MONT 106N – Identifying Patterns Seminar
Lab on Correlation, Regression, Data Analysis
November 11 and 13, 2009

Background and Goals

In this lab, we want to use some of the ideas we have been discussing about correlation, regression, and so forth to look at some actual real world data and try to understand the patterns that are there. We will use many of the statistical features of the Microsoft Excel spreadsheet program.

General Information on Excel

To get ready for work on this lab project and launch Excel in Beaven 335a,

- Log on the campus network with your username and password.
- (This step can be done before the lab.) Launch Novell Groupwise, look for an email from me with three spreadsheet file attachments, and extract and save them to your campus network P : drive. You will be working extensively with those files for this assignment.
- From the desktop, double click the shortcut for *Excel 2007* to launch the spreadsheet program.

(The other one, Excel NP, is an older version that does most of the same things, but is set up somewhat differently. If you are familiar with that one, it should be OK to use it, but the directions below refer specifically to Excel 2007.)

Take a look at the overall layout of the of the Excel window. There are tabs, menus, etc. similar to many Windows programs, but there are some differences too. In particular note the large "Office Button" at the upper left. This is where all of the usual File options are now 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.xlsx` that you extracted from my email:

- Press the "Office Button" at the upper left of the Excel window,
- then Open,
- Find your P : drive in the folder box at the top of the Open window, highlight the file `First.xlsx`,
- 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" and A13 has the text "SD," but there are no numbers in those rows (yet!). We are going to use Excel to compute the averages and SD's, 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.
- The Excel command for SD is STDEV. Follow what we did for the averages to compute the SDs of the columns B through E.

- *Technical Note:* Excel uses a slightly *different* formula for computing SD's than we have talked about. Here's is Excel's way:

$$SD_x = \sqrt{\frac{(x_1 - \text{ave}_x)^2 + \cdots + (x_n - \text{ave}_x)^2}{n-1}}$$

  (note the $n-1$ in the denominator rather than the n we have used). This means that the results of SD computations will be slightly different than we would get by our formula! But the results will be largely equivalent as a measure of "spread" of data. Why on Earth are there *two formulas* in common use for computing SD's? I'm tempted to say, "don't ask(!)" But here's an explanation: The two formulas estimate "spread" in two slightly different ways. Each has corresponding good properties. So statisticians have never really settled on a single way to do this(!)
- In doing the averages and the SD's 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?)
- Now that all the course "stats" are computed, let's do some analysis. Say we want to know how well correlated the scores on the Midterm and the Final (columns $C$ and $D$) were. Generate a scatter plot by highlighting the cells in those columns and rows 2 through 9, 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 arrange them if you want.
- The Excel command for computing the correlation coefficient r is `CORREL`. In any convenient cell, enter `=CORREL(C2:C9,D2:D9)`. Does this match what you think the scatter plot indicated?
- Now, say you want to perform a regression with $x =$ the Midterm score and $y =$ the Final score. 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 $C2 : C9$ and y correspond to $D2 : D9$), select options to plot residuals and line fit and check the box for labels. Press OK and a new "page" 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 meaning of all the output!*

Lab Investigations

A) As we discussed in class on Monday, regression analysis, in conjunction with various data transformations, is often performed to try to determine whether various sorts of functional relations exist between x and y.

- The *basic regression* of y versus x can be used to look for *linear relations* $y = mx + b$.
- Applying a logarithm function to y and doing regression of $\ln(y)$ versus x can be used to look for *exponential relations* $y = ca^x$ or equivalently $y = ce^{kx}$.
- Applying logarithms to both y and x, and doing regression of $\ln(y)$ versus $\ln(x)$ can be used to look for *power laws* $y = cx^\alpha$.

The data in the file `USMSAs.xlsx` that you downloaded from my email has the populations of the 75 largest SMA's in the US in 2005 (extrapolated from the 2000 census and other measurements). The SMA's are standard metropolitan areas used by government agencies and many others to study demographic and economic trends. They are designed to coincide with the major concentrations of population, not administrative boundaries. Thus, for instance, the SMA for Boston contains not just the City of Boston, but also the first several "rings" of suburban towns around the city.

When you open the spreadsheet file `USMSAs.xlsx`, you will see two long columns of data – the populations, and their ranks. As you can guess, the first few are the MSA's corresponding to New York, Los Angeles, Chicago etc. Fairly early in the list are perhaps unexpected places like Riverside-San Bernardino-Ontario, CA (not traditional "big cities," but major concentrations of population). Not surprisingly, high population states such as California, Texas, and Florida tend to have a lot of the larger MSAs. Worcester, MA is included in the MSA ranked 65!) Investigate the data and try to develop answers to these questions:

1) Does it seem like there is a linear relation between x = rank and y = population? Look at the value of r and the residuals for the regression – strong patterns in the residuals indicate a "lack of fit."
2) What about a linear relation between $\ln(y)$ and x? (Note: to compute the natural log of the cell $C4$ (for instance) in Excel, you can use `=LN(C4)` in another cell.) Again, look at the value of r and the residuals.
3) What about a linear relation between $\ln(y)$ and $\ln(x)$? Once again, look at the value of r and the residuals.
4) It should be fairly clear from the data that the 8 or so largest MSA's are somewhat unrepresentative of the rest. What happens if you repeat 1,2,3 on just the 9th through the 75th MSA's?
5) What can you say about a functional relation between x = rank of the MSA, and y = population based on what you have found?
6) (Extra Credit) Does the pattern you found here hold in other countries as well? It is fairly easy to find data online about the largest cities in countries around the world.

Try the United Kingdom, India, etc. Is the pattern pretty general, or are there differences?

B) The data in the spreadsheet file `SemiCond.xlsx` was developed by Jacqueline M. Hughes-Oliver, Department of Statistics, North Carolina State University, Raleigh, NC. I found it in an online database of interesting datasets maintained by the American Statistical Association. The data comes from an experiment about techniques for manufacturing the semiconductors used in computer chips and other electronic components.

The data consists of measurements of polysilicon thickness at 13 sites on 22 wafers processed using Rapid Thermal Chemical Vapor Deposition. The processing conditions are determined by:

(a) thickness of oxide applied to the wafer prior to deposition of polysilicon, and
(b) deposition time (i.e. how long the chemical vapor containing the polysilicon was allowed to deposit on the wafers.

The columns are given in the following order: wafer label, oxide thickness (in angstroms), deposition time (in seconds), polysilicon thickness (in angstroms) for locations 1 to 13.

The goal of the study was to study how the combination of thickness of oxide and deposition times affected the *uniformity* of the thickness of the layer of polysilicon across all the sites on the wafers. (This is probably an important consideration for using the wafers to produce microchips, but I can't say I understand every aspect of the technical background! The same would generally be true of statisticians called in to consult on a study like this one.)

1) What statistic could we use to measure uniformity of the polysilicon thickness across all the sites on one of the wafers? Compute this for each of the wafers.
2) If you look carefully at the data, there is a good case to be made for *excluding* the data from site 13 on each wafer as an unrepresentative "outlier." How would you make that case in the most convincing way, *from the data*? (That is, it is *not necessary, or relevant* to speculate about a possible cause for the discrepancy between the site 13's and the other sites. Just argue "from the numbers.")

In this example, since the oxide thickness and the deposition time *both vary*, we really want to study a possible relation of the form

$$y = m_1 x_1 + m_2 x_2 + b \tag{1}$$

where y = your measure of uniformity from 1), x_1 = oxide thickness, and x_2 = deposition time. We are going to use a variant of regression, called *multiple regression* for this. (Note: It would be possible to do separate regressions of y against x_1 and y against x_2, but that is *never the right approach* in a problem like this! It's a total statistical "no-no" in fact! As you might guess, the point is that we want to take the values of x_1 and x_2 into account when looking at how they affect y.)

3) To do a multiple regression in Excel, you use the same Data Analysis/Regression commands we talked about before, but instead of making the x be data from a single

column, you put in range of cells coming from a *rectangular block* of several columns. Fortunately, our spreadsheet file is set up well for that – the two variables x_1 and x_2 are in columns B and C. The output from the regression is generated in a separate page as before, and if you look closely, you will see the estimated values of the two coefficients m_1, m_2 and intercept b in (1) above.

4) Try to interpret the output you are getting from Excel. Is the relation (1) a good fit for the data or not?
5) What do the *signs of the coefficients* m_1 and m_2 indicate about the roles of the oxide thickness and the deposition time in determining the measure of polysilicon thickness uniformity?

Assignment

Submit your finished spreadsheets for Investigations A and B by email. (The first worked example will *not be submitted* – it was mainly practice to get you "up to speed" working in Excel.) To help me keep track of everyone's work, *please rename the spreadsheet files with your last name, and the letter A or B* – for instance `MonfetteA.xlsx` for Greg's final work on investigation A. Also, write up a short *report* (1 or 2 pages) on your conclusions in a separate Word document. This should contain your answers to the questions posed in the problems above. Attach this to your email as well. The completed worksheets and reports will be due on Friday, November 20.