**Multiple Regression Notes**

I. [Imputing Missing Values](http://en.wikipedia.org/wiki/Missing_data)

You cannot run a regression with any missing values. If there are some missing values, you can:

* + - Use the average of the existing values for each of the missing values - which renders the variable useless if there are many missing values
    - Or, do a linear regression with that variable (for the rows that you do have values) and another variable (other than your y-variable!) to estimate the missing values
    - Research to find the values
    - Worst case – eliminate the rows with missing data, but only do this if you have a lot of rows

II. [Correlation](http://en.wikipedia.org/wiki/Correlation)— determining if there is a statistical relationship between two variables (this step is optional but is a good idea to see what is correlated with your y-variable)

* Excel function “correl” {correl(array1, array2)} will give correlation as r-value. r can be between -1 (negative correlation) and 1 (positive correlation) inclusive. The direction of the relationship is very important. You are most interested in the correlation between the y-variable (response variable) and the x-variables (candidate variables).
* When you square r, you get the percentage of change in the variable y (output) that can be explained by the variable x (input). The R-squared can be between 0 and 1 inclusive. Higher is better.
  + Correlation in the [Excel Data Analysis Toolpak Add-in](https://support.office.com/en-us/article/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6) you can run many correlations at once.

Data->Data Analysis->Correlation

Check “labels in first row” so you know what you’re looking at

Gives a correlation matrix

[CORRELATION DOES NOT PROVE CAUSATION](http://en.wikipedia.org/wiki/Correlation_causation)—things can have similar trends without causing each other

III. [Multiple Regression Analysis](http://en.wikipedia.org/wiki/Regression_analysis) in Excel

* + In your dataset: y-value column needs to be first, x-value columns need to be contiguous, you can run up to 16 input variables at a time
  + Transformations – in a separate column for each transformed variable, transform the values of a variable then use in the regression (it is atypical, but acceptable, to have the original variable and its transformation in a model)
    - Ladder of Powers (of commonly used transformations)

1/x : reciprocal

LN(x) : natural logarithm

x½ : square root

x² : square

* + - You can look at a graph of the y-variable vs. the x-variable to determine the shape of the relationship to help determine which transformation, if any, may best represent the data.
    - Alternatively, you can do all 4 transformations then:
      * Throw all of them into multiple regression and remove iteratively (see below)
      * Use correlation to determine the best transformation
  + [Interactions](http://en.wikipedia.org/wiki/Interaction_%28statistics%29) – the effect of two variables may not be additive, the synergy created by the interaction of the two variables may be found by multiplying the two variables (Col1\*Col2) in a third column and using the result in a regression. It is very common for the interaction term and at least one of the other individual variables to be in the final model.
  + In Excel: Data->Data Analysis->Regression (in the [Excel Data Analysis Toolpak Add-in](https://support.office.com/en-us/article/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6))
  + Let the regression do the work in deciding which variables should be included in the model

Click on ‘Labels’ to capture them (otherwise, refers to data as X1, X2, etc.)

Typically, do not use “Constant is Zero”

Keep Confidence Level: 95%

Hit “OK”

SUMMARY OUTPUT analysis:

* **Variables in model must be significant: P-values for each variable must be under approx. 0.05.**
* If it’s over 0.05 (ie, insignificant), it should NOT be included in your final equation
* Ignore the p-value for the Intercept
* Iterative process:

1. Remove worst variable (highest p-value)
2. Rerun the regression
3. Repeat steps #1-2 until all remaining variables are significant (all p-values<0.05)

* Coefficients for each variable will be in ‘Coefficients’ column

{ie, y = Intercept + (Coefficient1\*X Variable 1) + (Coefficient2\*X Variable 2)+ …}

* Also, the sign of the coefficients must make sense – you can use the correlation to confirm.
* Compare models using ‘Adjusted R-squared’ – higher is better

**Multiple Regression Guides:**

[A Quick Guide to Using Excel 2007’s Regression Analysis Tool](http://faculty.fuqua.duke.edu/~pecklund/ExcelReview/Use%20Excel%202007%20Regression.pdf)

[EXCEL 2007: Multiple Regression](http://cameron.econ.ucdavis.edu/excel/ex61multipleregression.html)

[YouTube Video: Regression Analysis in Excel Part](http://www.youtube.com/watch?v=nyCVa6M1IFo)