MONT 105Q – Mathematical Journeys
A "Crash Course" on Excel, Including Linear Regression
March 16, 2016

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 in class for the rest of the semester to perform a number of different data analysis and statistical tasks. You will probably want to use it for the final projects for the course as well.

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 class.) Launch gmail, look for an email from me with two spreadsheet file attachments `First.xls` and FirstRegressionEx.xls, and extract and save them. The following discussion will refer to those spreadsheet files.
- From the desktop (or the apps listing if you have not put a shortcut on the desktop), launch *Excel.*

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 (or pairs of letters, etc. if you go past 26 columns), and the rows are labeled by numbers. You can also have several "sheets" within a spreadsheet file. (Some commands generate output on additional sheets too.) These are accessed via the tabs at the bottom.

- 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 in rows 1 through 6 and columns A through M. The text labels Mean, Minimum, Q1, Median, Q3, Max, Standard Deviation in column A will identify basic statistics that we'll compute now.

- In cell B9, enter the formula =`AVERAGE(A3:M6)`. 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.
- In the rest of column A, enter
- =`MIN(A3:M6)` to compute the minimum,
- =`QUARTILE(A3:M6;1)` to compute the first quartile,
- =`MEDIAN(A3:M6)` to compute the minimum,
- =`QUARTILE(A3:M6;3)` to compute the third quartile,
- =`MAX(A3:M6)` to compute the maximum,
- =`STDEV(A3:M6)` to compute the standard deviation, (Note: Excel uses the "$n-1$"-formula for the SD, not the "n"-formula we discussed in class.)
- Now let's construct a histogram for this data set, using 7 bins corresponding to the ranges of values 7,8,9, then 10,11,12, then 13,14,15, etc. up to 25,26,27. (Note: these are entered in column D for your reference.)
- Enter the command =`COUNTIF(A3:M6,"<=9")` in cell E9 next to the 7,8,9 label. (This should compute the value 6 in that cell indicating that there are 6 numbers in the data set in that range.)
- Then enter =`COUNTIF(A3:M6,"<=12")-E9` in cell E10 next to the 10,11,12 label. Do you see what this does?
- Continue to compute the rest of the rest of the counts by the same method. *This should yield the values: 6,10,15,8,8,4,1*
- *Comment*: Some of you may know other ways to produce counts for a histogram. This is just one of the ways it is possible to do this computation. I'm choosing this one

instead of a short-cut alternative because I think it's good practice for understanding the layout and mechanics of a spreadsheet if that is not already familiar(!)

- Now suppose we want to produce a frequency histogram for this data set. Highlight the counts in column E and use Insert/Chart. Select the option Column to see a basic histogram plot. You can also experiment to add a title to this plot, add labels on the x-axis corresponding to the ranges of values in column D, and so forth. (Note: By placing the cursor over it and left-clicking, the chart can be moved around to locate it anywhere in the spreadsheet.)
- Looking ahead to your next assignment (Problem Set 3), here's an additional feature of spreadsheets that it's good to understand. Suppose we want the average of just the numbers in cells A3-A6 (the first column). We can compute that and place it in cell A7 by entering `=AVERAGE(A3:A6)`. Note that I *did not* put in the dollar signs this time. That means that these are *relative* addresses. If you highlight this formula, copy and paste it into column B, you will see that the column label gets changed, so that the command now computes the average of the entries in column B(!) You can actually paste the command into all of columns B - M by highlighting the whole block on row 7 and pasting.

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 for regression and other sorts of calculations. 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 for class on March 16.

An alternate (optional) 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)