Using Excel to perform regression analysis

This set of notes is just a basic description of how to use Excel to conduct regression analysis. If you already know how to use Excel to do this, then that is fantastic. If you have some other program that you use for statistical analysis (SAS, Stata, Eviews, Matlab, R, etc.), that is fine too – the great thing about regression analysis is that all the programs give the same results because regression analysis basically just involves inverting matrices and multiplying them together to obtain the coefficient estimates (other types of econometrics, such as limited dependent variable analysis, involve convergence based on tolerance levels which can and do differ across programs).

If you have not used Excel for regression analysis, and you do not have other statistical software to use, then read below (or read the Excel help file). I am using the "old-fashioned" method – I believe some versions of Excel have a "point and click" method for regression analysis which I don't have access to on my computers (then again, I don't use Excel for estimating models).

Instructions:

To the best of my knowledge, multiple regression analysis works best in Excel when all of the data being used as independent variables are in columns that are next to each other (if there were 5 independent variables, they should be in columns C – G). The dependent variable does not have to be in a column next to the independent variables (it could be in column A or column K, for instance). If you have your data set up in the preceding manner, then it is straightforward to estimate a multiple regression model in Excel.

1. Highlight a block of cells that is 5 x (# of independent variables + constant term). So if you have 6 independent variables, then you will want to highlight a 5x7 block of cells.

2. Start typing the multiple regression command for Excel. It is =LINEST(Range of dependent variable, range of independent variables, TRUE, TRUE). The command "LINEST" is for "Linear estimation". The range of the dependent variable should be in a single column (C2:C49). The range of the independent variables should be in the adjacent columns, and may be something like (D2:I49). The first "TRUE" value is for including a constant, and the second "TRUE" value is for including statistics. Thus, your entry should look like:

   ```
   =LINEST(C2:C49,D2:D49,TRUE,TRUE)
   ```

   Note that as you are typing this command the entry will only appear in one of the cells. Before "unhighlighting" the cells, continue to step 3.

3. The LINEST command requires entry as an array. With the cells highlighted, press "Ctrl+Shift+Enter". I would press "Ctrl" first, keep holding it down, then press "Shift", and keep holding both buttons down, and then press "Enter". The key is to NOT press "Enter" first

(otherwise you won't get the full results). Like magic, your highlighted block of cells should fill in with numbers.

4. I'm going to reproduce the results from the Mrs. Smyth's pie example here:

2815.494751	0.030258034	2.042729	29866.59	5.837648	-122607	529773.7
4539.242557	0.0039448	3.762305	13449.22	1.650494	16422.38	271330.9
0.871042264	67583.77555	#N/A	#N/A	#N/A	#N/A	#N/A
46.15560361	41	#N/A	#N/A	#N/A	#N/A	#N/A
1.26491E+12	1.8727E+11	#N/A	#N/A	#N/A	#N/A	#N/A

What I find to be the most challenging aspect of Excel's estimation is remembering what variables the rows and columns correspond to. The first row will be coefficient estimates for your independent variables. The second row will be standard errors for those same variables.

Note that these variables are in reverse order of how they appear in the spreadsheet. For instance, my independent variables ranged from D2:I49, so I had columns D, E, F, G, H, and I as independent variables, plus I included a constant. The columns D, E, F, G, H, and I correspond to the variables Price, Advertising Expenditures, Competitor's Price, Income, Population, and Time. The first column of results (with the coefficient estimate of 2815.49 and standard error of 4539.24) is for the LAST independent variable, Time. The second column is for Population, the third column for Income, etc. The very last column, with 529773.7 and 271330.9, contains the estimates for the intercept and its associated standard error. Note that none of the reported results are t-statistics, so that you will have to construct your own t-statistics.

There are 6 other cells with numbers in them. I will reproduce those six cells and have a corresponding set of 6 cells that contain a description of the information provided.

0.871042264	67583.77555	R^2	SEE
46.15560361	41	F-stat	DF
1.26491E+12	1.8727E+11	RSS	ESS

SEE is the standard error of the estimate, F-statistic is the F-statistic for the joint significance of ALL the regressors, DF is the degrees of freedom, RSS is the regression sum of squares, and ESS is the error sum of squares.

5. The process for producing the estimates is fairly straightforward, but I find it difficult to interpret the resulting output. Other statistical software produces much clearer output, so if you are familiar with other software I do not mind if you use it.