MONT 104N – Modeling The Environment Seminar
A "Crash Course" on Excel, Including Linear Regression
October 15, 2012

General Information on Excel

Excel is a general-purpose spreadsheet and data analysis tool included in the Microsoft Office suite of programs. We will be using it pretty extensively this year to construct and work with several different types of mathematical models and to perform a number of different data analysis tasks.

The instructions below are tailored to Excel 2010 running on a Windows PC. If you are working with an earlier version (e.g. Excel 2007) or if your computer is not a Windows machine, everything should be there, but perhaps accessed in a slightly different way.

- (This step can be done before the lab.) Launch gmail, look for an email from me with two spreadsheet file attachments, and extract and save them. The following discussion will refer to those spreadsheet files.
- From the desktop, double click the shortcut for *Excel* to launch that program.

Take a look at the overall layout of the of the Excel window. There are tabs, menus, etc. similar to many programs, but there are some differences too. In particular note the File tab at the upper left. This is where all of the usual File options are located (i.e. the controls for reading in or saving files, printing, etc.)

Like all spreadsheet programs, Excel gives you a workspace that is composed of a 2D grid of "cells" identified by location – by an *address*. The columns are labeled by capital letters, and the rows are labeled by numbers.

- A single cell is referenced by the column, followed by the row, for instance $B23$ is the cell in column B and row 23.
- A range of cells is referenced by giving the "starting cell," a colon, and the "final cell" in the range. For instance $B2 : B45$ indicates the cells in column B and rows 2 through 45. $B2 : F2$ indicates the cells in row 2 and columns B through F. Similarly, $B2 : D10$ indicates all the cells in a *rectangular block* with upper left corner at cell $B2$ and lower right corner at cell $D10$.
- The addresses seen so far are all *relative addresses*. In other words, they are set up so that if we perform an operation in one cell that depends on the entries to the left in its row, then it is possible to copy and paste that operation to other rows and the entries in the new row will be used. If you want to specify a *fixed* address then put in \$ characters: $\$C\$5$ means the cell with fixed address in column C and row 5. (We will see several examples of this in a while; if it is not clear why we need this distinction, wait until you see the examples!)

The contents of a cell can be a text label identifying what the data in a row or column represents, a number, or a formula indicating how to perform a desired calculation using

other information in different cells within the spreadsheet. When you finish entering a formula this way and press the Enter key, the indicated computation is performed and the result is displayed in that cell. One *very nice* feature of spreadsheets is that if you change the contents of a cell that is used to compute a value this way, then the calculation is automatically performed again to update the value displayed. We will also see this in a moment.

A First Worked Example

Begin by reading in the spreadsheet file `First.xls` that you extracted from my email:

- Press the File tab at the upper left of the Excel window,
- then Open,
- Find the folder where you saved the file `First.xls`,
- Highlight it and press Open at the bottom.

You should now see a rectangular block of cells filled with names, text, and numbers at the upper left of the spreadsheet in rows 1 through 10 and columns A through E. Think of this as the grade book for a small class with 8 students (the rows are labeled with their names) who have had four assignments as in the labels for columns B through E. Note that $A12$ has the text "Average" but there are no numbers on that row (yet!). We are going to use Excel to compute the averages, on each assignment.

- In cell $B13$, enter the formula `=AVERAGE(B2:B9)`. As you type, you will see this showing up in the cell and in the input box above the grid. When you are done press Enter, and the average will be computed and displayed.
- Now we will use the same method to compute the average on each of the other assignments: Highlight cell $B13$ by clicking the left mouse button over that cell. Make sure the Home tab at the top of the Excel window is active, press Copy (next to the "Office Button"), drag the highlighting box so that all the cells in row 13, columns B to E are highlighted, and press Paste (next to Copy). You should now see the averages for each column.
- In doing the averages we were making use of the *relative addressing* mentioned above. Copying the formula in one cell and pasting it into another also changed the addresses of the cells that the formula was applied to. Now, we are going to perform an operation where we want to use contents of a fixed cell on multiple rows. Start by filling in new information in row 14: Put a text label "Weights" in $A14$ and the constants .3 in $B14$, .25 in $C14$, .4 in $D14$, and .05 in $E14$.
- In cell $F1$ add the text label "Course Average." In $F2$ enter the formula

  ```
  =$B$14*B2 + $C$14*C2 + $D$14*D2 + $E$14*E2
  ```

  You should see the weighted average displayed.
- You can now copy and paste that formula to the other cells in column F and rows 3 through 9 to do the same computation for the other students in the class. (Note that

the weights always come from the same row, hence the fixed addresses. Can you see what would happen if we did not do it that way?)

Computing scatterplots, regression lines, etc. with Excel

To create a scatter plot of a bivariate data set (that is a collection of (x_i, y_i) points for $i = 1, \ldots, N$), you will follow these steps:

1) Enter the x_i and y_i values into the spreadsheet in two consecutive columns. (To help you understand what you did if you come back to the spreadsheet later, it is often helpful to enter text headings in the cells at the top of the columns, but that is not necessary.)
2) Highlight the range of cells containing the data by holding down the left "mouse" button and dragging the cursor.
3) With the data highlighted, press the Insert tab of the Excel window, and choose the option Scatter from the Plots group.
4) You should see a bare-bones version of the plot generated at this point.

You will almost always want to edit your plot to add axis labels, a chart title, trendline(s), equation(s), etc. To do this you will use various options in the Layout tab of the Chart Tools group.

5) With the Layout tab of the Chart Tools group highlighted, you should see Chart Title, Axis Titles, Legend, etc. Each of those buttons produces a pulldown menu that you use to add or remove features of the chart. The title options, for instance, add text boxes overlaying the graph that you type in to add the title you want.
6) The Trendline menu contains options that we will be using very extensively for constructing linear and other models. The Linear Trendline button just adds the trendline, though. If you want to be able to generate the equation of the line overlaid on the plot, go to the bottom option in the menu (Other Trendline Options), select the trend/regression type you want, and check the box that says Display Equation on Chart.

To practice these operations, open the `FirstRegressionEx.xls` file from my email in Excel, and try to reproduce the charts in the handout from class on October 5.

An alternate method for computing the regression line

There may be times when the steps above generate more information than you really want or need. Excel can also compute the m, b coefficients in the equation of the regression line $y = mx + b$ by another strictly spreadsheet calculation without any graphics. This is slightly tricky, but once you get the hang of it, it should become automatic:

1) As before, you will need to enter the x and y data into some range of cells of the spreadsheet (two consecutive columns is always a good choice!)

2) Outside that range of cells, highlight a"block" of two consecutive cells on one row
3) Say the y-values are in column B, rows 3 through 13 and the x-values are in column A, rows 3 through 13. You would enter the formula $= LINEST(B3 : B13, A3 : A13, 1, 0)$, then
4) Press Control/Shift/Enter simultaneously. (Note: On a Mac, this step is different – hold down the apple key and return together.)
5) The slope and intercept of the regression line will be printed out in the cells you highlighted in step 2 (with the slope on the left and the intercept on the right)