE,FALSE)` and press SHIFT/CTRL/ENTER together. Note: the y-values *must* come first, then the x-values, or else you will be computing a regression line of the form $x = My + B$. That actually does makes sense, but it is not what we want here(!)
  3. The slope value m will be computed and put into the left cell where you entered the LINEST command. The intercept value b will be put into the cell next to that one to the right.
- The Excel command for computing the correlation coefficient r is `CORREL`. With data as above, in any convenient cell, enter `=CORREL(A3:A25,B3:B25)`.
- There is also a *second, alternative* way to generate the numbers we want from a linear regression, together with a lot of additional statistical information including some useful plots. From the Data tab, select Data Analysis, highlight Regression in the small Data Analysis window that comes up, and press OK. Fill in the input ranges for the x and y as indicated above (that is, make x correspond to $A3 : A25$ and y correspond to $B3 : B25$), select options to `Plot Residuals` and `Line Fit` and check

the box for labels. Press OK and a new "sheet" will be generated giving all the results of the regression. Note the tabs at the bottom of the grid saying `Sheet 1` and `Sheet 2`. You can toggle back and forth between the main file and the regression results by pressing those tabs.

- We will discuss the *residual plot* in more detail on Friday. For now, the main thing that you need to know is that this plot shows the difference between the y-value from each data point, and the "predicted" y-value for the corresponding x given by the equation of the regression line $y = mx + b$. In other words the residual plot shows $y_i - (mx_i + b)$ for each i, $i = 1, \ldots, n$.
- *We will discuss the meaning of some of the other output later!*

Lab Work

Your goal today is to do as much as possible (hopefully all) of the spreadsheet work for problems 17 and 19 from Chapter 4 on this week's problem set. (You will probably want to print out your work in the lab and attach those sheets to the rest of the problem set.) Use the graphing features of Excel to produce the scatter plots in part a of each problem. Ignore parts c in both problems. (These ask you to use the "straightedge" or "eyeball" method to draw an approximate straight line fitting a data set.