ChE 263
Handout: **Data Analysis** Using Excel

Statistical Functions

Excel contains about eighty intrinsic functions for data analysis. They can each be invoked by clicking ***f***f_x on the formula bar or in the "Formulas" ribbon, and then selecting the Statistical category. Some of the most common are described below. In these descriptions, the arguments *A*,*B*,*C*,…, may be either single values or ranges. Function names are case independent.

- Average(*A*,*B*,*C*,…) The arithmetic mean of the elements of *A*,*B*,*C*,…

$$\text{Average} = \frac{1}{\text{n}}\sum_{\text{n}} A_i$$

- Correl(*A*,*B*) The Pearson correlation coefficient, R, for the data sets *A* and *B*. When $R^2 = 0$, no correlation exists between the sets; when $R^2 = 1$, *A* and *B* are perfectly correlated. R^2 can be considered equal to the fraction of the variations in the dependent variable that are the result of variations in the independent variable. R^2's are easily determined using Excel's Trendlines. However, they must be calculated separately when Solver or other regression functions are employed.
- RSQ(*A*,*B)* Instead of using "=(correl(A,B))^2" to find the R^2 value you can use this function. See more detailed description above under Correl. R^2 can also be known as the "coefficient of determination".
- Count(*A*,*B*,*C*…) The total number of cells in the ranges *A*,*B*,*C*,… that contain numeric values, including dates and times.
- Max(*A*,*B*,*C*…) The maximum number of all the elements of *A*,*B*,*C*,…
- Median(*A*,*B*,*C*…) The element of *A*,*B*,*C*,… that has an equal number of larger and smaller values. If there is an even number of elements, the average of the two mid-range values is found.
- Min(*A*,*B*,*C*…) The minimum number of all the elements of *A*,*B*,*C*,…
- Mode(*A*,*B*,*C*…) The value of the elements of *A*,*B*,*C*,… that occurs most often.
- Rand() An evenly distributed random number between zero and one. A new random number is produced every time the worksheet is calculated.
- Randbetween(*a*,*b*) An evenly distributed random <u>integer</u> between *a* and *b*. A new random number is produced every time the worksheet is calculated.
- Var(*A*,*B*,*C*…) The variance of the elements of *A*,*B*,*C*,…

$$\text{Var} = \frac{1}{\text{n}}\sum_{n}\left(A_i - \text{Average}(A)\right)^2$$

- Stdev(*A*,*B*,*C*…) The standard deviation of the elements of *A*,*B*,*C*,…

$$\text{Stdev} = \sqrt{\text{Var}}$$

Based Upon Dr. Hugh Hales' "Computer Skills For Engineers and Scientists" and used with Permission

Regression Analysis (Curve Fitting) Using Charts

Regression analysis is particularly easy to do with a variety of functions on data which have been plotted in Excel. Just select the chart; under “Chart Tools” ribbon, select the “Layout” tab; then click on the “Trendline” button and select “More Trendline Options…”. Now you can select the Trendline type, specify whether to display the trendline equation and R^2 value or set intercept value. Several regression functions are available:

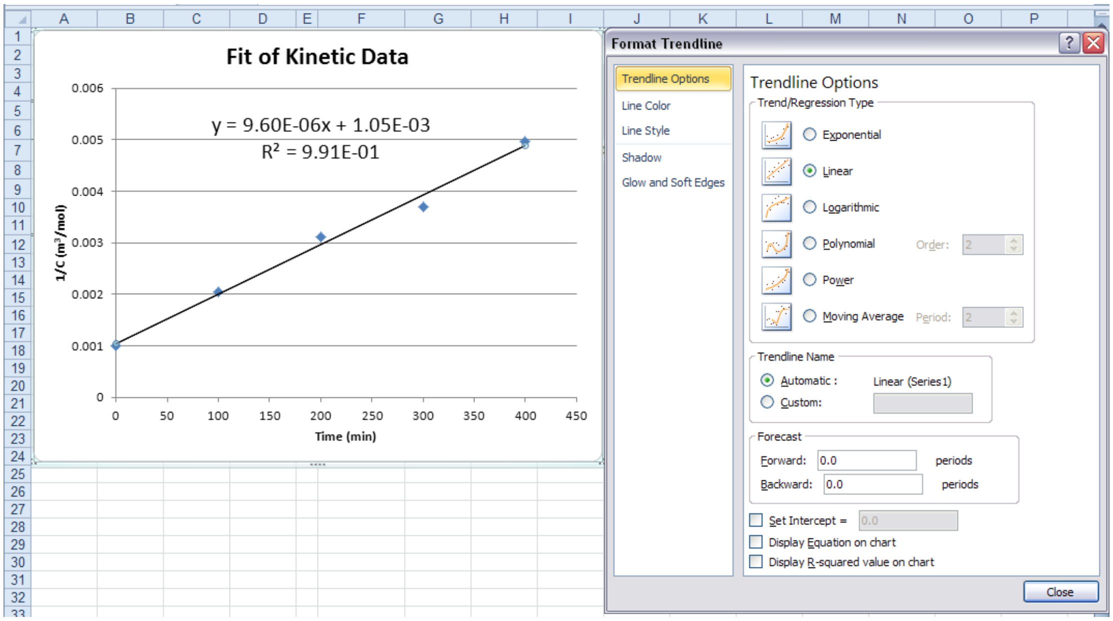
- Linear $y = c_0 + c_1 x$
- Polynomial $y = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_6 x^6$
- Exponential $y = c\, e^{bx}$
- Logarithmic $y = c \ln(x) + b$
- Power $y = c\, x^b$

The approximating equation and its correlation coefficient (R^2) can be displayed on the chart and the trendline can be edited by right clicking on the Trendline and selecting “Format Trendline…”

- Moving Average is not a regression technique. Rather, it simply soothes the data by plotting the average of the data point and its (Period – 1) previous neighbors. These averages are connected with straight lines.

Note: When using trendline to determine the equation coefficients and R^2 value, only a few significant digits are shown. To see more significant digits do the following:

1. Right click on the box containing the equation and/or R^2 value on the chart
2. Select “Format Trendline Label….”
3. Select Either “Number” or “Scientific” under Category
4. Change “Decimal Places:” to display the number of desired significant digits.



Trendline Options

Trend/Regression Type

○ Exponential
● Linear
○ Logarithmic
○ Polynomial Order: 2
○ Power
○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (Series1)
○ Custom:

Forecast

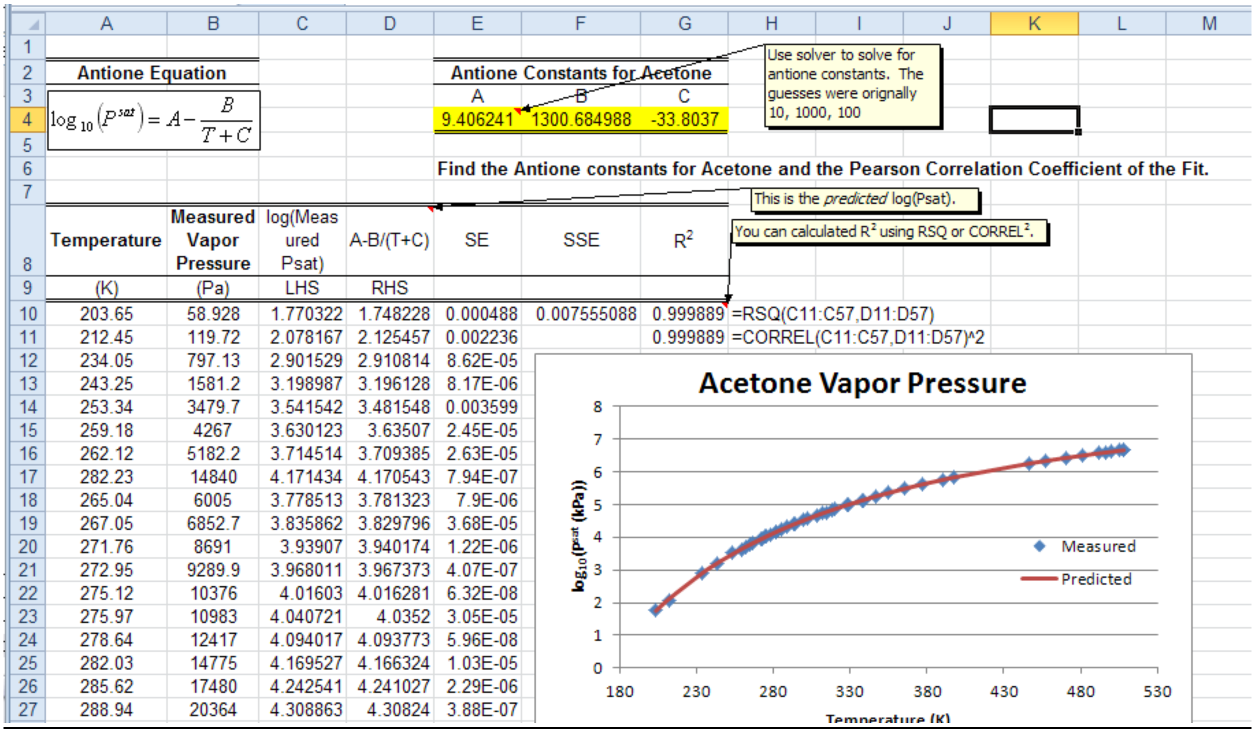
Forward: 0.0 periods
Backward: 0.0 periods

□ Set Intercept = 0.0
□ Display Equation on chart
□ Display R-squared value on chart

Regression Analysis (Curve Fitting) Using Solver

Regression analysis can also be performed using solver and statistical functions. This type of regression analysis is performed when we are fitting an equation that does not match up with the types available with trendline (e.g. y=tan(x), y=ln(x)+2x^2, etc.). In this case we will use SSE (sum square of errors) where the square of errors is between the measured value and the value we calculate based upon our equation. The equations constants are then solved for by forcing the SSE close to zero using "Set Objective" in solver and selecting the equation constants as "By Changing Cells"" in solver. Good guess values are particularly important when the equations are complex. One technique is to try very large numbers, very small numbers, negative numbers and then different combinations. It is typically helpful to see a graph of the measured values and the line from the equation you are using to see what guess values you should change before you try solver again as you seek to get solver to converge.

The R^2 value can be determined using RSQ or the square of the CORREL functions and inputting both the measured and equation calculated data (e.g. "=RSQ(*measured_data_array*,*calculated_data_array*)").



Antione Equation

$$\log_{10}\left(P^{sat}\right) = A - \frac{B}{T+C}$$

Antione Constants for Acetone

A	B	C
9.406241	1300.684988	-33.8037

Use solver to solve for antione constants. The guesses were orignally 10, 1000, 100

Find the Antione constants for Acetone and the Pearson Correlation Coefficient of the Fit.

This is the *predicted* log(Psat).

You can calculated R^2 using RSQ or CORREL^2.

Temperature	Measured Vapor Pressure	log(Measured Psat)	A-B/(T+C)	SE	SSE	R^2	
(K)	(Pa)	LHS	RHS				
203.65	58.928	1.770322	1.748228	0.000488	0.007555088	0.999889	=RSQ(C11:C57,D11:D57)
212.45	119.72	2.078167	2.125457	0.002236		0.999889	=CORREL(C11:C57,D11:D57)^2
234.05	797.13	2.901529	2.910814	8.62E-05			
243.25	1581.2	3.198987	3.196128	8.17E-06			
253.34	3479.7	3.541542	3.481548	0.003599			
259.18	4267	3.630123	3.63507	2.45E-05			
262.12	5182.2	3.714514	3.709385	2.63E-05			
282.23	14840	4.171434	4.170543	7.94E-07			
265.04	6005	3.778513	3.781323	7.9E-06			
267.05	6852.7	3.835862	3.829796	3.68E-05			
271.76	8691	3.93907	3.940174	1.22E-06			
272.95	9289.9	3.968011	3.967373	4.07E-07			
275.12	10376	4.01603	4.016281	6.32E-08			
275.97	10983	4.040721	4.0352	3.05E-05			
278.64	12417	4.094017	4.093773	5.96E-08			
282.03	14775	4.169527	4.166324	1.03E-05			
285.62	17480	4.242541	4.241027	2.29E-06			
288.94	20364	4.308863	4.30824	3.88E-07			

Data Analysis Tools

Nineteen additional data analysis tools are available in Excel through the Data Analysis Toolpak. They can be selected from the dialog box that appears when Data Analysis is selected from the Data ribbon. If Data Analysis is not listed in the Analysis section of the Data ribbon, add it by doing the following: 1) Select File ribbon, 2) Select Options, 3) Select Add-Ins, 4) Click on Analysis ToolPak (not to be confused with the VBA version), 5) Click the Go button. The following are some useful Data Analysis Tools:

Random Number Generator is available from the Data Analysis dialog box. This tool generates random numbers corresponding to several probability distribution functions. Random numbers are often useful when creating synthetic data sets. Excel's Rand and Randbetween functions also generate random numbers. However, their random numbers are uniformly probable. The Random Number Generator tool offers the following distributions: **Uniform, Normal, Bernoulli, Poisson, Patterned, Discrete**

Rand and Randbetween also regenerate their numbers every time the worksheet is recalculated. The Random Number Generator creates numbers that do not change.

The procedure is as follows:

1. Select Random Number Generator from the Data Analysis dialog box. Click OK.
2. Input the desired Number of Variables (# of columns you would like the data to be inputted into) and Number of Random Numbers (# of rows you would like the data to be inputted into) blank
3. Select the Distribution type.
4. Specify the desired Parameters.
5. Specify the Output Options. Typically select Output Range: and input the top left cell in the range of cells you would like the data to be inputted.
6. Click OK.

Histogram another tool available from the Data Analysis dialog box. Histograms are charts that summarize the number of occurrences of values in a data set. Histograms provide an easily understood depiction of the distribution of the data. The procedure for creating a histogram is as follows:

1. Select Histograms from the Data Analysis dialog box, Click OK.
2. In the Histogram dialog box that appears, specify the range of the data to be analyzed for the histogram in the Input Range text box. If these values contain header labels, check the Labels box. Although specification of the address of a list of data bounds in The Bin Range text box is optionally, its use is generally recommended. The specified range should contain the histogram bars' delimiters (i.e. the bins' upper and lower limits). Select an output range. Click OK.

a. Note: The Bin Range can be initially confusing to students. The Histogram compares the data to the set bin range and **counts the number of data points that are the same as or below a given bin but above the next lowest bin value**.

3. A table of frequencies and bins is generated. The table contains one more frequencies than bins. The first frequency corresponds to the number of occurrences that are less than the first bin. The second frequency is the number of occurrences between the first and second bins. The third, the number between the second and third bins, etc. The last frequency is the number of occurrences greater than the last bin. To graph the actual histogram, create a Column Chart of the frequencies vs. the bins. Do not use the Chart Output option as it has been finicky in the past and caused excel to freeze when trying to modify or delete the outputted chart. So far with Excel 2010 it hasn't been a problem, but I haven't tested it extensively.